Digitalni tekstni plagijati

Tedo Vrbanec
Učiteljski fakultet Zagreb, Odsjek u Čakovcu
tedo.vrbanec@gmail.com

Sažetak. *Rad daje pregled domene plagiranja digitalnih tekstnih dokumenata. Opisuje porijeklo pojma plagijata, daje prikaz definicija te objašnjava plagijatu srodne pojmove. Ukazuje na širinu domene plagiranja te se potom ograničava na podomenu plagiranja digitalnih tekstnih dokumenata za koje predlaže taksonomiju prema više kriterija: prema porijeklu i namjeni, prema tehničkoj provedbi plagiranja, prema posljedicama plagiranja, prema složenosti otkrivanja i prema (više)jezičnom porijeklu. Rad prikazuje vrste i metode plagiranja, tipove i kategorije plagijata, pristupe i faze otkrivanja plagiranja. Potom opisuje klasifikaciju metoda i algoritama otkrivanja plagijata te nudi prijedlog metodologije otkrivanja plagijata.*

Ključne riječi: akademski plagijati, metode plagiranja, metodologija otkrivanja plagijata, taksonomija plagijata, tekstni plagijati.

I. Uvod

Vjerojatnost da dvije osobe bez međusobnog utjecaja napišu identičan ne-trivijalan tekst ili naprave identično ne-trivijalno djelo je vrlo mala, a neka istraživanja poput [1, str. 133–134] dokazuju da je to i nemoguće. U slučaju preuzimanju tuđih misli, riječi ili djela, bez jasne naznake izvora, postupak se naziva plagiranje, a proizvod plagijat. Plagijat (lat. *plagiare* = ukrasti; lat. *plagere* = oteti; lat. *plagiarius* = otmičar) je djelomično ili u cijelosti preuzeti tuđi intelektualni ili umjetnički rad, bez jasne naznake tuđeg autorstva. Nije svako korištenje tuđeg djela nedozvoljeno i neetično, ali postoje norme na koji se način takvo korištenje treba i naznačiti [2, str. 1]. Stoga je u današnje internetsko doba kada su dokumenti i informacije izuzetno lako dostupni, problematika plagiranja u akademskoj zajednici te u istraživačkim organizacijama od velikog značaja i interesa [3, str. 1].

U svim je državama plagiranje zakonom zabranjeno [3, str. 1]. Prema hrvatskom Zakonu o autorskom pravu i srodnim pravima [4, Par 2]: "autorsko pravo pripada, po svojoj naravi, fizičkoj osobi koja stvori autorsko djelo" i nastavlja [4, Par 3]: "autorsko djelo je originalna intelektualna tvorevina iz književnoga, znanstvenog i umjetničkog područja koja ima individualni karakter, bez obzira na način i oblik izražavanja, vrstu, vrijednost ili namjenu".

Postoji visoko suglasje o definiciji pojma plagijata. S vrlo malim varijacijama izričaja, većina izvora (poput [3, str. 99], [5, str. 1], [6, str. 2], [7, str. iii], [8, str. 565], [9, str. 1]) koriste definiciju koja se nalazi i u Merriam-Webster rječniku [10], a koja definira plagijat kao djelo nastalo korištenjem tuđih riječi ili ideja, bez davanja zasluga izvornom autoru. Encyclopedia Britannica definira plagijat [11] kao čin uzimanja spisa druge osobe te njegovu predaju kao vlastitog. Rječnici Cambridgea [12] i Oxforda [13] te Sveučilište u Oxfordu [14] definiraju plagijat kao korištenje ideja ili rada drugih osoba pretvarajući se da su vlastita. Meuschke and Gipp definiraju akademski plagijat [15] kao preuzimanje tuđih ideja ili izričaja bez davanja dužnog priznanja izvornim autorima ili izvorima prema akademskim principima. Plagiarism.org [16] smatra da su plagiranje i plagijat prevara te da uključuju krađu tuđeg te potom laganje o toj krađi. Prema međunarodnoj organizaciji Action Plagiarius [17], plagijat je imitacija proizvoda u svrhu gospodarskog korištenja.

Uz plagijat se veže nekoliko srodnih pojmova [17]: krivotvorina, piratiziranje dizajna, piratiziranje branda i replika. Krivotvorina ili imitacija je proizvod koji krivotvoritelj potencijalnom kupcu predstavlja kao original, dakle kupac se nastoji uvjeriti da je riječ o originalnom proizvodu. Krivotvorenje je kazneno djelo. Piratiziranje dizajna je marketinški koncept kojim se proizvođači koriste kako bi u kratkom roku iskoristili veliki interes kupaca za neki proizvod na način da dizajnom svojeg proizvoda jako podsjeća na neku poznatu marku. Piratiziranje branda je situacija kada proizvođač ne može zaštititi svoje ime i proizvode u nekoj zemlji, jer je to prethodno učinio netko drugi s kime je nužno postići financijski sporazum. Replika je nova izrada nekog proizvoda od izvornog proizvođača ili vlasnika prava.

Plagijati se često koriste u poslovnom svijetu kako bi se:

- bez većeg ulaganja došlo do novijih proizvoda, dok još imaju znatnu profitabilnost,
- na nezakonit način iskoristio tuđi brand ili
- okoristilo tuđim dizajnom ili idejom.

Ekonomske posljedice industrijskih plagijata su teške, a neke procjene govore [17] da 10% svjetske trgovine čine krivotvorine i plagijati, te da se godišnje zbog toga izgubi 200-300 milijardi € i 200 tisuća radnih mjesta.

Plagijatima se često koriste učenici i studenti tijekom svoga obrazovanja za izradu programskih rješenja, zadaća ili seminara. Razlozi mogu biti:

nedostatak vremena, nedostatak sposobnosti, lijenost ili neznanje da je takvo postupanje nemoralno, nedopušteno i kažnjivo.

Jednom otkriveni, plagijati u pravilu nose značajne i dalekosežne negativne posljedice za plagijatore, u obliku javne sramote, financijskih kazni, izbacivanja iz obrazovnih institucija, poništenja diploma, gubitka radnog mjesta, sudskih postupaka i osuda, negativnim ili smanjenim ocjenama te otežanom polaganju ispita. Ovaj rad se fokusira na plagijate digitalnih akademskih radova tekstnog formata (akademski digitalni tekstni plagijati, DT plagijati), pa se takvi u nastavku podrazumijevaju, ukoliko nije drugačije eksplicitno napisano. Takvi su naime, najčešći objekt plagiranja tijekom obrazovanja te u akademskim radovima.

Prema Alzahrani i suradnicima (2012) [1, str. 133] prvi radovi o plagiranju tekstova i izvornog programskog kôda datiraju iz 1970-tih godina. Oni su se pretežito bavili otkrivanjem plagijata u izvornim programima pisanim u programskim jezicima Pascal i C. Dvadesetak godina kasnije pojavili su se radovi u kojima su prezentirane statističke računalne metode otkrivanja kopiranja i u prirodnim jezicima. 1990-tih godina znanstveni istraživači počinju ozbiljnije publicirati radove o akademskim plagijatima pa tako Samuelson (1994) [18] polemizira o etičnosti i kršenju autorskih prava izdavača u slučaju autoplagijarizma. Autori na prijelazu tisućljeća uglavnom se bave problemima pronalaženja plagijata u zatvorenim sustavima unutar akademskih ustanova i web plagiranjem. Suvremeni istraživači pokušavaju (1) dotjerati postojeće sustave kako bi bili efikasniji i efektniji, (2) koriste semantičke i stilističke sličnosti dokumenata i (3) pronalaze načine izvlačenja znanja iz njih.

U prvom poglavlju rada definirani su ili opisani plagijat i srodni pojmovi, posljedice plagijarizma, najčešća mjesta pojavnosti plagiranja te počeci istraživanja ove domene. U drugom poglavlju opisane su metode plagiranja i njihovi rezultati - vrste plagijata. Treće poglavlje predlaže detaljnu taksonomiju plagijata koja je provedena prema više kriterija. Četvrto je poglavlje posvećeno metodama plagiranja. postupcima otkrivanja plagiranja: pristupima, fazama, metodama i algoritmima za otkrivanje plagijata te konačno prijedlog metodologije za otkrivanje plagijata. U petom poglavlju se raspravlja o pristupima problemima otkrivanja plagiranja, o ograničenjima pristupa, ograničenjima ovog rada te o područjima daljnjih istraživanja. Šesto poglavlje donosi završne i zaključne misli.

Pri otkrivanju plagijata od presudne su važnosti postupci njihova otkrivanja, čiji su ključni elementi metode i algoritmi otkrivanja plagiranja.

II. Metode plagiranja

Prema Alzahraniju i suradnicima (2012) [1, str. 135], pod uvjetom da nisu ispravno referencirani originalni izvori, plagirani dijelovi mogu nastati parafraziranjem, sažimanjem originalnog teksta, kombiniranjem, restrukturiranjem, generalizacijom ili specifikacijom koncepata.

Maurer i suradnici (2006) [19, str. 1051–1052] navode sljedeće metode plagiranja:

- Metoda kopiranja i lijepljenja (engl. *copy-paste*): doslovno kopiranje teksta.
- Plagiranje ideja: korištenje sličnih koncepata i misli koji nisu opće poznati.
- Parafraziranje: gramatičke izmjene, korištenje istoznačnica, promjena redoslijeda riječi u rečenici, korištenje drugih riječi i izraza za iste misli.
- Umjetničko plagiranje: korištenje drugih medija za suštinski isto djelo.
- Plagiranje kôda: korištenja programskog kôda, algoritama, klasa ili funkcija bez dozvole ili referenciranja.
- Nedostatak poveznica na izvore: postojanje navodnika, ali ne i dovoljno informacija o izvoru, poveznice koje više ne vrijede.
- Nepravilno/neprecizno korištenje navodnika.
- Dezinformiranje referencama: referenca upućuje na krivi ili nepostojeći izvor.
- Plagiranje prijevodom: prijevod bez reference.

III. Taksonomija plagijata

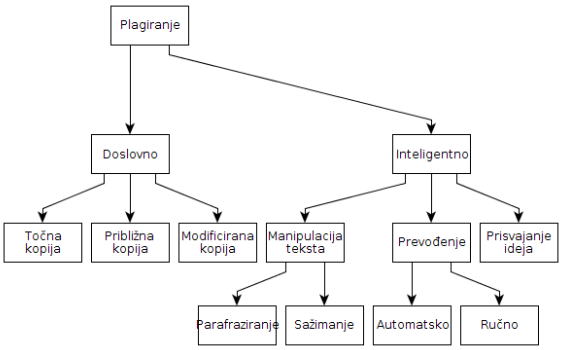
Svi autori razlučuju više tipova DT plagijata [20], [21]. Tako Maurer i suradnici (2006) [19, str. 1051] plagijate dijele u kategorije ovisne od namjere plagijatora: slučajni, nenamjerni, namjerni i autoplagijat.

Schwarzenegger i Wohlers (2006) [22, str. 3] razlikuju sedam tipova plagijata: potpuni plagijat, plagijat prijevodom, copy/paste plagijat, parafraziranje, autoplagijat, *ghostwriter* te citiranje izvan konteksta.

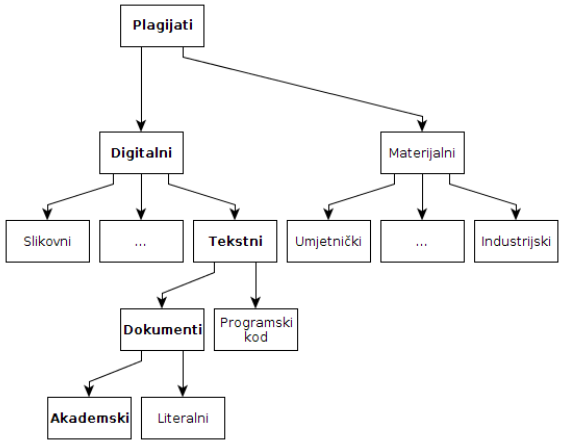
Kakkonen i Mozgovoy (2010) nude [23, str. 4] prilično drugačiju podjelu: doslovno kopiranje, plagijat parafraziranjem, tehnički prikriveni plagijat (vidi IV.II. Metode i algoritmi otkrivanja plagijata), namjerno netočno korištenje literature i teški plagijat, pri čemu u posljednju kategoriju uključuje a) korištenje tuđih ideja, koncepata, mišljenja; b) prijevod, c) *ghostwriter* i d) umjetnički plagijat.

Alzahrani i suradnici (2012) predlažu taksonomiju plagiranja [1, str. 134] prikazanu na slici 1. Temelj njihove taksonomije je ponašanje autora prilikom plagiranja, odnosno način plagiranja.

Plagijate možemo dijeliti na tipove ili vrste prema više kriterija. Tako **prema porijeklu i namjeni** (slika 2), objekti plagiranja mogu biti materijalni (industrijski, umjetnički) i nematerijalni. Nematerijalni mogu biti u izvorno digitalnom obliku (tekstovi, izvorni programi i sl.) ili se mogu digitalizirati (umjetničke slike, pjesme i sl.).



Slika 1. Taksonomija plagiranja (Alzahrani)



Slika 2. Objekti plagiranja

Tekstne plagijate možemo dijeliti na akademske i literalne [15]. Literalni plagijati izazivaju umjetničku i financijsku štetu izvornom autoru. Akademski plagijati mogu izazvati akademsku i posrednu financijsku štetu. Sustavna provjera DT dokumenata i sustavna borba protiv plagiranja vodi se upravo u akademskim krugovima.

DT plagijati mogu se **prema kriteriju tehničke realizacije plagiranja**, dijeliti na tipove [20], [21]:

1. Klon ili potpuni plagijat (engl. *clon*) – podmetanje tuđeg dokumenta kao svojeg.
2. Prijevod (engl. *translation*) – prijevod tuđeg dokumenta s drugog jezika bez navođenja autorstva i dozvole autora.
3. Kopija (engl. *copy*) – dokument koji sadrži znatan udio teksta iz jednog izvora, bez značajnije promjene.
4. Supstitut (engl. *find/replace*) – originalnom su dokumentu zamijenjene ključne riječi i izričaji, ali je dokument zadržao prvobitni smisao i sadržaj izvornog dokumenta.
5. Spoj (engl. *remix*) – dokument u kome su parafrazirani drugi dokumenti te sastavljeni na način da djeluju kao smislena cjelina.
6. Autoplagijat (engl. *recycle*) – korištenje vlastitih ranijih dokumenata bez odgovarajuće naznake.
7. Hibrid (engl. *hybrid*) – dokument u kome su kombinirani korektno citirani dijelovi i oni kopirani.
8. Mješavina (engl. *mashup*) – nekonzistentna mješavina dokumenata različitih izvora bez korektnog citiranja.
9. Otpad (engl. *error*) – dokument koji uključuje citate iz nepostojećih ili netočnih izvora.
10. Nakupina (engl. *aggregator*) – dokument u kome se pravilno citiraju izvori, ali ne sadrži originalnost.
11. Ponavljanje (engl. *re-tweet*) – dokument koji uključuje odgovarajuće citate, ali se previše veže na tekst ili strukturu izvornih dokumenata.
12. Proizvod (engl. *ghostwriter*) – dokument koji je zapravo (najčešće plaćena) usluga nekog drugog autora onom potpisanom.

Prethodna podjela po tipu mogla bi se, **prema kriteriju potencijalne težine posljedica**, reducirati na tri kategorije:

- Drski plagijat (klon, prijevod, kopija, supstitut i proizvod). Ovdje su i namjera i potencijalna šteta od plagiranja najveći, a plagijator najbezobzirniji ili najnaivniji.
- Pravi plagijat (spoj, hibrid, mješavina, otpad). U akademskoj su zajednici ovakvi pokušaji plagiranja učestali, posebice kod realizacije studentskih obaveza. Teško je razlučiti namjernost, neznanje ili naivnost autora plagijata, a teško je i njihovo otkrivanje.
- Blagi plagijat (autoplagijat, nakupina, ponavljanje). S moralnog, etičkog i pravnog stajališta, ova kategorija plagijata je najbenignija, ali svakako ne i dozvoljena ili opravdana.

Te dvije podjele nisu međusobno potpuno neovisne. To je lakše vidjeti ukoliko uvedemo još jedan pragmatični kriterij podjele: mogućnost automatiziranog otkrivanja, tj. **kriterij složenosti otkrivanja**. Dakle, unutar podjele po tipu, sve tipove plagijata možemo podijeliti na one koji se lako ili teško otkrivaju, što rezultira matricom prikazanom tablicom 1. Lako otkrivanje podrazumijeva da ih je moguće otkriti automatiziranim programskim sustavima za provjeru plagijata, a složeno otkrivanje podrazumijeva da je za njihovo otkrivanje potrebna analiza ljudskog eksperta.

Tablica 1. Matrica kategorije/složenost okrivanja

	Složenost otkrivanja	
	Jednostavno	**Složeno**
Drski	klon, kopija, supstitut	prijevod, proizvod
Pravi	-	spoj, hibrid, mješavina, otpad
Blagi	autoplagijat, ponavljanje	nakupina

Prema jezičnom porijeklu, plagijate možemo dijeliti na istojezične i plagijate prijevodom. Plagijati prijevodom mogu nastati plagiranjem dokumenata s jednog ili više jezika, a preduvjet za njihovo programsko otkrivanje je pristup programskim sustavima za automatsko prevođenje.

IV. Otkrivanje plagiranja

IV.I. Pristupi i faze otkrivanja plagiranja

Lancaster (2005) [24] je klasificirao pristupe otkrivanja plagijata prema pet kriterija:

1. Tradicionalna klasifikacija prema kojoj se dokumentima izračunavaju ili svojstva (engl. *Attribute Counting Systems*) ili struktura (engl. *Structure Metric Systems*). Lancaster [24, str. 4] smatra takvu klasifikaciju nedorečenom, jer neki sustavi imaju pristup koji ne pripada niti jednoj od dviju klasa.
2. Klasifikacija prema tipu korpusa koji se obrađuje. Ovdje imamo više podjela.
   - Prema vrsti dokumenata koji se obrađuju korpus dokumenata mogu činiti izvorni tekst programa, tekstni dokumenti ili oboje.
   - Prema izvoru dokumenata, korpusi mogu biti interni (dokumenti dostupni organizaciji), eksterni (svi izvori s Interneta) ili mješoviti.
   - Prema načinu rada, pristupi mogu biti sa ili bez tokenizacije[1].
3. Klasifikacija prema dostupnosti sustava za otkrivanje plagijata
   - Prema smještaju, mogu biti lokalni ili na webu.
   - Prema otvorenosti, mogu biti javni ili privatni.
4. Klasifikacija prema broju dokumenata koje istovremeno obrađuje korištena metrika
   - Metrike mogu biti singularne (jednodim.), parne (dvodim.) te korpusne (n-dimenzionalne; n=broj dokumenata u korpusu).
5. Klasifikacija prema složenosti korištenih metrika
   - Metrike mogu biti površinske ili strukturalne.

Prema Maureru i sur. [19, str. 1056–1061] strategiju otkrivanja plagijata trebalo bi provesti kroz tri faze (Maurer ih naziva metodama):

1. Korištenje lokalnog repozitorija dokumenata, tj. uspoređivanje provjeravanog dokumenta riječ-po-riječ s potencijalnim izvorima plagiranja.
2. Uspoređivanje provjeravanog dokumenta sa svim dostupnim web izvorima na način da se uspoređuju karakteristični dijelovi ili rečenice, a ne cijeli dokumenti.
3. Korištenje stilometrije tj. algoritma za jezičnu analizu koji uspoređuje stil sljednih odlomaka provjeravanog dokumenta te upozorava na nedosljednost odnosno promjenu stila, što ukazuje na povećanu vjerojatnost plagiranja.

Culwin i Lancaster (2001) [6, str. 3–6] prepoznaju četvero-fazni model otkrivanja plagijata: (1) faza prikupljanja u kojoj dokumenti pune repozitorij svih relevantnih dokumenata, (2) faza detekcije u kojoj programski sustav prepoznaje sumnjive parove dokumenata, (3) faza potvrde u kojoj ljudski ekspert potvrđuje ili odbacuje sumnju u plagiranje te (4) faza istraživanja u kojoj ljudski ekspert potvrđuje plagiranje i određuje sankcije za plagijatora.

[1] Tokenizacija je faza prije obrade dokumenta u kojoj se termovi (domenski objekti) mijenjaju simbolima.

IV.II. Metode i algoritmi otkrivanja plagijata

Idealan algoritam za otkrivanje plagijata trebao bi moći utvrditi sljedeće [23, str. 4]:

1. Doslovno kopiranje
   - izvorno digitalnih dokumenata i
   - digitaliziranih analognih izvora.
2. Otkrivanje parafraziranja u oblicima:
   - dodavanje ili uklanjanje riječi ili slova,
   - dodavanje namjernih pravopisnih ili gramatičkih grešaka,
   - zamjena riječi sinonimima,
   - mijenjanje poretka riječi u rečenicama ili izrazima,
   - promjene u gramatici i stilu.
3. Otkrivanje tehničkih trikova kojima se nastoje iskoristiti slabosti postojećih automatskih sustava za otkrivanje plagijata, poput:
   - korištenje fontova koji su slični po izgledu, no različiti po kôdu,
   - umjesto razmaka korištenje slova bijele boje koja čitatelj ne vidi, ali zbune SW za prepoznavanje plagijata,
   - korištenje slike teksta umjesto teksta, itd.
4. Namjerno krivo referenciranje:
   - krivo ili neprecizno označavanje navodnicima,
   - namjerno netočne ili nepostojeće reference,
   - korištenje isteklih poveznica na izvore.
5. Teško plagiranje:
   - plagiranje ideja (slični koncepti ili mišljenja izvan općepoznatog, bez ispravnog referenciranja),
   - plagiranje prevedenog teksta (prijevod bez priznanja izvornog autora),
   - korištenje testa *ghostwritera*,
   - umjetničko plagiranje (tuđi rad na drugom mediju).

Takvom idealnom algoritmu približavamo se razvojem postojećih i novih metoda, algoritama i metodologija. Do sada razvijene metode možemo klasificirati u dvije klase: **vanjske** (**ekstrinzične**) te **unutarnje** (**intrinzične**), već prema tome da li se plagiranje traži uspoređujući potencijalni plagijat s potencijalnim izvornikom ili se unutar samog dokumenta traže dokazi za plagiranje [25, str. 1].

Lukashenko i sur. (2007) [9, str. 1] navode da se metode za otkrivanje plagiranja mogu svrstati u dvije klase: **metode za prevenciju** koje su vremenski zahtjevne ali imaju dogoročne učinke i **metode za otkrivanje** koje su kratkoročne i imaju brze učinke.

Metode prevencije su [9, str. 1] "mjere opreza kojima je cilj spriječiti razvoj bolesti". One ne djeluju tako brzo kao metode otkrivanja plagiranja, no njihov je učinak dugoročan, stoga i vrlo poželjan. **Politika iskrenosti i poštenja** je nastojanje cijelog društva da

utječe na svijest, savjesnost, moral, etičnost, stav, ..., odgoj svih ljudi ili dionika sustava. U krajnjem ili minimalističkom slučaju umjesto cjelokupnog društva može se utjecati na neki organizirani dio poput sustava znanosti i visokog obrazovanja ili sveučilišta koja potiču vrijednosti akademskog integriteta. **Sustav kažnjavanja** podrazumijeva donošenje propisa i kazni za njihovo kršenje na razini društva ili sustava. Ove dvije metode djeluju kao preventiva i liječenje.

Statističke metode ne teže za tim da "razumiju" dokument. Ove metode iz dokumenata izvlače ne baš uvijek strogo statističke veličine. Pored frekvencije riječi, računaju se i njihove težinske vrijednosti (ponderi). Neki autori u statističke veličine ubrajaju i različite mjere udaljenosti [26, str. 1–2]: Hammingova udaljenost, Euklidova udaljenost, Lempel-Ziv udaljenost, kompresijska udaljenost te dvije najznačajnije: informacijska udaljenost i normalizirana informacijska udaljenost [26, str. 2–4]. Statističke su metode najčešće sastavni dio drugih metoda.

Metode otkrivanja kopiranja obuhvaća algoritme koji se mogu svrstati u četiri podkategorije [27, str. 1–2], [28], [29, str. 430], [30], [31, str. 23–81]

1. Klasični algoritmi ili **algoritmi za uspoređivanje znakovnih nizova** su brojni. U radovima se spominju Brute-Force (Naive), Knuth-Morris-Pratt, Boyer-Moore, Boyer-Moore-Smith, Boyer-Moore-Horspool, Boyer-Moore-Horspool-Raita, Simon, Colussi, Galil, Apostolico-Giancarlo, Turbo-BM, Reverse Colussi, Sunday algoritmi (Quick Search, Optimal Mismatch, Maximal Shift) i Ratcliff/Obershelp. Neki od algoritama su u stanju pretraživati sličnost teksta sa više uzoraka, npr. Commentz-Walter, Hume, Baeza-Yates

2. Algoritmi sufiksa konačnih automata (engl. Suffix automata algorithms) su Reverse Factor, Turbo Reverse Factor, Suffix Tree i Aho-Corasick algoritam.

3. Posmični algoritmi (engl. Bit-parallelism algorithms) su Shift-Or algorithm, Shift-And i BNDM.

4. Algoritmi i metode korištenja sažetaka su primjerice algoritmi Harrison, Karp-Rabin, Running Karp-Rabin Greedy String Tiling, Las Vegas, Monte Carlo, metoda prosijavanja (engl. *Winnowing*) [32], Wu-Manberov algoritam za višestruke uzorke [33], metoda sjeckanja (engl. *chunking*) [29, str. 430] itd.
Oni koriste kriptografske nepovratne (engl. *hash*) funkcije poput MD5 kako bi se iz manjih ili većih dijelova teksta dobili sažeci. Veličina teksta određuje osjetljivost algoritma. I najmanja promjena teksta mijenja sažetak. Iz sažetaka se stvaraju matrice sličnosti [30] dvaju dokumenata koje se uspoređuju. Metode su zahtjevne prema potrebnim računalnim resursima [29, str. 430].

Metode otkrivanja parafraziranja i semantičke sličnosti dvije su grupe srodnih metoda, a zajedno su ovdje stoga što otkrivajući parafraziranje otkrivamo i semantičku sličnost, dok semantička sličnost razotkriva parafraziranje. Primjeri metoda su metode obrade prirodnih jezika [25] (engl. *Natural Language Processing*, NLP), morfološka analiza (engl. *Morphological Analysis*), sintaktičko raščlanjivanje (engl. *Syntactic Parsing*), metoda uspoređivanja meta-informacija dokumenata, metoda uspoređivanja ključnih riječi (engl. *Keyword Similarity*) [29, str. 430–431] metoda tokenizacije ili opojavničenja [34, str. 74]. U području semantike snažno se razvijaju metode umjetne inteligencije, obrade prirodnih jezika, rudarenja podataka, metode stilometrijske analize teksta, metode izvlačenja i prezentacije znanja [35]–[37] iz dokumenata, podataka i prirodnih jezika (grafičke metode prezentacije znanja poput BG i NOK, podatkovni modeli, semantičke mreže, neuralne mreže, metoda MultiNets, HSF metoda za predstavljanje uzoraka u prirodnim jezicima). Stilometrijske metode su postale toliko pouzdane da se njima izvršene analize priznaju u nekim zakonodavstvima (SAD, UK, Australija) [38].

Ovaj kratki pregled metoda i algoritama koji mogu poslužiti za otkrivanje plagijata zasigurno nije potpun te zaslužuje istraživanje koje bi utvrdilo njihov potpuni skup te za svaku prednosti, nedostatke, domene djelovanja, složenost, efikasnost, efektnost te međusobno usporedilo one istih klasa u radu nad istom domenom. Podjela nije jednoznačna jer dio algoritama koristi elemente po kojima bi mogli pripadati u više klasa.

IV.III. Prijedlog metodologije otkrivanja plagiranja

Na temelju dosadašnjih spoznaja, predlaže se metodologija otkrivanja plagiranja. Nakon što sustav za otkrivanje plagiranja zaprimi dokument, nad njime se provode metodološki koraci prikazani dijagramom toka na slici 3.

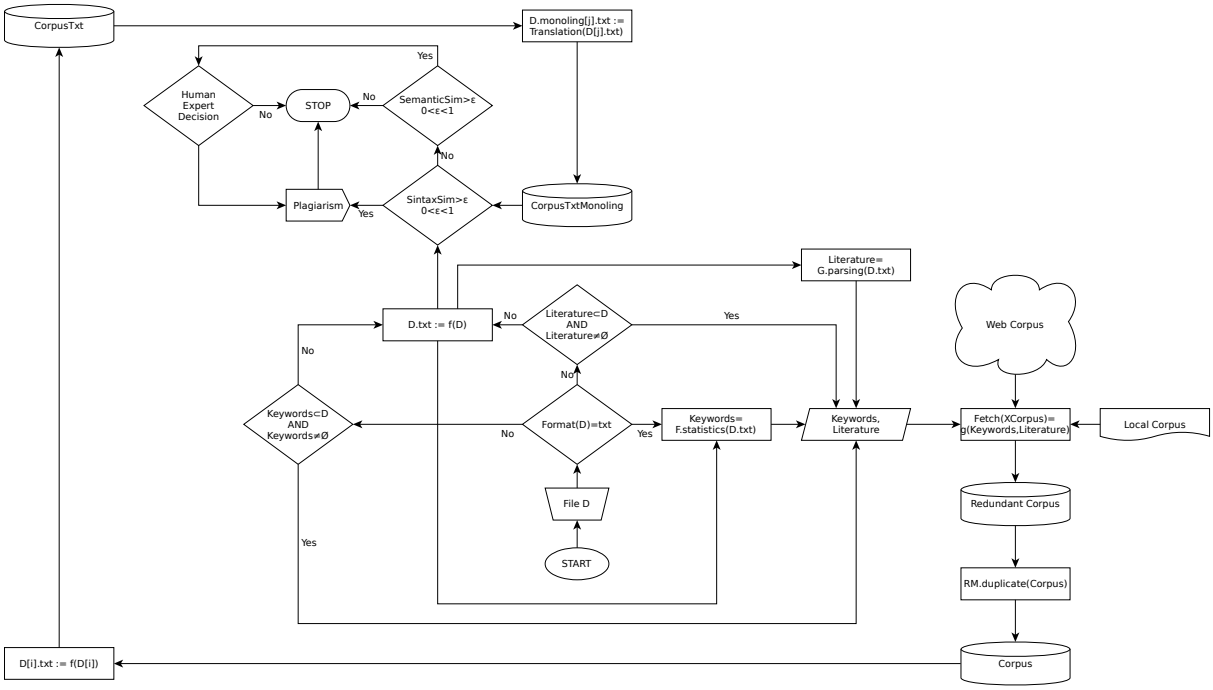
1. Parsiranje provjeravanog dokumenta

1.a) Pronalaženje ključnih riječi

U idealnom slučaju dokument već ima izdvojene ključne riječi. U slučaju da ih nema, ide se na korak 1.b). i potom vrši statistička obrada dokumenta koja izvlači najfrekventnije riječi pri čemu se izuzimaju trivijalne. Ovaj se korak može unapređivati korištenjem algoritama za morfološku normalizaciju [34], [39], uvođenjem novih algoritama, korištenjem algoritama poput Ratcliff/Obershelp ili korištenjem postojećih online servisa [40]. Izazov ovoga koraka je eliminirati česte ali ne-sadržajne riječi te pojmove sastavljene od više riječi.

1.b) Izvlačenje teksta

Provjeravani dokument pretvara se u tekstnu inačicu postojećim besplatnim alatima otvorenog kôda: *catdoc* ili *antiword* za doc format [41, str. 24], *docx2txt* za docx format, *pdftotxt* za pdf format [42, str. 19], *odt2txt* za odt format te *html2text* za html format ulaznog dokumenta.



Slika 3. Dijagram toka metodologije

1.c) Izvlačenje podataka o literaturi

U idealnom slučaju, dokument ima popis korištene literature. U popis se dodaju i sve URL poveznice, uz uklanjanje duplikata. U najgorem slučaju rezultata nema, tj. popis literature je prazan skup.

2. Stvaranje korpusa dokumenata

2.a) Stvaranje korpusa

Iz URL poveznica, ključnih riječi i popisa korištene literature stvara se web korpus, tj. dohvaćaju se sve datoteke (html, pdf, doc(x), odt) koje su potencijalni izvori plagiranja. Dodatno se može definirati i lokalni korpus dokumenata, posebno ako se provjerava više dokumenata iste ili slične tematike.

2.b) Uklanjanje duplikata

Eventualni duplikati se detektiraju pomoću nepovratne *hash* f-je (npr. MD5) te uklanjaju iz korpusa.

2.c) Izvlačenje teksta iz dokumenata korpusa

Dokumenti korpusa pretvaraju se u tekstne inačice po principu opisanom u 1.b).

2.d) Rješavanje problema višejezičnosti

U ovom se koraku utvrđuje zastupljenost pojedinih jezika u korpusu te njihovo prevođenje u dominantni jezik. Za svaki dokument algoritamski se utvrđuje na kojem je jeziku napisan, a potom se automatski prevode oni koji su napisani nekim drugim jezikom. To svakako nisu zadovoljavajući prijevodi za čitanje ili publiciranje, no ključne riječi i znanje uglavnom ostaje dovoljno dobro sačuvano.

3. Otkrivanje plagiranja

3.a) Otkrivanje sintaktičke sličnosti

Postupak otkrivanja plagiranih dijelova - sintaktičke sličnosti parova dokumenata koristi jedan ili više algoritama za uspoređivanje znakovnih nizova.

3.b) Utvrđivanje semantičke sličnosti

Postupak utvrđivanja semantičke sličnosti parova dokumenata odvija se korištenjem algoritama sažetaka koji onda služe kao brzo sito za selektivnu primjenu metoda otkrivanja parafraziranja, semantičke sličnosti i metoda stilometrijske analize.

3.c) Rezultat

Izračun opće ocjene i kategorizacije plagijata za jednostavnije oblike plagiranja ili procjena semantičke sličnosti uz preporuke ljudskom ekspertu. Opća ocjena može biti na skali 0-1 (1 predstavlja jednakost dokumenata). Preporuka ljudskom ekspertu trebala bi sadržavati sumnjive (potencijalno plagirane) dijelove dvaju ili više dokumenata.

V. Rasprava

U otkrivanju DT plagijata postoje dva pristupa u pronalaženju sličnosti ili istovjetnosti:

- pronalaženje plagiranja **izražavanja** ideja (lat. *forma*),
- pronalaženje plagiranja samih **ideja** (grč. *idea*).

Prvi je pristup tehnički i logički lakše izvediv, pa je u tom pristupu zabilježen veći razvoj, kako u teorijskom smislu (algoritmi, metode) tako i u programskim sustavima za njihovo otkrivanje, no taj pristup, iako ima još razvojnih mogućnosti, nije i neće biti u mogućnosti otkriti sve plagijate. Uvidjevši ograničenja prvog pristupa, suvremeni istraživači se sve više okreću drugom pristupu koji je mnogo teži i još nije ostvario puni potencijal razvoja. U njemu se istraživači sve više okreću metodama kojima se izvlači semantika ili čak i znanje iz dokumenata. Usporedba se tako prenosi na višu, semantičku razinu - potragom za semantičkom sličnošću, stilističkom nedosljednošću, uspoređivanjem znanja, što je algoritamski i programski složenije, zahtjevnije je i u

pogledu potrebite računalne snage za koju je sve češće potrebno upregnuti mrežne operacijske i programske sustave (homogene ili heterogene), kako bi se provjera plagiranja provela u razumnom vremenu. Efikasnosti radi, oba pristupa će u nekom razvijenom sustavu trebati objediniti. Nagrada za uspješno svladavanje te sljedeće evolucijske faze borbe protiv plagiranja bila bi mogućnost otkrivanja većeg postotka najsloženijih i najprofinjenijih vrsta plagijata, za koje danas nema zadovoljavajućih rješenja izvan domene ljudskog eksperta, koji će, bez obzira na napredak automatiziranih sustava za pronalaženja plagijata i dalje ostati konačni prosuditelj je li neki dokument doista plagijat ili nije.

Važna područja daljnjeg istraživanja koja izlaze iz okvira ovog rada su (1) utvrđivanje jedinstvenih i zajedničkih mjera sličnosti koje su svojstvene algoritmima pronalaženja plagijata, (2) usporedba postojećih programskih sustava te (3) usporedba algoritama i metodološkog okvira koji postojeći programski sustavi koriste. Metode stilometrijske analize te izvlačenja znanja iz DT dokumenata najperspektivniji su istraživački smjerovi. No, velika se korist može izvući iz "klasičnih" metoda i algoritama pronalaženja sličnosti. Potencijalima posebno obećavaju algoritmi (poput Ratcliff/Obershelp), koji uspoređujući ulazni string s kontrolnim mogu zanemariti određene razlike ili greške, poput pravopisnih. Nadovezujući algoritam tako može ulazne riječi uspoređivati s rječnikom te prepoznati pojmove i izbjeći mnogobrojne tehničke komplikacije u početnoj fazi izvlačenja znanja iz DT dokumenata. To bi olakšalo i postojeće algoritme i sustave normalizacije i lematizacije teksta [39], [43].

VI. Zaključak

S obzirom na mnogovrsnost tipova plagijata te načina i složenosti njihova stvaranja, otkrivanje plagiranja DT dokumenata kompleksan je zadatak. Za (raz)otkrivanje nekih vrsta plagiranja danas postoje pouzdani i efikasni postupci, metode, algoritmi i pretežito komercijalni programski sustavi (koji u pravilu ne pružaju garanciju povjerljivosti ili postoje ograničenja u pogledu broja predanih dokumenata). Plagijati često istovremeno pripadaju u više tipova i to dodatno otežava problem njihova otkrivanja. Naime, neke metode plagiranja ili njihove kombinacije rezultiraju plagijatima čije otkrivanje je vrlo teško. Za te vrste plagiranja još nema pravog odgovora, tj. rješenja problema njihova pronalaženja putem algoritama pa onda i njihovo programski realizirano pronalaženje. No, razvijaju se pristupi, metode i algoritmi koji obećavaju napredak u njihovom otkrivanju tj. da problem otkrivanja plagiranja konvergira u kategoriju rutinski rješivih, računalno podržanih problema.

Literatura

[1] S. M. Alzahrani, N. Salim, i A. Abraham, „Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods“, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, tom 42, izd 2, str. 133–149, mart 2012.

[2] P. Clough i others, „Old and new challenges in automatic plagiarism detection“, u *National Plagiarism Advisory Service, 2003; http://ir. shef. ac. uk/cloughie/index. html*, 2003.

[3] R. Kumar i R. C. Tripathi, „An Analysis of Automated Detection Techniques for Textual Similarity in Research Documents“, *International Journal of Advanced Science and Technology*, tom 56, str. 99–110, 2013.

[4] Hrvatski Sabor, „Zakon o autorskom pravu i srodnim pravima“. Narodne novine, 2003.

[5] F. Culwin i T. Lancaster, „Plagiarism, prevention, deterrence & detection“, *CiteSeerX*, 2000.

[6] F. Culwin i T. Lancaster, „Plagiarism issues for higher education“, *VINE*, tom 31, izd 2, str. 36–41, jun 2001.

[7] T. Lancaster, „Effective and efficient plagiarism detection - Phd thesis“, South Bank University, 2003.

[8] S. M. Zu Eissen i B. Stein, „Intrinsic plagiarism detection“, u *Advances in Information Retrieval*, Springer, 2006, str. 565–569.

[9] R. Lukashenko, V. Graudina, i J. Grundspenkis, „Computer-based plagiarism detection methods and tools: an overview“, u *Proceedings of the 2007 international conference on Computer systems and technologies*, 2007, str. 40.

[10] Merriam-Webster Dictionary, „Plagiarism - Definition and More from the Free Merriam-Webster Dictionary“, 2014. [Na internetu]. Pristupačno na: http://www.merriam-webster.com/dictionary/plagiarism. [Pristupljeno: 13-avg-2014].

[11] Encyclopedia Britannica, „Plagiarism“, 23-okt-2013. [Na internetu]. Pristupačno na: http://www.britannica.com/EBchecked/topic/462640/plagiarism. [Pristupljeno: 04-jan-2015].

[12] Cambridge University Press, „Meaning of “plagiarize” in the Cambridge English Dictionary“, 2015. [Na internetu]. Pristupačno na: http://dictionary.cambridge.org/dictionary/english/plagiarize?q=plagiarism. [Pristupljeno: 12-sep-2015].

[13] Oxford dictionary, „Plagiarism: definition of plagiarism“, 17-dec-2014. [Na internetu]. Pristupačno na: http://www.oxforddictionaries.com/definition/english/plagiarism. [Pristupljeno: 06-jan-2015].

[14] University of Oxford, „Plagiarism“, 20-nov-2011. [Na internetu]. Pristupačno na: http://www.ox.ac.uk/students/academic/guidance

/skills/plagiarism#. [Pristupljeno: 06-jan-2015].
[15] N. Meuschke i B. Gipp, „State-of-the-art in detecting academic plagiarism“, *International Journal for Educational Integrity*, tom 9, izd 1, 2013.
[16] Plagiarism.org, „What is Plagiarism?“, 2014. [Na internetu]. Pristupačno na: http://www.plagiarism.org/plagiarism-101/what-is-plagiarism/. [Pristupljeno: 06-jan-2015].
[17] Action Plagiarius, „Plagiarius | Innovation contra Imitation“, 2014. [Na internetu]. Pristupačno na: http://plagiarius.de/. [Pristupljeno: 08-jan-2015].
[18] P. Samuelson, „Self-plagiarism or fair use“, *Communications of the ACM*, tom 37, izd 8, str. 21–25, 1994.
[19] H. A. Maurer, F. Kappe, i B. Zaka, „Plagiarism-A Survey“, *J. UCS*, tom 12, izd 8, str. 1050–1084, 2006.
[20] iParadigms, „White Paper - The Plagiarism Spectrum: Instructor Insights into the Ten Types of Plagiarism“, USA, 2012.
[21] V. Juričić, „Detekcija plagijata u višejezičnom okruženju - doktorska disertacija“, University of Zagreb, Zagreb, 2012.
[22] C. Schwarzenegger i W. Wohlers, „Quellen zitieren, nicht plagiieren“, *Universität Zürich, Unijournal 4/06*, str. 3, 2006.
[23] T. Kakkonen i M. Mozgovoy, „Hermetic and web plagiarism detection systems for student essays—an evaluation of the state-of-the-art“, *Journal of Educational Computing Research*, tom 42, izd 2, str. 135–159, 2010.
[24] T. Lancaster i F. Culwin, „Classifications of plagiarism detection engines“, *Innovation in Teaching and Learning in Information and Computer Sciences*, tom 4, izd 2, maj 2005.
[25] M. Chong, L. Specia, i R. Mitkov, „Using natural language processing for automatic detection of plagiarism“, u *Proceedings of the 4th International Plagiarism Conference (IPC 2010), Newcastle, UK*, 2010.
[26] M. Li, X. Chen, X. Li, B. Ma, i P. M. B. Vitanyi, „The Similarity Metric“, *IEEE Transactions on Information Theory*, tom 50, izd 12, str. 3250–3264, dec. 2004.
[27] P. D. Michailidis i K. G. Margaritis, „On-line string matching algorithms: survey and experimental results“, *International Journal of Computer Mathematics*, tom 76, izd 4, str. 411–434, jan. 2001.
[28] V. Alfred, „Algorithms for finding patterns in strings“, *Algorithms and Complexity*, tom 1, str. 255, 2014.
[29] B. Stein i S. M. zu Eissen, „Near Similarity Search and Plagiarism Analysis“, u *From Data and Information Analysis to Knowledge Engineering*, M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, i W. Gaul, Prir. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, str. 430–437.
[30] B. Stein, „Principles of hash-based text retrieval“, u *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, str. 527–534.
[31] G. A. Stephen, *String search*. University College of North Wales, 1992.
[32] S. Schleimer, D. S. Wilkerson, i A. Aiken, „Winnowing: local algorithms for document fingerprinting“, u *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 2003, str. 76–85.
[33] S. Wu, U. Manber, i others, „A fast algorithm for multi-pattern searching“, 1994.
[34] R. Lujo, „Lociranje sličnih logičkih cjelina u tekstualnim dokumentima na hrvatskome jeziku - magistarski rad“. Fakultet elektrotehnike i računarstva Sveučilišta u Zagrebu, 2010.
[35] M. Pavlic, A. Jakupovic, i A. Mestrovic, „Nodes of Knowledge Method for Knowledge Representation“, *Informatologia*, tom 46, izd 3, str. 206, 2013.
[36] A. Jakupovic, M. Pavlic, A. Mestrovic, i V. Jovanovic, „Comparison of the Nodes of Knowledge method with other graphical methods for knowledge representation“, u *Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on*, 2013, str. 1004–1008.
[37] M. Pavlić, A. Meštrović, i A. Jakupović, „Graph-based formalisms for knowledge representation“, u *Proceedings of the 17th world multi-conference on systemics cybernetics and informatics (WMSCI 2013)*, 2013, tom 2, str. 200–204.
[38] M. R. Brennan i R. Greenstadt, „Practical Attacks Against Authorship Recognition Techniques.“, u *IAAI*, 2009.
[39] J. Šnajder, „Postupci morfološke normalizacije u pretrazivanju informacija i klasifikaciji teksta“. Fakultet elektrotehnike i računarstva Sveučilišta u Zagrebu, 2008.
[40] M. Tadić, „Hrvatski morfološki leksikon“, 2005. [Na internetu]. Pristupačno na: http://hml.ffzg.hr/hml/info.php?show=hlp. [Pristupljeno: 02-nov-2015].
[41] S. Byers, „Information leakage caused by hidden data in published documents“, *IEEE security & privacy*, izd 2, str. 23–27, 2004.
[42] C. S. Burns, „Characteristics of a Megajournal: A Bibliometric Case Study“, *Journal of Information Science Theory and Practice*, tom 3, izd 2, str. 16–30, jun 2015.
[43] S. Beliga, M. Pobar, i S. Martinčić-Ipšić, „Normalization of Non-Standard Words in Croatian Texts“, *arXiv preprint arXiv:1503.08167*, 2015.