

Metódy detekcie plagiátorstva

Plagiarism Detection Methods

Ondrej PLOČICA – Henrieta TELEPOVSKÁ

Abstract

Tackling plagiarism in the academic environment is now devoted increased attention. Automated plagiarism detection helps teachers in a fairer assessment of students and forces students to avoid this undesirable phenomenon. There are several types of plagiarism, which can detect only review by man. Automated detection is based on experience that most cases of plagiarism are not difficult to uncover - usually a direct copy of text, with some reordering of words. This contribution is dedicated to the classification of plagiarism detection methods in text documents. The classification is based on the need of a repository of documents and the type of plagiarism that method can detect. The paper contains a short description of used methods, techniques and algorithms and deals with their exploitation at Slovak universities.

Keywords

Plagiarism detection, classification of detection methods, stylometry, string comparison, natural language processing.

Úvod

Internet obsahuje veľké množstvo jednoducho dostupných dokumentov napísaných v rôznych jazykoch, čo vytvára vhodný priestor pre jednoduchšie a rýchlejšie vyhotovenie písomných prác (záverečných, semestrálnych, a pod.) využitím plagiátorských techník. Hoci zatiaľ neexistuje hodnoverný prieskum o rozsahu plagiátorstva na Slovensku, je pravdepodobné, že tento rozsah nie je nízky.

Na odhaľovanie podvodníkov sa využíva najmä automatizovaná detekcia plagiátorstva. Na slovenských univerzitách sa zatiaľ takéto softvérové systémy používajú len ojedinele. Hlavným problémom ich nasadenia je to, že tieto systémy nie sú dostatočne vhodné pre práce písané v slovenčine. Projekt Centra vedecko-technických informácií (CVTI) na zabezpečenie plagiátorského softvéru pre vysoké školy a univerzity jasne naznačuje, že aj na Slovensku sa boju proti plagiátorstvu venuje čoraz väčšia pozornosť.

Tento príspevok podáva prehľad o metódach a technikách používaných pri detekcii plagiátorstva. Vznikol na základe našej účasti pri realizácii štúdie uskutočniteľnosti anti-plagiátorského projektu.

Klasifikácia metód detekcie plagiátorstva

Pre automatizovanú detekciu plagiátorstva bolo navrhnutých mnoho metód, techník a algoritmov. Automatizovaná detekcia je založená na skúsenostiach, že najčastejšie prípady plagiátorstva nepatria medzi ťažko odhaliteľné – zvyčajne ide o priame kópie textu (copy - paste), s prípadným poprehadzovaním slov vo vete alebo viet v odsekoch [4].

Existuje viacero klasifikácií detekčných metód. V tomto príspevku berieme za základ pre klasifikáciu metód potrebu nejakého úložiska dokumentov a typ plagiátorstva, pre detekciu ktorého sú vhodné. Prehľad metód pre jednotlivé typy plagiátorstva je uvedený v tabuľke Tab.1. Jednotlivé riadky tabuľky zodpovedajú typom plagiátorstva a stĺpce zodpovedajú typu používaného úložiska.

	Bez úložiska	Vlastné úložisko	Verejné úložisko (Internet)
Jednoduchá kópia dokumentu alebo jeho časti	Analýza štýlu	Porovnávanie obsahu	„Online“ vyhľadávanie
Modifikovaná kópia dokumentu alebo jeho časti		Lokálna analýza podobnosti	
Parafrázovaná kópia dokumentu alebo jeho časti		Sémantická analýza	
Preklad z iného jazyka		Skrytá sémantická analýza	
		Medzi-jazyková analýza	Spätný preklad a „online“ vyhľadávanie

Tab.1 Klasifikácia detekčných metód

Detekčné metódy sú zatiaľ úspešné najmä pre prvé dva typy plagiátorstva, keď v preverovanom dokumente existujú presné kópie alebo jednoducho modifikované (vynechanie alebo doplnenie slova, zmena poradia slov, zmena poradia viet) časti nejakého dokumentu. Metódy pre ďalšie dva typy plagiátorstva sú zatiaľ v štádiu experimentovania.

Charakteristika metód detekcie plagiátorstva

- Analýza štýlu (štýlometria)

Táto metóda je založená na rozdelení dokumentu na jeho prirodzené časti (kapitoly, odseky, vety) a analýze určitých charakteristík štýlu písania textu. Vychádza z predpokladu, že každý autor má vlastný štýl písania, teda spôsob konštrukcie viet a používaný slovník.

Daný text sa porovnáva s typickým štýlom písania autora, získaným z jeho predošlých textov alebo sa porovnáva štýlu daného odseku s celkovým štýlom dokumentu. Táto metóda je na rozdiel od ostatných metód schopná odhaliť plagiátorstvo aj bez porovnávania s inými dokumentmi. Medzi základné štýlometrické charakteristiky patria podľa [7] textová štatistika na úrovni znakov, syntaktická štatistika na úrovni viet, štatistika používania vetných členov, štatistika používania špeciálnych slov, štruktúrna štatistika dokumentu.

- Lokálna analýza podobnosti na základe odtlačkov

Táto metóda je vhodná pre analýzu podobnosti dokumentov. Pod odtlačkom dokumentu sa rozumie množina zakódovaných reťazcov slov (úsekov) vybraných z dokumentu, ktorá slúži na jednoznačnú identifikáciu dokumentu. Analýza podobnosti dokumentov A a B na základe ich odtlačkov $h(A)$ a $h(B)$ je mierou veľkosti prieniku odtlačkov [3]:

$$j_{LOC}(A, B) = \frac{|h(A) \cap h(B)|}{|h(A) \cup h(B)|}$$

Odtlačok $h(d)$ dokumentu možno považovať za špeciálny model dokumentu d . Výhodou tohto modelu je kompaktnosť (vzhľadom na veľkosť dokumentu) a vernosť (vzhľadom na analýzu podobnosti).

- **Techniky porovnávania obsahu**

Karp-Rabinov algoritmus vyhľadávania je založený na využití hašovania na hľadanie podreťazca v texte [4]. Pri rozdelení textu na úseky sa realizuje dlaždicový prístup, kde výsledkom sú neprekrývajúce sa podreťazce (od iniciálnej dĺžky po maximálnu dĺžku).

Princíp metódy úsekovania [5] spočíva v tom, že sa porovnávajú iba časti textu (úseky) alebo ich odtlačky. Najčastejšie sa používajú prekrývajúce sa úseky slov (n -gramy).

Cieľom metódy Winnowing [6] je efektívny algoritmus pre výber hašovaných hodnôt z množiny všetkých kontinuálnych podreťazcov dĺžky n (n -gramy) tak, aby bola zaručená čiastočná detekcia každej dostatočne dlhej plagiarizovanej časti textu.

- **Medzi-jazyková analýza**

Tento prístup je nutný v prípade, že je potrebné detekovať plagiátorstvo z dokumentov, ktoré sú napísané v inom jazyku. Medzi-jazyková analýza použitím štatistického modelu [1] je založená na troch hlavných krokoch. Najskôr sa pre daný dokument a referenčný korpus v inom jazyku, vyberie podmnožina kandidátskych dokumentov z korpusu, ktoré môžu byť zdrojom plagiarizovaných fragmentov daného dokumentu. Potom sa vykoná sémantická analýza medzi sekciami dokumentu a kandidátskych dokumentov. Nakoniec sú podobné sekcie analyzované, aby sa odfiltrovali prípady správnej citácie. Nevýhodou uvedeného prístupu je potreba vytvorenia referenčného korpusu a nutnosť prekladu sekcie princípom slovo za slovom.

- **„Online“ vyhľadávanie**

Táto metóda je založená na myšlienke, že ak nejaké dokumenty obsahujú niekoľko rovnakých dostatočne dlhých fragmentov textu, tak pravdepodobne ide o plagiátorstvo. Vhodne zvolené podreťazce slov dokumentu sa vyhľadávajú na Internete pomocou verejne dostupných služieb ako je Google.

- **Sémantická analýza**

Využitie sémantickej analýzy pre zlepšenie detekcie podobnosti je opísaný v práci [2]. Ich systém vyberá kľúčové slová (podstatné mená, slovesá, prídavné mená) reprezentujúce štruktúrne charakteristiky dokumentov. Synonymické rady pre kľúčové slová sa hľadajú v systéme WordNet, pričom každý rad má priradenú číselnú hodnotu. Všetky kľúčové slová, ktoré sa vyskytujú v štruktúrnom charakteristickom strome dokumentu nesú svoju hodnotu. Potom pri porovnávaní zdrojov, binárne porovnanie týchto hodnôt určuje, či sú to synonymá. Systém najskôr vykonáva porovnávací algoritmus na strome celého dokumentu. Ak sú podobnosti vyššie ako istá hranica, potom sa spustí porovnávanie na úrovni viet. Tak je systém schopný detekovať podobnosť aj s malými sémantickými modifikáciami.

- **Skrytá sémantická analýza**

Je to štatistická technika pre výber a reprezentáciu podobnosti významu slov a pasáží textu na základe veľkého množstva textu. Matica slov a segmentov, ku ktorým sa vzťahujú sa využíva na vytvorenie sémantického doménového priestoru. Text, ktorý sa má porovnať s týmto doménovým priestorom, sa tiež reprezentuje v tvare dokumentového vektora. Ak je

tento vektor podobný modelovému vektoru v doméne, potom má dokument vyšší stupeň podobnosti. Aj napriek možnosti redukovať slovnú a kontextovú maticu, rozmery matice sú stále veľké a analýza vektorového priestoru je výpočtovo náročná.

Použitie detekčných metód pre slovenský jazyk

Detekčné metódy boli vyvíjané a testované predovšetkým pre texty napísané v anglickom jazyku. Sú použiteľné aj pre texty napísané v slovenskom jazyky, hlavne pre odhaľovanie presných a minimálne modifikovaných kópií iných dokumentov alebo ich častí.

Pre odhaľovanie sémanticky modifikovaných kópií je potrebné tieto metódy prispôbiť slovenčine, ktorá má komplikovanejšiu gramatickú štruktúru. Riešením je použitie prístupov spracovania prirodzeného jazyka a využitie budovaného jazykového korpusu pre slovenčinu.

Slovenských textov vhodných pre plagiátorstvo je na Internete pomerne málo, takže veľmi dôležité je odhaľovanie preložených kópií textov. Učiteľ dokáže odhaliť zlý doslovný preklad z iného jazyka, automatizovaná detekcia pomocou softvéru nie je zatiaľ úspešne vyriešená. Jednoduchým riešením by mohlo byť využitie verejne dostupných prekladových služieb na preklad slovenských textov do iných jazykov (predovšetkým angličtiny a češtiny) a potom využitie „online“ vyhľadávania.

Vzhľadom na fakt, že sa väčšinou jedná o preverovanie písomných prác jednotlivých autorov, vhodným riešením je prispôbenie metódy analýzy štýlu, najmä výber štylometrických charakteristík vhodných pre slovenčinu.

Záver

Tento sa príspevok stručne charakterizuje metódy detekcie plagiátorstva a naznačuje možnosti ich prispôbenia pre texty napísané v slovenskom jazyky. Je dôležité si však uvedomiť, že tieto metódy len upozorňujú na podozrivé znaky, že môže ísť o plagiátorstvo. Problémy, ktoré systém detekcie plagiátov odhalí musia byť analyzované doménovými odborníkmi.

Abstrakt

Boju proti plagiátorstvu v akademickom prostredí je v súčasnosti venovaná zvýšená pozornosť. Automatizovaná detekcia plagiátorstva pomáha učiteľom pri spravodlivejšom hodnotení študentov a núti študentov vyhýbať sa tomuto nežiaducemu javu. Existuje viacero typov plagiátorstva, ktoré zatiaľ dokáže objaviť len preskúmanie človekom. Automatizovaná detekcia je založená na skúsenostiach, že najčastejšie prípady plagiátorstva nepatria medzi ťažko odhaliteľné – zvyčajne ide o priame kópie textu s prípadným jeho poprehadzovaním. Tento príspevok je venovaný klasifikácii metód detekcie plagiátorstva v textových dokumentoch. Základom pre klasifikáciu je potreba nejakého úložiska dokumentov a typ plagiátorstva, pre detekciu ktorého sú vhodné. Príspevok obsahuje krátky opis používaných metód, techník a algoritmov a zaoberá sa možnosťami ich využitia na slovenských univerzitách.

Kľúčové slová

Detekcia plagiátorstva, klasifikácia metód detekcie, štylometria, porovnávanie reťazcov, spracovanie prirodzeného jazyka

Literatúra

- [1] BARRON-CEDENO, A. - ROSSO, P. - PINTO, D. – JUAN, A. 2008. On cross-lingual plagiarism analysis using a statistical model. In: Proceedings of PAN-08.
- [2] IYER, P. - SINGH, A. 2005. Document Similarity Analysis for a Plagiarism Detection Systéme. 2nd Indian International Conference on Artificial Intelligence (IICAI –2005). pp. 2534-2544
- [3] MONOSTORI, K.- FINKEL, R. - ZASLAVSKY, A. - HODÁSZ, G. – PATAKI, M. 2002. Comparison of overlap detection techniques. In LNCS, volume 2329.
- [4] MOZGOVOY, M. 2007. Enhancing Computer-Aided Plagiarism Detection, Dissertatio., University of Joensuu.
- [5] PATAKI, M. 2003. Plagiarism Detection and Document Chunking Methods. The Twelfth International World Wide Web Conference. Budapest.
- [6] SCHLEIMER, S. - WILKERSON, D. S. – SUKEN, A. 2003. Winnowing: local algorithms for document fingerprinting. In Proceedings of ACM SIGMOD Int. Conf. on Management of Data.
- [7] STEIN, B. - ZU EISSEN, S. M. 2006. Near Similarity Search and Plagiarism Analysis. In Proceedings of 29th Annual Conference of the GfKI. Springer.

Kontakt

Ing. Ondrej Pločica , KPI FEI TU Košice, Letná 9, 042 00 Košice, 055/6025150

Ondrej.Plocica@tuke.sk

Ing. Henrieta Telepovská, PhD., KPI FEI TU Košice, Letná 9, 042 00 Košice, 055/6022568

Henrieta.Telepovska@tuke.sk

Recenzent: Mgr. Ján Skalka, PhD., UKF Nitra