Detekcia plagiátorstva

Z Wikipédie, voľnej encyklopédie

Prejsť na navigáciuPridať do vyhľadávania

Tento článok môže vyžadovať vyčistenie, aby spĺňalo normy kvality Wikipédie. Nebol špecifikovaný dôvod vyčistenia. Pomôžte vylepšiť tento článok, ak môžete. (December 2010) (zistite, ako a kedy odstrániť túto správu šablóny)

Detekcia plagiátorstva je proces lokalizácie prípadov plagiátorstva v rámci práce alebo dokumentu. Široké používanie počítačov a príchod internetu umožnili plagiarizovať prácu iných. Väčšina prípadov plagiátorstva sa nachádza v akademickej oblasti, kde dokumenty sú zvyčajne eseje alebo správy. Plagiátorstvo však možno nájsť prakticky vo všetkých oblastiach vrátane románov, vedeckých prác, návrhov umenia a zdrojového kódu.

Detekcia plagiátorstva môže byť buď manuálna, alebo pomocou softvéru. Ručná detekcia si vyžaduje značné úsilie a vynikajúcu pamäť a je nepraktická v prípadoch, keď je potrebné porovnať príliš veľa dokumentov alebo nie sú k dispozícii porovnávacie dokumenty. Detekcia pomocou softvéru umožňuje porovnávať obrovské zbierky dokumentov, čím je úspešná detekcia oveľa pravdepodobnejšia.

Prax plagiátorstva pomocou použitia dostatočných slovných substitúcií na vyhnutie sa detekčnému softvéru je známa ako rogeting. [1]

obsah

• 1Systémová detekcia

o 1.1In textové dokumenty

 1.1.1 Účinnosť v oblasti vysokoškolského vzdelávania

 1.1.2Approaches

 1.1.2.1Fingerprinting

 1.1.2.2Rozhodnutie reťazca

 1.1.2.3Buľa slov

 1.1.2.4 Analýza citácie

 1.1.2.5Stylometry

 1.1.3Performance

 1.1.4Software

o 1.2v zdrojovom kóde

• 2Pozrite sa tiež

• 3References

• 4Literature

• 5Externé odkazy

Detekcia pomocou softvéru [upraviť]

Počítačová detekcia plagiátov (CaPD) je úloha získavania informácií (IR) podporovaná špecializovanými IR systémami, označovanými ako systémy detekcie plagiátov (PDS).

V textových dokumentoch [upraviť]

Systémy detekcie textového plagiátorstva implementujú jeden z dvoch všeobecných prístupov detekcie, z ktorých jeden je vonkajší, druhý je vnútorný. [2] Externé detekčné systémy porovnávajú podozrivý dokument s referenčnou zbierkou, čo je súbor dokumentov, o ktorých sa predpokladá, že sú pravé. [3] Na základe vybraného modelu dokumentu a preddefinovaných kritérií podobnosti je úlohou detekcie získať všetky dokumenty, ktoré obsahujú text, ktorý je podobný stupňu nad zvolenou prahovou hodnotou, aby obsahoval text v podozrivom dokumente. [4] Vlastné PDS analyzuje iba text, ktorý sa má vyhodnotiť, bez porovnania s externými dokumentmi. Cieľom tohto prístupu je rozpoznať zmeny v jedinečnom štýle písania autora ako indikátor potenciálneho plagiátorstva. [5] PDS nie sú schopné spoľahlivo identifikovať plagiátorstvo bez ľudského úsudku. Podobnosti sa vypočítavajú pomocou vopred definovaných modelov dokumentov a môžu predstavovať falošné pozitíva. [6] [7] [8] [9] [10]

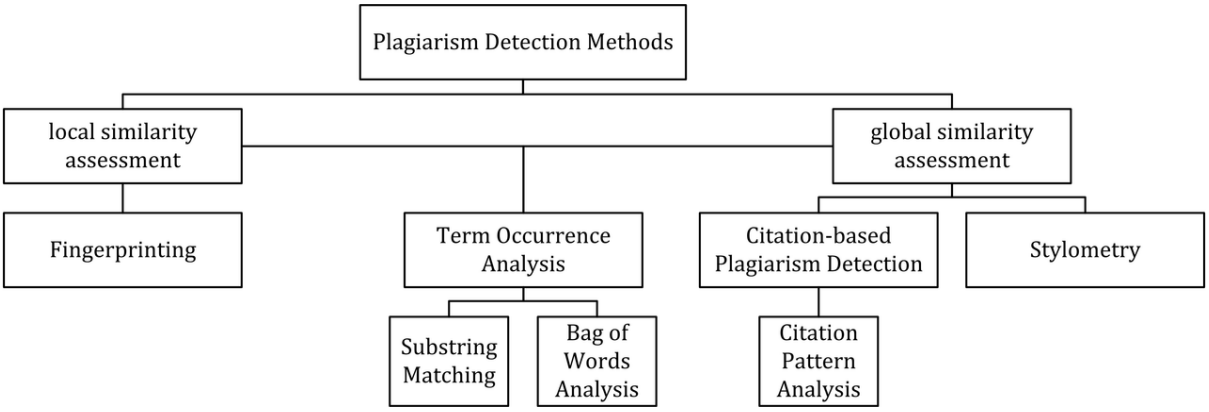
Účinnosť v oblasti vysokoškolského vzdelávania [upraviť]

Táto časť sa z veľkej časti alebo úplne opiera o jediný zdroj. Príslušnú diskusiu možno nájsť na stránke diskusií. Pomôžte vylepšiť tento článok vložením citácií na ďalšie zdroje. (December 2017)

Bola vykonaná štúdia na overenie účinnosti softvéru na detekciu plagiátov v prostredí vysokoškolského vzdelávania. Jedna časť štúdie priradila jednej skupine študentov, aby napísali príspevok. Títo študenti sa prvýkrát vzdelávali o plagiátstve a informovali, že ich práca má prebiehať pomocou systému detekcie plagiátorstva. Druhá skupina študentov bola pridelená na písanie príspevku bez akýchkoľvek informácií o plagiátstve. Vedci očakávali, že nájdu nižšie sadzby v skupine jedna, ale zistili približne rovnaké miery plagiátorstva v oboch skupinách. [11]

Prístupy [upraviť]

Nasledujúci obrázok predstavuje klasifikáciu všetkých prístupov detekcie, ktoré sa v súčasnosti používajú na detekciu plagiátorstva pomocou počítača. Tieto prístupy sú charakterizované typom posúdenia podobnosti, ktoré vykonávajú: globálne alebo lokálne. Globálne prístupy na posúdenie podobnosti využívajú charakteristiky prevzaté z väčších častí textu alebo dokumentu ako celku na výpočet podobnosti, zatiaľ čo lokálne metódy skúmajú predvolené textové segmenty iba ako vstup.

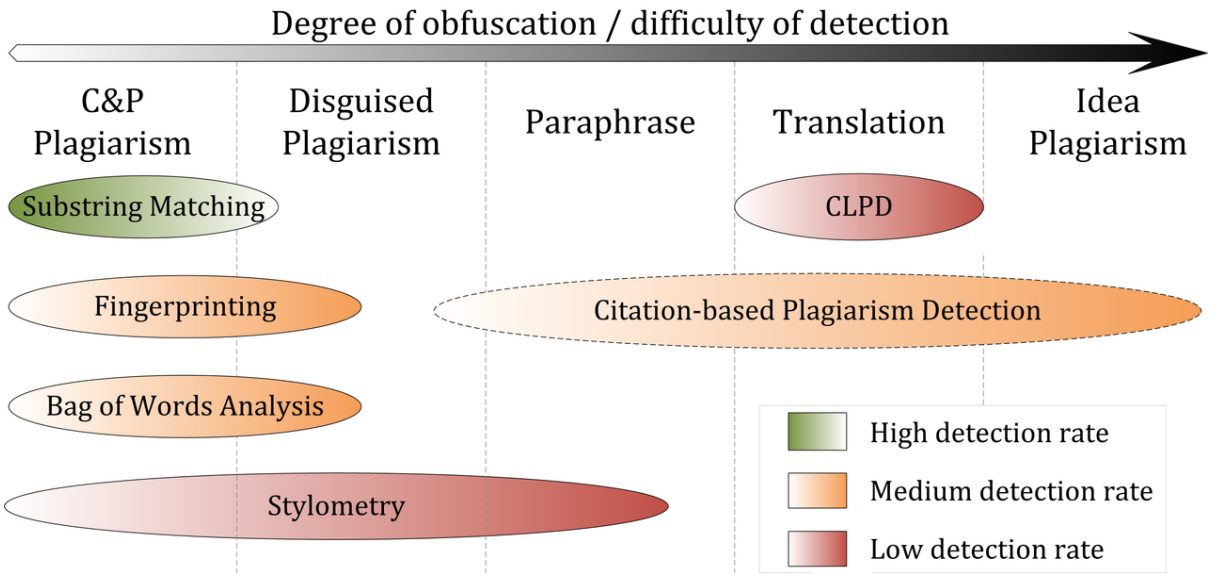


Klasifikácia metód detekcie plagiátorstva pomocou počítača

Snímanie odtlačkov prstov [upraviť]

Odtlačky prstov sú v súčasnosti najpoužívanejším prístupom k detekcii plagiátorstva. Táto metóda vytvára reprezentatívne záznamy dokumentov výberom množiny viacerých podčiarkov (n-gramov) z nich. Sady predstavujú odtlačky prstov a ich prvky sa nazývajú markanty. [12] [13] Podozrivý dokument je skontrolovaný pre plagiátorstvo pomocou výpočtu jeho odtlačku prstov a dotazovania markantov s predkompilovaným indexom odtlačkov prstov pre všetky dokumenty referenčnej zbierky. Zhodujúce sa záznamy s ostatnými dokumentmi naznačujú zdieľané textové segmenty a naznačujú možný plagiát, ak prekročia zvolený prah podobnosti. [14] Výpočtové zdroje a čas sú obmedzujúce faktory pre odtlačky prstov, čo je dôvod, prečo táto metóda zvyčajne porovnáva len podmnožinu markantov, aby urýchlila výpočet a umožnila kontroly vo veľmi veľkej zbierke, ako je internet. [12] String matching [edit] String matching je prevládajúci prístup používaný v informatike. Pri použití na problém detekcie plagiátov sa porovnávajú dokumenty na doslovné prekrývanie textu. Na riešenie tejto úlohy boli navrhnuté mnohé metódy, z ktorých niektoré boli prispôsobené externej detekcii plagiátorstva. Kontrola podozrivého dokumentu v tomto nastavení vyžaduje výpočet a ukladanie efektívne porovnateľných reprezentácií pre všetky dokumenty v referenčnej zbierke, aby ste ich porovnali párovým spôsobom. Vo všeobecnosti boli pre túto úlohu použité prípony modelov dokumentov, ako napríklad prípony stromov alebo prípony. Napriek tomu zostava podreťaze zostáva výpočtovo nákladná, čo z neho robí nerealizovateľné riešenie pre kontrolu veľkých zbierok dokumentov. [15] [16] [17] Bag of words [upraviť] Bag of words analysis predstavuje prijatie vyhľadávania vektorových priestorov, tradičný IR koncept, do oblasti detekcie plagiátorstva. Dokumenty sú reprezentované ako jeden alebo viac vektorov, napr. pre rôzne časti dokumentu, ktoré sa používajú na výpočty podobnosti párov. Výpočet podobnosti sa potom môže opierať o tradičnú mieru podobnosti kosín alebo o sofistikovanejšie podobnostné opatrenia. [18] [19] [20] Citation analysis [edit] Citácia založená plagiátorová detekcia (CbPD) [21] je jediný prístup k detekcii plagiátorstva, ktorý sa nespolieha na podobnosť textu. [22] CbPD skúma citačné a referenčné informácie v textoch na identifikáciu podobných vzorov v citačných sekvenciách. Ako taký je tento prístup vhodný pre vedecké texty alebo iné akademické dokumenty, ktoré obsahujú citácie. Citačná analýza na odhaľovanie plagiátov je pomerne mladý koncept. Nie je prijatý komerčným softvérom, ale existuje prvý prototyp systému detekcie plagiátov založeného na citáciách [23]. Podobná poradia a blízkosť citácií v skúmaných dokumentoch sú hlavnými kritériami používanými na výpočet podobností citačných vzorov. Citačné modely predstavujú subsekvencie, ktoré neobsahujú výlučne citácie zdieľané porovnávanými dokumentmi. [22] [24] Faktory, vrátane absolútneho čísla alebo relatívneho podielu zdieľaných citácií v schéme, ako aj pravdepodobnosť, že sa citácie vyskytujú v dokumente, sa tiež považujú za kvantifikáciu stupňa podobnosti vzory [22] [24] [25] [26] Stylometria

[editovať] Stylometria zahŕňa štatistické metódy na kvantifikáciu autorského jedinečného štýlu písania [27] [28] a používa sa hlavne na autorské pripisovanie alebo vnútorné CaPD. Vytváraním a porovnávaním stylometrických modelov pre rôzne textové segmenty sa dajú odhaliť pasáže, ktoré sú štylisticky odlišné od ostatných, a teda potenciálne plagiarizované. [5] Výkon [upraviť] Porovnávacie hodnotenia plagiátorových detekčných systémov [3] [29] 31] [32] [33] naznačujú, že ich výkon závisí od typu plagiátorstva (pozri obrázok). Okrem analýzy citačných vzorov sa všetky detekčné prístupy spoliehajú na textovú podobnosť. Je preto príznačné, že presnosť detekcie znižuje počet prípadov plagiátov, ktoré sú zamaskované.



Detekčný výkon prístupov CaPD v závislosti od typu plagiátorstva

Zreteľné kópie, napríklad plagiátorstvo kopíruj a vlož (c & p) alebo mierne zamaskované prípady plagiátorstva, môžu byť s vysokou presnosťou zistené súčasným externým PDS, ak je zdroj prístupný softvéru. Najmä procesy spájania podreťazcov dosahujú dobrú výkonnosť pre plagiátorstvo c & p, pretože bežne používajú bezstratové modely dokumentov, napríklad suffix trees. Výkonnosť systémov, ktoré používajú fingerprinty dokumentov alebo analýzy typu bag of words pri detekcii kópií, závisí od straty informácií spôsobených používaným modelom dokumentu. Aplikovaním flexibilnej stratégie združovania slov a selekcie sú lepšie schopné detekovať mierne formy maskovaného plagiátorstva v porovnaní s procedúrami porovnávania podreťazcov.

Detekcia intrinsic plagiátorstva pomocou štýlometrie môže do určitej miery prekonať hranice textovej podobnosti porovnaním lingvistickej podobnosti. Vzhľadom na to, že štylistické rozdiely medzi plagiarizovanými a pôvodnými segmentmi sú významné a môžu sa spoľahlivo identifikovať, môže štýlometria pomôcť identifikovať maskované a parafrázované plagiáty. Štýlometrické porovnania pravdepodobne zlyhajú v prípadoch, keď sú segmenty výrazne parafrázované až do bodu, v ktorom sa viac podobajú štýlu osobného písania plagiátora alebo ak bol text zostavený viacerými autormi. Výsledky medzinárodných súťaží na detekciu plagiátorstva, ktoré sa konali v rokoch 2009, 2010 a 2011 [3] [32] [33], ako aj pokusy vykonané Steinom [34] naznačujú, že štýlometrická analýza funguje spoľahlivo iba na dĺžky dokumentov niekoľko tisíc alebo desiatok tisíc slov, čo obmedzuje použiteľnosť metódy na pre účely CaPD.

Rastúce množstvo výskumov sa uskutočňuje na metódach a systémoch schopných odhaliť translačné plagiátorstvo. V súčasnosti sa detekcia plagiátorstva v rámci viacerých jazykov (CLPD) nepovažuje za vyspelú technológiu [35] a príslušné systémy nedokázali v praxi dosiahnuť uspokojivé výsledky detekcie. [31]

Zisťovanie plagiátorstva založené na citáciách pomocou analýzy citačných vzorov je schopné identifikovať silnejšie parafrázy a preklady s vyššou mierou úspešnosti v porovnaní s inými detekčnými prístupmi, pretože je nezávislé od textových charakteristík. [22] [25] Avšak vzhľadom na to, že analýza citačných vzorov závisí od dostupnosti dostatku citačných informácií, je obmedzená na akademické texty. Je slabšia v porovnaní s textovým prístupom pri zisťovaní kratších plagiovaných pasáží, ktoré sú typické pre prípad plagiátorstva typu kopírovania a vkladania alebo pretrepávania a vkladania; druhá sa týka zmiešania mierne zmenených fragmentov z rôznych zdrojov. [36]

Softvér

Návrh softvéru na detekciu plagiátov na použitie s textovými dokumentmi sa vyznačuje viacerými faktormi:

Faktor	Popis a alternatívy
Rozsah vyhľadávania	Internet za použitia vyhľadávačov / Inštitucionálne databázy / Lokálne, systémovo-špecifické databázy
Čas analyzovania	Čas medzi odovzdaním dokumentu do systému a dodaním výsledku
Kapacita dokumentov / Dávkové spracovanie	Počet dokumentov, ktorý dokáže systém spracovať za jednotku času
Kontrola intenzity	Ako často a pre ktoré typy fragmentov dokumentu (odseky, vety, sekvencie slov pevnej dĺžky) systém vyhľadáva externé zdroje, napríklad vyhľadávače.
Typ porovnávacieho algoritmu	Algoritmy definujúce spôsob ako systém porovnáva dokumenty voči sebe
Precision and Recall	Počet dokumentov správne označených ako plagiáty v porovnaní s celkovým počtom dokumentov označených ako plagiáty a celkový počet dokumentov, ktoré boli skutočne plagiarizované. Precision znamená, že bolo zistených niekoľko falošných plagiátov a Recall znamená, že niekoľko falošných negatívov zostalo nezistených.

Väčšina rozsiahlych systémov na detekciu plagiátov používa veľké interné databázy (okrem iných zdrojov), ktoré rastú s každým dodatočným dokumentom predloženým na analýzu. Toto však niektorí považujú za porušenie autorských práv študentov.

V zdrojovom kóde

Plagiátorstvo v zdrojovom kóde počítača je tiež časté a vyžaduje si iné nástroje než tie, ktoré sa používajú na porovnávanie textu v dokumente. Významný výskum bol venovaný plagiátorstvu zdrojového kódu v akademickom prostredí. [37]

Výrazným aspektom plagiátorstva zdrojového kódu je, že neexistujú databázy statí (paper mills, essay mills), ktoré využívajú pri tradičné plagiátorstve. Keďže väčšina programových úloh očakáva od študentov, aby napísali programy s veľmi špecifickými požiadavkami, je veľmi ťažké nájsť existujúce programy, ktoré ich spĺňajú. Keďže integrácia externého kódu je často ťažšia ako písanie od nuly, väčšina plagiarizujúcich študentov to získava od svojich rovesníkov.

Podľa Roya a Cordyho [38] môžu byť algoritmy detekcie podobnosti zdrojového kódu klasifikované na základe

- Reťazce – hľadajú sa presné zhody segmentov, napr. pätíc slov. Je to rýchle, ale môže byť pomýlené premenovaním identifikátorov.
- Tokeny – podobne ako s reťazcami, ale pomocou lexer-u (robí lexikálnu analýzu) pre prevod programu do tokenov. Vylučujú sa medzery, komentáre a názvy identifikátorov, čo robí systém robustnejším pri jednoduchom nahrádzaní textu. Väčšina akademických systémov na detekciu plagiátov funguje na tejto úrovni pomocou rôznych algoritmov na meranie podobnosti medzi sekvenciami tokenov.
- Parse trees (zobrazenie štruktúry vety alebo reťazca pomocou grafu stromu) - robia a porovnávajú stromy. To umožňuje zistiť podobnosti na vyššej úrovni. Napríklad porovnávanie stromov môže normalizovať podmienené výkazy a detegovať ekvivalentné konštrukcie, ktoré sa navzájom podobajú.
- Grafy závislostí v programe (PDG – program dependency graphs) - PDG zachytáva skutočný tok riadenia v programe a umožňuje nájsť ekvivalenty vyššieho druhu, pri vyšších nákladoch v zložitosti a v čase výpočtu.
- Metrika - metriky zachytia "skóre" segmentov programu podľa určitých kritérií; napríklad "počet slučiek a podmienok" alebo "počet použitých premenných". Metriky sú jednoduché na výpočet a môžu sa rýchlo porovnávať, ale môžu tiež viesť k falošným pozitívnym výsledkom: dva fragmenty s rovnakými výsledkami na množine metrík môžu robiť úplne iné veci.
- Hybridné prístupy - napr. parse trees a suffix trees stromov môžu kombinovať detekčnú schopnosť parse stromov s rýchlosťou poskytovanou suffix stromami, typ dátovej štruktúry zodpovedajúcej reťazcom.

Predchádzajúca klasifikácia bola vyvinutá pre refaktorovanie kódu programu a nie pre akademickú detekciu plagiátorstva (dôležitým cieľom refaktorovania je vyhnúť sa duplicitnému kódu, ktorý sa označuje ako klon kódu v literatúre). Uvedené prístupy sú účinné proti rôznym úrovniam podobnosti; nízka podobnosť sa týka rovnakého textu, zatiaľ čo podobnosť na vysokej úrovni môže byť spôsobená podobnými špecifikáciami. V akademickom prostredí, keď sa očakáva, že všetci študenti budú kódovať rovnaké špecifikácie, je úplne očakávaný funkčne ekvivalentný kód (s vysokou úrovňou podobnosti) a iba nízka podobnosť sa považuje za dôkaz podvádzania.