Využití moderních přístupů pro detekci plagiátů

Zdeněk Češka

Katedra informatiky a výpočetní techniky, Fakulta aplikovaných věd
Západočeská univerzita v Plzni
Univerzitní 22, 306 14 Plzeň, Česká republika
zceska@kiv.zcu.cz

Abstrakt. *Plagiátorství je v současnosti nejvíce skloňovaným pojmem, se kterým se můžeme setkat v každé oblasti lidské tvůrčí práce. Školství je jednou z důležitých oblastí, kde je nutné tomuto zamezit. V tomto článku se zabýváme moderními přístupy pro detekci plagiátů textových dokumentů. Naše metoda využívá normalizaci textu a latentní sémantickou analýzu pro nalezení skrytých vztahů mezi dokumenty. Dále uvádíme předběžné experimenty provedené na testovacím korpusu, který obsahuje 950 textových dokumentů o politice. Předběžné experimenty naznačují výhodnost naší metody a zlepšení výsledků oproti ostatním přístupům. V závěru článku diskutujeme využití WordNet tezauru pro zlepšení přesnosti současných metod a možnosti identifikace plagiátů, které byly přeloženy do jiných jazyků.*

1 Úvod

Plagiátorství je současným problémem, se kterým se potýkáme v každé oblasti tvůrčí lidské práce. Jedním z možných řešení, které se často prosazuje, jsou nejrůznější ochrany zabraňující kopírování digitálních médií. Ačkoli takovýchto ochran existuje nepřeberné množství, vždy se podaří nalézt nějakou slabinu a patřičnou ochranu deaktivovat. Internet můžeme považovat za zvláštní případ média, kde jsou jakékoli informace volně dostupné. Na Internetu lze bez problemů nalézt obsah většiny CD, DVD a dalších médií v nechráněné podobě, díky čemuž tyto ochrany pozbývají smysl.

Cílem není chránit informace, ale vyhledávat plagiátory a patřičně je trestat. Hlavní výhoda tohoto přístupu spočívá v psychologii, kdy každý plagiátor může být odhalen, porovná-li se jeho práce s databází již existujících děl. Školství je jednou z oblastí, kde kopírování cizích prací velmi škodí a brání tak přirozené tvořivosti studentů.

Clough [2] a Maurer [7] provedli malé srovnání aktuálního stavu metod pro detekci plagiátů mezi textovými dokumenty. V tomto článku jdeme hlouběji a popisujeme moderní metodu pro detekci plagiátů s využitím latentní sémantické analýzy (Latent Semantic Analysis – LSA) spolu s normalizací textu pro odhalování skrytých sémantických asociací mezi frázemi. Zvláštním rysem naší metody je zpracování celého korpusu najednou. Při tomto zpracování se využívá globální statistika všech obsažených dokumentů a zlepšuje se přesnost detekce plagiátů.

Další text v článku je organizován tímto způsobem. Sekce 2 popisuje současný stav v oblasti detekce plagiátů. Sekce 3 navrhuje metodu založenou na LSA. Porovnání naší metody s ostatními je uvedeno v Sekci 4. Sekce 5 popisuje budoucí práce na naší metodě a konečně Sekce 6 je souhrnem našich dosažených výsledků.

2 Současné metody

Metody pro detekci plagiátů lze rozdělit na metody pro zpracování psaného textu a metody pro zpracování zdrojových kódů. Detekce plagiátů zdrojových kódů je v současné době již poměrně vyřešená, což způsobuje především pevná struktura kódu. V tomto článku se budeme nadále zabývat psaným textem, a to z důvodu jeho uplatnění ve školství na nejrůznější semestrální, bakalářské a diplomové práce. Tabulka 1 pak prezentuje rozdělení metod pro detekci textových plagiátů dle složitosti použitého algoritmu a počtu dokumentů, které daná metoda zpracovává najednou. Toto rozdělení bylo původně publikováno Lancasterem [3].

Tabulka 1. Rozdělení metod pro detekci plagiátů textových dokumentů

Typ rozdělení		Popis
Složitost použité metody	Povrchní	Metrika je počítána bez jakékoli znalosti lingvistických pravidel nebo struktury dokumentů
	Strukturní	Metrika je počítána s částečným porozuměním dokumentů
Počet dokumentů které, se zpracovávají u dané metody	Jednotlivá	Pro výpočet této metriky se zpracovává pouze jeden dokument, tj. dvě jednotlivé metriky mohou být využity pro výpočet párové podobnosti
	Párová	Dva dokumenty se zpracovávají současně pro výpočet metriky
	Multidimenzionální	*M* dokumentů se zpracovává společně pro výpočet metriky
	Korpální	Všechny dokumenty obsažené v korpusu se zpracovávají společně pro výpočet metriky

Jedním z nejpopulárnějších systémů je SCAM [10] založený na modelu relativních frekvencí slov, tzv. RFM modelu. Tato metoda může být klasifikována jako Povrchní, Párová. Podobně lze zařadit systém „Detection of Duplicate Defect Reports" [8], který pracuje na principu vektorového modelu, tzv. VSM. Ačkoli systém Ferret [5] využívá slovních trigramů pro nalezení překryvu textu mezi dvěma dokumenty, jedná se stále o Povrchní, Párovou metodu. Důvodem tohoto zařazení je porovnávání trigramů bez hlubšího porozumení souvislostí uvnitř textu.

V následující textu se zabýváme metodou, která je založana na LSA. Tuto metodu lze klasifikovat jako Strukturální a Korpální z důvodu sofistikovaného předzpracování textu a jeho následné hlubší analýzy.

3 Detekce plagiátů s využitím LSA

Námi navrhovaná metoda využívá LSA [4] pro odvození skrytých sémantických asociací mezi frázemi obsaženými v textu. Každá fráze je v našem případě reprezentována slovním N-gramem, který si lze představit jako posloupnost *n* slov následujících bezprostředně za sebou. V dalším textu popisujeme jednotlivé kroky procesu zpracovávající textové dokumenty.

3.1 Předzpracování textu

Předzpracování je jedním z klíčových kroků pro dosažení kvalitních výsledků u úloh zabývajících se zpracováním přirozeného jazyka (Natural Language Processing – NLP). V našem případě využíváme techniky pro mazání stop-slov a lematizaci. Mazání stop-slov je základní NLP technika, která odstraňuje všechna bezvýznamná slova v závislosti na definovaném slovníku. Lematizace [11] je následný proces pro získání základního tvaru slova, takzvaného lemmatu.

3.2 Extrakce frází

V dalším kroce extrahujeme fráze (ve smyslu N-gramů) předem zvolené délky z předzpracovaného textu. V našich předběžných experimentech jsme se zaměřili na N-gramy délky 1 až 5. N-gramy délky 1 jsou ve skutečnosti pouze jednotlivá slova a používáme je pro srovnání s metodami VSM a RFM.

3.3 Analýza a redukce frází

Čím delší fráze extrahujeme, tím vzniká větší množství unikátních frází, které musí být porovnávány napříč všemi dokumenty. Z tohoto důvodu velké množství frazí neuměrně zvyšuje časové požadavky na výpočet při aplikaci LSA. Pro redukci frází na přijatelnou úroveň jsme vytvořili filter založený na počtu dokumentů, ve kterých se daná fráze vyskytuje, tzv. DF filter. V závislosti na tomto filtru určujeme, zda je daná fráze důležitá či ne. Fráze, které se nacházejí pouze v jednom dokumentu jsou odstraněny okamžitě, protože nemohou být plagiovány v ostatních dokumentech. V dalším kroce odstraňujeme fráze, které se vyskytují ve více než $\mu + \sigma$ dokumentech, kde μ je střední hodnota počtu dokumentů, ve kterých se daná fráze nachází a σ je směrodatná odchylka od střední hodnoty. Tento krok odstraňuje všechny velmi často se opakující fráze, které lze považovat za bezvýznamné.

Tabulka 2. Počet frází před a po aplikaci DF filtru. Experiment byl proveden na vzorku 1000 zpráv standardního ČTK korpusu.

Délka fráze	Počet původních frází	Počet frází po redukci	Průměrný výskyt stejné fráze
1	30550	15343	7.45
2	128449	28206	1.76
3	169093	23337	1.34
5	189621	18281	1.18
7	195999	15549	1.13
9	199421	13536	1.10

Tabulka 2 zobrazuje počty frází před a po aplikaci DF filtru. DF filter významně ovlivňuje delší fráze, kde se s rostoucí délkou výrazně zvyšuje redukční poměr. Důvodem jsou především dlouhé fráze, které se vyskytují pouze v jednom dokumentu. V případě frází délky 5 lze dosáhnout až 10-ti násobného redukčního poměru.

3.4 Vytvoření zjednodušeného modelu dokumentů

Dále vytvoříme zjednodušený model vztahů mezi frázemi a dokumenty, který může být popsán maticí A. Nechť A je n x m obdélníková matice složená z n vektorů $[A_1, A_2, \ldots, A_n]$, kde vektor A_i představuje fráze obsažené v dokumentu i. Vektor A_i se skládá z m prvků $a_{i,j}$, kde každý prvek představuje váženou frekvenci výskytu fráze j v dokumentu i, jak naznačuje rovnice (1). Tato rovnice je modifikací standardního TF-IDF váhování [9].

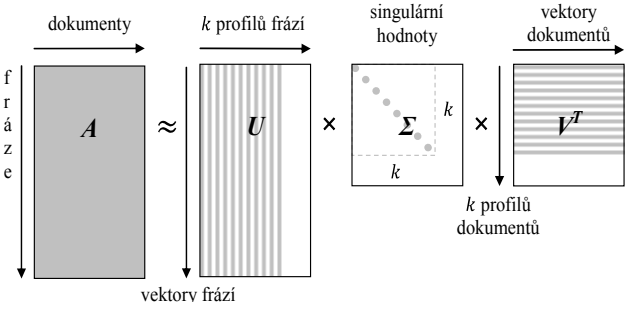
$$a_{i,j} = \begin{cases} \frac{1}{2} + \frac{PF_{i,j} \cdot \log\left(\frac{|n|}{DF_j}\right)}{2 \cdot \max_i\left(PF_{i,j}\right) \cdot \log\left(|n|\right)} & \text{jestliže se fráze } j \text{ nachází v dokumentu } i \\ 0 & \text{jinak} \end{cases} \quad (1)$$

$PF_{i,j}$ představuje frekvenci výskytu fráze j v dokument i, DF_j označuje počet dokumentů, ve kterých se nachází fráze j a $|n|$ je celkový počet zkoumaných dokumentů. Rozdíl oproti TF-IDF spočívá v IDF normalizaci tak, aby $a_{i,j} \in <0.5,1>$. V případě, že fráze j se nenachází v dokumentu i, $a_{i,j} = 0$. Tento způsob váhování dosahuje nejlepších možných výsledků v dalším kroce, který se zabývá dekompozicí matic.

3.5 Latentní sémantická analýza

V tomto kroce se odvozují skryté sémantické asociace mezi frázemi, které jsou obsaženy ve zkoumaných dokumentech. Pro odhalení těchto vztahů využíváme metodu singulární dekompozice (Singular Value Decomposition – SVD), která rozkládá matici A na tři nezávislé matice U, Σ a V^T. Všechny tyto matice mohou být dekomponovány s redukovaným skrytým prostorem k pro získání nejlepší k-té aproximace A, viz [1]. Toho docílíme přepsáním singulární hodnot $\sigma_{k+1}, \sigma_{k+2}, \ldots, \sigma_m$ číslem 0, kde $1 \leq k \leq m$. V našem případě matice U je n x k sloupcově ortonormální, jejíž sloupečky představují singulární vektory frází. Σ je k x k diagonální matice bez záporných a nulových hodnot, které představují singulární hodnoty. A konečně matice V^T je k x m řádkově ortonormální, jejíž řádky představují singulární vektory dokumentů.

Obrázek 1 zobrazuje dekompozici matice A mnohem detailněji. Ve výsledku matice V^T obsahuje jednotlivé profily dokumentů a je základním stavebním prvkem pro výpočet podobností mezi dokumenty.



Obrázek 1. Dekompozice matice frází zastoupených v dokumentech prostřednictvím metody SVD

3.6 Normalizace podobností mezi dokumenty

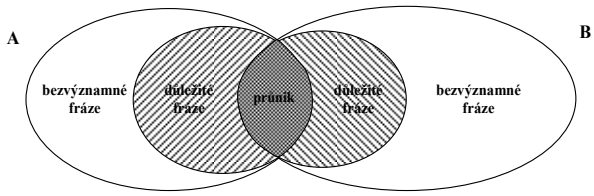
V posledním kroce počítáme podobnosti mezi jednotlivými páry dokumentů. Nejdříve je nutné přenánásobit matici V^T singulárními hodnotami pro získání správného rozměru jednotlivých prvků v profilech dokumentů, což popisuje rovnice (2).

$$B = \Sigma \times V^T \tag{2}$$

Korelační matice podobností mezi dokumenty se vypočte dle rovnice (3), kde sloupečky matice B musí být normalizovány. Výsledná matice sim_{SVD} je symetrická, kde pro každý pár dokumentů je obsažena jejich procentuální podobnost.

$$sim_{SVD} = \left\|B\right\|^T \times \left\|B\right\| \tag{3}$$

Ačkoli se může zdát, že výpočet je v současnosti hotov, je nutné se zamyslet nad vlivem DF filtru pro redukci fází. Obrázek 2 zachycuje situaci, kde část frází je označena jako bezvýznamná a nejsou tím pádem uvažovány během výpočtu. Následný výpočet probíhá nad menší množinou, tudíž sim_{SVD} dosahuje nižšího procentuelního ohodnocení, které nemůže odpovídat realitě.



Obrázek 2. Průnik dvou množin frází

Rovnice (4) modifikuje sim_{SVD} pro získání správného ohodnocení podobnosti mezi dokumenty R a S. Podobnosti jsou váženy poměrem mezi počtem původních frází $|ph_{orig}|$ a počtem frází po redukci $|ph_{red}|$.

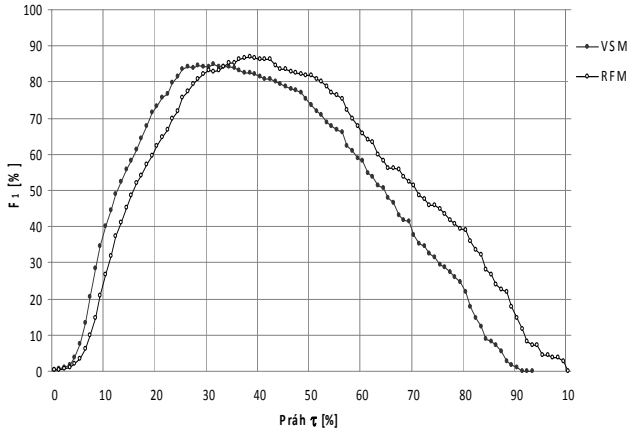
$$sim(R,S) = sim_{SVD}(R,S) \cdot \sqrt{\frac{|ph_{orig}(R)|}{|ph_{red}(R)|} \cdot \frac{|ph_{orig}(S)|}{|ph_{red}(S)|}} \tag{4}$$

4 Experimenty

Pro naše počáteční experimenty jsme shromáždili kolekci 150 plagiovaných dokumentů v českém jazyce. Tato kolekce byla vytvořena manuálně studenty. Ze standardního ČTK korpusu jsme náhodně vybrali 300 článků o politice a použili je jako základ pro vytvoření plagiovaných dokumentů. Výsledný korpus čítající 950 dokumentů jsme namíchali ze 150 plagiovaných dokumentů, 300 původních článků a 500 dalších náhodně vybraných článků o politice.

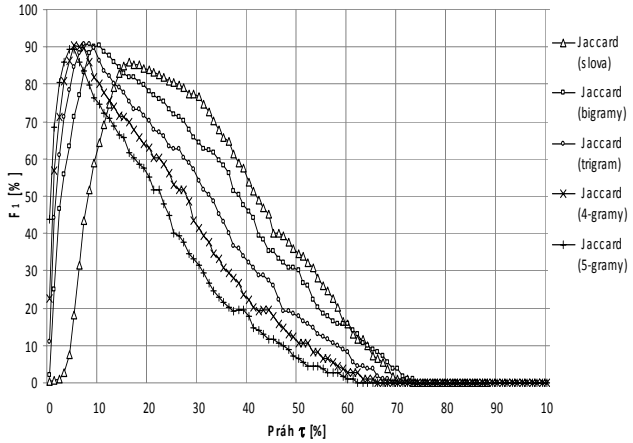
Obrázek 3 zobrazuje závislost míry F_1 na prahu τ pro metodu VSM [8] a RFM [10]. Obě křivky jsou poměrně široké, tudíž není problém stanovit správný práh τ, kterým rozhodujeme, zda je daný dokument plagiát či ne. Nicméně dosahované skóre pro F_1 je nižší než u ostatních metod.

Následující metoda využívá Jaccard-Tanimoto koeficient [6], který byl použit v systému Ferret [5]. Obrázek 4 zachycuje rozličné závislostní křivky při použití jednotlivých slov, bigramů, trigramů, 4-gramů a 5-gramů. Při porovnání s předchozím grafem jsou všechny křivky výrazně užší než VSM a RFM. Jak lze vidět z grafu, zvětšující se délka N-gramu (v našem případě fráze) snižuje práh τ potřebný pro dosažení nejlepších možných výsledků.



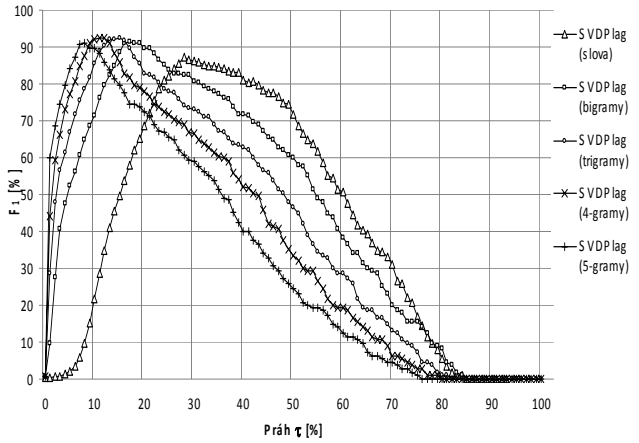
Obrázek 3. Závislost míry F_1 na prahu τ při detekci plagiátů metodami VSM a RFM

Obrázek 5 zobrazuje závislostní křivky míry F_1 na prahu τ pro naší metodu SVDPlag, která využívá LSA. Oproti systému Ferret jsou křivky širší, a tudíž je snazší určit správný práh. Naše experimentální metoda SVDPlag dosahuje mnohem lepších hodnot F_1 oproti ostatním metodám.



Obrázek 4. Závislost míry F_1 na prahu τ při detekci plagiátů Jaccard-Tanimoto koeficientem s využitím slov, bigramů, trigramů, 4-gramů a 5-gramů

Tabulka 3 je sumarizací nejlepších získaných hodnot pro F_1. Jak je vidět, naše experimentální metoda dosahuje nejlepších výsledků pro fráze skládající se ze čtyř následujících slov (4-gram). Systém Ferret dosahuje nejlepších výsledků pro fráze o třech slovech, kdy míra F_1 je 90,82% oproti 92,57% v porovnání s naším systémem. V případě trigramů a 4-gramů dosahuje naše metoda velmi obdobných výsledků. Podobně se chová i metoda založená na Jaccard-Tanimoto koeficientu pro bigramy, trigramy a 4-gramy. Přestože jsou výsledky okolo sekvence tří slov velmi podobné, doporučujeme raději volit 4-gramy. Delší fráze lépe separují nesouvisející dokumenty a redukují šum nacházející se v hlavičkách a patičkách dokumentů.



Obrázek 5. Závislost míry F_1 na prahu τ při detekci plagiátů metodou SVDPlag s využitím slov, bigramů, trigramů, 4-gramů a 5-gramů

Tabulka 3. Nejlepší dosažené výsledky pro míru F_1

Metoda	Práh τ	F_1
VSM	30%	84,97%
RFM	37%	87,03%
Jaccard (slova)	16%	85,84%
Jaccard (bigramy)	10%	90,56%
Jaccard (trigramy)	8%	90,82%
Jaccard (4-gramy)	6%	90,53%
Jaccard (5-gramy)	4%	89,36%
SVDPlag (slova)	28%	87,31%
SVDPlag (bigramy)	17%	91,36%
SVDPlag (trigramy)	15%	92,48%
SVDPlag (4-gramy)	12%	92,57%
SVDPlag (5-gramy)	8%	91,03%

5 Budoucí práce

Náš další výzkum se ubírá směrem k využití WordNet tezauru. V nejjednodušší variantě plánujeme nahrazovat slova jejich multijazykovými indexy (InterLingual Index - ILI), kde každý index označuje skupinu stejných synonym (tzv. synset). Daný index je společný pro různé jazyky, tudíž bude možné v budoucnu zahrnout i vícejazykovou podporu.

WordNet tezaurus má nicméně daleko širší uplatnění. Jednotlivé synsety jsou provázány nejrůznějšími odkazy, představující například hyperonyma, hyponyma, antonyma, odvozeniny atd. Z našeho pohledu jsou nejzajímavější hyperonymické odkazy, díky kterým lze nalézt obecnější tvary slov. Tyto odkazy vlastně tvoří stromovou strukturu slova od nejkonkrétnějšího významu po nejobecnější. Představme si slova „kočka“ a „pes“, jejichž společné hyperonymum je slovo „zvíře“. Struktura WordNetu je samozřejmě mnohem detailnější a ke slovu „zvíře“ se dostaneme až po několika posunech v hierarchii, kdy jdeme například přes slova „savec“ a „obratlovec“.

Zmíněným postupem můžeme velice zjednodušit slovní zásobu a zobecnit veškerá slova na libovolnou požadovanou úroveň. Jediným problémem, se kterým se potýkáme, je volba vhodné úrovně. Podstatná slova mívají obvykle hloubku stromu 5 až 8. Naproti tomu slovesa 1 až 3. Z tohoto důvodu je nutné zacházet s každým slovním druhem zvlášť, což je otázkou budoucnosti.

6 Závěr

V tomto článku jsme představili metodu pro detekci plagiátů využívající LSA. Tato metoda na základě asociací mezi frázemi odhaluje podobnost mezi dokumenty. Naší metodu lze klasifikovat jako Strukturální a Korpální, viz Tabulka 1.

Z našich experimentů provedených na korpusu čítajícím 950 dokumentů o politice je zřejmé, že SVDPlag překonává ostatní metody pro detekci plagiátů. Nejlepší výsledky jsme získali pro 4-gramy a 12% práh, kdy za těchto podmínek míra F_1 dosahuje 92,57%.

Jedním z klíčových faktorů naší budoucí práce bude využití WordNet tezauru pro pokročilou normalizaci slov. Dále plánujeme návrhnout složitější model textových dokumentů, který by podchytil vztahy slov obsažených uvnitř frází. Posledním cílem je rozšíření stávajícího korpusu o více plagiovaných dokumentů a přidání dalších témat.

Poděkování

Tato práce byla částečně podporována z prostředků Národního Programu Výzkumu II, projekt 2C06009 (COT-SEWing).

Reference

[1] M. Berry, S. Dumais, G. O'Brein, „Using Linear Algebra for Intelligent Information Retrieval“, *SIAM Review*, vol. 37 issue 4, pp. 573-595, Society for Industrial and Applied Mathematics, Philadelphia, USA, 1995. ISSN 0036-1445.

[2] P. Clough, „Plagiarism in natural and programming languages: An overview of current tools and technologies“, *Internal Report CS-00-05*, Department of Computer Science, University of Sheffield, 2000.

[3] T. Lancaster, F. Culwin, „Classification of plagiarism detection engines“, *E-journal ITALICS*, vol. 4 issue 2, 2005. ISSN 1473-7507.

[4] T. Landauer, P. Foltz, D. Laham, „An introduction to Latent Semantic Analysis“, *Discourse Processes*, vol. 25, pp. 259-284, 1998.

[5] P. Lane, C. Lyon, J. Malcolm, „Demonstration of the ferret plagiarism detektor“, *Proceedings of the 2nd International Plagiarism Conference*, Newcastle, 2006.

[6] C. Manning, H. Schutze, „Foundation of statistical natural language processing“, *The MIT Press*, Massachusetts Institute of Technology, Cambridge MA, 1999.

[7] H. Maurer, F. Kappe, B. Zaka, „Plagiarism – A survey“, *Journal of Universal Computer Science*, vol. 12 issue 8, pp. 1050-1084, 2006.

[8] P. Runeson, M. Alexanderson, O. Nyholm, „Detection of duplicate defect reports using natural language processing“, *Proceedings of the IEEE 29th International Conference on Software Engineering*, pp. 499-510, 2007.

[9] G. Salton, C. Buckley, „Term-Weighting Approaches in Automatic Retrieval“, *Journal of Information Processing and Management*, vol. 24 issue 5, pp. 513-523, 1988.

[10] N. Shivakumar, H. Garcia-Molina, „SCAM: A copy detection mechanism for digital documents“, *Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries*, Austin, 1995.

[11] M. Toman, R. Tesar, K. Jezek, „Influence of word normalization on text classification“, *Proceedings of the 1st International Conference on Multidisciplinary Information Sciences & Technologies*, vol. 2, pp. 354-358, Merida, Spain, 2006. ISBN 84-611-3105-3.