Plagijarizam i kako ga izbeći

Seminarski rad u okviru kursa
Metodologija stručnog i naučnog rada
Matematički fakultet

Ana Mitrović, Nikola Vidič
amitrovic01@gmail.com, nikolavidich@gmail.com

12. april 2015.

Sažetak

Plagijarizam je pojava koja se danas javlja u sve većem broju oblasti, tako da je postala značajna i u programiranju. Većina ljudi se jako često susreće sa njom, a isto tako i učestvuje u samom činu plagijarizma, svesno ili nesvesno. Iz tih razloga razvijaju se nove metode za detekciju plagijarizma i poboljšavaju već postojeće. Glavni problem je razumeti sam pojam plagijarizma i razlikovati radnje koje smemo od onih koje ne smemo raditi. Često je granica između ovih radnji jako tanka, pa nenamerno pravimo greške stvarajući rad bilo koje vrste. Najsigurnije je slediti određene korake kako bismo se zaštitili i bili sigurni da nismo prisvojili tuđe ideje i dela kao svoje, jer ni sami ne bismo želeli da budemo na mestu onoga čiji je rad iskopiran a trud zanemaren.

Ključne reči: plagijarizam, tipovi plagijarizma, detekcija plagijarizma, izbegavanje plagijarizma, algoritmi za detekciju plagijarizma

Sadržaj

1 Uvod

Plagijarizam (eng. *plagiarism*) je termin koji je teško definisati [6]. Postoje mnoge definicije, od kojih će neke biti navedene, međutim mnoge od njih nisu potpuno kompletne iz više razloga. Prvi razlog je što, ako plagijarizam definišemo kao krađu to i nije sasvim tačno. Kada se od nekog nešto ukrade ta osoba više nema ukradenu stvar, a kada neko ukrade od nekog autora (npr. pasus) to je više kopiranje nego krađa – autor i dalje ima taj pasus, ali ga sada ima i osoba koja ga je ,,ukrala". Sa druge strane, ako kažemo da je plagijarizam pozajmica, ni to nije dobro, jer se to ,,pozajmljeno" ne vraća. Drugi razlog iz kog su definicije plagijarizma nekompletne je što je plagijarizam širok pojam koji može da se odnosi na mnoge stvari (npr. tekstualna dokumenta (književna dela), muziku, slike, ideje, programski kod...). Mnoge definicije isključuju neke od ovih stvari, koncentrišući se na taj način na druge (najčešće samo na tekst).

U nastavku rada sledi opšta priča o plagijarizmu. Prvo ćemo probati da objasnimo šta je plagijarizam (2) i koji sve tipovi postoje (3). Nakon toga ćemo pričati o plagijarizmu teksta, odnosno o nekim metodama za njegovu detekciju (4) i o plagijarizmu programskog koda. Ovo je možda najzanimljiviji deo za informatičare, tako da ćemo obraditi metode za detekciju, kao i jedan od algoritama za detekciju (5). Na kraju slede saveti za izbegavanje plagijarizma (6), što je značajno svakome bez obzira na profesiju ali naročito studentima jer se oni uglavnom prvi put susreću sa pisanjem radova upravo na fakultetu neznajući dovoljno o plagijarizmu.

2 Pojam plagijarizma

Sa obzirom na to da ne postoji univerzalna definicija plagijarizma, pokušaćemo da Vam približimo ovaj pojam sa više definicija. Neke od najčešćih definicija su:

Definicija 2.1 *,,Plagijarizam je čin prisvajanja pisanih dela druge osobe i predstavljanja istih kao svoje. Ova vrsta prevare je usko vezana sa falsifikovanjem i piraterijom, radnjama koje dovode do kršenja autorskih prava [2]."*

Definicija 2.2 *Plagijarizam je [5]:*

- *Krađa i prosleđivanje ideja ili reči nekog drugog kao svoje.*
- *Korišćenje tuđeg dela bez navođenja izvora.*
- *Čin krađe književnog dela.*
- *Predstavljanje već postojeće ideje kao nešto novo i originalno.*
- *Kopiranje tuđih reči ili ideja bez priznavanja zasluga.*
- *Nestavljanje znakova navoda na mestu citata.*
- *Davanje pogrešnih informacija o izvoru citata.*
- *Menjanje reči u rečenici, tako da kontekst ostane isti, bez navođenja izvora.*
- *Kopiranje velikog broja reči ili ideja sa izvora tako da čine većinu rada, bez obzira na to da li su navedeni izvori.*
- *Korišćenje slika sa drugih sajtova bez navođenja izvora.*
- *Pravljenje video snimka korišćenjem delova tuđih snimaka ili slobodno korišćenje muzike zaštićene autorskim pravima.*
- *Komponovanje muzike koja koristi delove druge kompozicije.*

3 Tipovi plagijarizma

Postoji više različitih tipova plagijarizma koji su rangirani (sortirani opadajuće) po ozbiljnosti na sledeći način [7]:

1. *Kloniranje* (eng. *clone*) - Predavanje tuđeg rada, od reči do reči, kao svog.
2. *Kopiranje* (eng. *ctrl+c*) - Rad sadrži veliku dozu teksta kopiranog sa jednog izvora (bez izmena).
3. *Pronađi i zameni* (eng. *find and replace*) - Menjanje ključnih reči i fraza uz zadržavanje suštine izvornog teksta.
4. *Remiks* (eng. *remix*) - Parafraziranje sa više izvora tako da se sve uklopi.
5. *Recikliranje* (eng. *recycle*) - Izdašno pozajmljivanje iz ranijeg rada na tu temu bez citiranja.
6. *Hibrid* (eng. *hybrid*) - Kombinovanje citiranih delova sa delovima koji nisu citirani.
7. *Pire* (eng. *mashup*) - Kopiranje materijala sa više izvora.
8. *Greška 404* (eng. *404 error*) - Citiranje iz nepostojećih ili netačnih izvora.
9. *Agregator* (eng. *aggregator*) - Rad sadrži odgovarajuće citate (izvori nisu nepostojeći), ali rad skoro da ne sadrži originalne ideje.
10. *Retvit* (eng. *re-tweet*) - Rad sadrži odgovarajuće citate, ali se previše oslanja na originalan tekst.

Za svaki tip plagijarizma određena je njegova frekvencija (broj koji govori koliko često se javlja u praksi) i njegova problematičnost (broj koji govori koliki problem predstavlja kada se javi). Definišemo proizvod prve dve karakteristike, kao što je prikazano u tabeli 1. Na osnovu ovih podataka možemo sortirati tipove po proizvodu i zaključiti da su oni sa većim proizvodom „teži" i nepoželjniji tipovi plagijarizma.

Tabela 1: Frekvencija i problematičnost plagijarizma na skali 1-10

Tip plagijarizma	Frekvencija	Problematičnost	Proizvod
Kloniranje	9,5	9,5	90,25
Kopiranje	8,9	7,4	65,86
Pire	9,1	4,4	40,04
Recikliranje	5,5	2,8	15,4
Agregator	2,8	2,9	8,12
Pronađi i zameni	3,9	1,2	4,68
Remiks	5,6	0,5	2,8
Retvit	4,4	0,5	2,2
Greška 404	0,6	1,3	0,78
Hibrid	0,5	1,1	0,55

4 Detekcija plagijarizma teksta

Plagijarizam teksta se najčešće javlja među studentima. Pisanje seminarskih, master i doktorskih radova neretko bude razlog zbog koga se

studenti predaju plagijarizmu. Osim toga, ova vrsta plagijarizma se javlja i pri pisanju radova bilo koje druge vrste (kao što su naučni radovi, novinski članci, poezija...). Razvijene su razne metode za detekciju plagijarizma teksta [1]:

1. *Metode zasnovane na gramatici* (eng. *Grammar-based method*) - Koncentrišu se na gramatičku strukturu dokumenata, a onda pomoću poređenja stringova računaju sličnost dva dokumenta. Ova metoda daje dobre rezultate kod kopiranja nekog dokumenta bez modifikacija. Međutim, ograničenje ove metode je to što ako je tekst izmenjen drugim rečima koje daju isto značenje onda ona neće detektovati plagijarizam.
2. *Metode zasnovane na semantici* (eng. *Semantics-based method*) - Računaju sličnost dokumenata pomoću jednog od dva najpoznatija pristupa detekciji plagijarizma, takozvanog algoritma uzimanja otisaka (eng. *fingerprints algorithm*). Slično prethodnoj, ovakva vrsta metode ne radi dobro ako je plagijarizovan samo deo dokumenta.
3. *Gramatičko-semantičke hibridne metode* (eng. *Grammar semantics hybrid method*) - Smatra se najvažnijom metodom za detekciju jer poboljšava prethodne i postiže dosta bolje rezultate. Ove metode prevazilaze ograničenja prethodnih: pogodne su i za kompletno kopiran tekst ali i za tekst koji je modifikovan drugim rečima. Takođe, moguće je i odrediti tačno koji deo teksta je plagijarizovan.
4. *Metod spoljašnje detekcije* (eng. *External plagiarism detection method*) - Ovaj metod koristi skup originalnih dokumenata iz kojih su delovi rada mogli da budu plagijarizovani. Sumnjivi dokument testiramo na plagijarizam tako što tražimo u njemu pasuse koji su delimični ili potpuni duplikat pasusa iz originalnih dokumenata. Nakon toga, sistem za detekciju plagijarizma prosleđuje rezultate pretrage osobi koja nadzire pretragu i donosi odluku da li je neki pasus plagijarizovan.
   Jedno rešenje pitanja pretrage sličnih pasusa je poređenje svakog pasusa iz sumnjivog dokumenta, sa svim pasusima iz dokumenata u skupu. Ovo rešenje ne valja jer je preskupo. Originalni skup dokumenata mora biti dovoljno velik da bismo pronašli što veći broj plagijarizovanih pasusa.
   Drugo rešenje se zasniva na heširanju (eng. *hashing*) ili uzimanju otisaka (eng. *fingerprinting*). Razlika između ove dve metode je u broju vrednosti (otisaka) koje se dobiju. Dobijene vrednosti opisuju pasuse u dokumentima. Sada se ne vrši poređenje celih pasusa, već samo vrednosti dobijenih heširanjem ili uzimanjem otisaka, pri čemu se smatra da slični pasusi imaju slične heš vrednosti. Metod spoljašnje detekcije se može zamisliti i kao problem k-najbližih suseda. U tom slučaju, susedi bi bili pasusi iz skupa dokumenata, a klase bi se dodeljivale pasusima sumnjivog teksta.
5. *Klasterovanje* (eng. *Clustering in plagiarism detection*) - Klasterovanje dokumenata je jedna od najvažnijih metoda korišćenih za pronalaženje informacija i predstavljanje rezultata. U oblasti detekcije plagijarizma se koristi da bi se smanjilo vreme pretrage. Iako klasterovanje pomaže i dalje postoje neki problemi i ograničenja vezana za vreme i količinu memorije koju klasterovanje zahteva.

5 Plagijarizam programskog koda

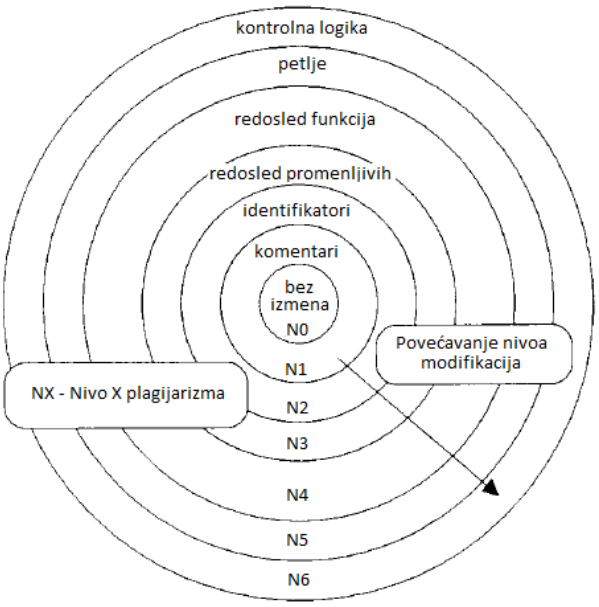
Plagijarizam programskog koda je tip plagijarizma s kojim se svaki informatičar u nekom trenutku svog obrazovanja ili karijere susreo. Ovaj tip plagijarizma je toliko zastupljen, da se javlja svakodnevno. Štaviše, skoro da ne postoji osoba koja, slučajno ili namerno, nije plagijarizovala tuđi kod.

Definicija 5.1 *„Plagijarizam programskog koda podrazumeva preduzimanje rutinskih transformacija na već postojeći kod [3].”*

Napomena: Ne treba mešati plagijarizam programskog koda sa pojmom rejuzabilnosti (eng. *reusability*). Za razliku od plagijarizma, rejuzabilnost je poželjna i podrazumeva korišćenje već postojećeg koda koji je za to i namenjen (biblioteke, šablone, optimizovane algoritme...).

Na osnovu toga kakve modifikacije se primenjuju, definisano je šest nivoa plagijarizma programskog koda:

- *Nivo 0*: originalni program bez modifikacija.
- *Nivo 1*: razlikuju se jedino komentari i uvlačenje.
- *Nivo 2*: razlikuju se nazivi promenljivih.
- *Nivo 3*: razlikuju se nazivi promenljivih, pozicije promenljivih, konstanti, procedura.
- *Nivo 4*: razlikuju se pozicije funkcija.
- *Nivo 5*: petlje se menjaju u druge koje su ekvivalentne (for, if, while...).
- *Nivo 6*: menjanje kontrolne logike.



Slika 1: Nivoi plagijarizma [3]

Postoji veliki broj algoritama za detekciju plagijarizma programskog koda. Mogu se klasifikovati u nekoliko grupa na osnovu metoda na kojima su zasnovani. Najčešće se koriste sledeće metode [1]:

1. *Stringovi* - Ovi algoritmi vrše jednostavno poređenje stringova. Algoritmi zasnovani na stringovima su brzi, ali lako može doći do zabune ako se preimenuju promenljive. Ove metode su slične metodama zasnovanim na gramatici za detekciju plagijarizma teksta jer obe vrše poređenje stringova.
2. *Tokeni* - Poređenje se vrši slično kao kod algoritama zasnovanih na stringovima, pri čemu se koristi lekser koji izdvaja tokene iz programa. Beline i komentari se zanemaruju, a sve promenljive su tokenizovane. Na taj način se eliminišu greške zbog preimenovanja. Sistemi za detekciju plagijarizma koji se koriste na univerzitetima, najčešće koriste ovaj metod. Postoji sličnost ovih sa gramatičko-semantičkim hibridnim metodama detekcije plagijarizma teksta. Obe metode su otporne na preimenovanje (promenljivih kod koda, reči kod teksta).
3. *Drveta parsiranja* - Izgradnjom stabla, moguće je primetiti dodatne sličnosti između dva programa kao što su uslovni iskazi.
4. *Grafovi zavisnosti u programu* (eng. *Program Dependency Graphs*) - Grafovi zavisnosti vide tok izvršavanja programa i na taj način omogućavaju uočavanje sličnosti u logici dva koda. Cena ovoga su veća složenost i vreme izračunavanja.
5. *Metrike* - Metrike broje pojavljivanja određenih delova (npr. operatora, različitih tipova, promenljivih datog tipa, petlji, itd.) u programskom kodu. Prednosti metrika su svakako lakoća izračunavanja, jednostavno i brzo poređenje. Sa druge strane, mana ovog pristupa je da metrike mogu dati pogrešan rezultat. Delovi koda sa podudarnim rezultatima u skupu metrika moraju biti isti.
6. *Hibridne metode* - Koristi se kombinacija dve ili više metoda da bi se maksimalno iskoristile prednosti koje te metode nude. Jedan takav metod može da koristi sposobnost uočavanja sličnosti koju imaju drveta parsiranja i brzinu algoritma zasnovanog na stringovima.

Ručna detekcija plagijarizma programskog koda može postati jako teška. Dovoljno je svega nekoliko jednostavnih operacija da učine program vizuelno dosta različitim od početnog. Iz tog razloga, bolje je detekciju ostaviti programima specijalizovanim za to. Neki od poznatijih su:

- PlagAware (`http://www.plagaware.com/`)
- PlagScan (`http://www.plagscan.com/`)
- iThenticate (`http://www.ithenticate.com/`)
- CheckForPlagiarism (`http://www.checkforplagiarism.net/`)
- PlagiarismDetection (`http://plagiarismdetection.org/`)

Postoje mnogi algoritmi za detekciju plagijarizma sa različitim karakteristikama. Sledeći algoritam oduzima dosta procesorskog vremena ali je jednostavan za implementaciju. Zasnovan je na poređenju stringova i sastoji se iz 4 koraka:

1. Obrisati sve komentare.

2. Ignorisati sve praznine i dodatne redove, osim ako su potrebni kao delimiteri.
3. Izvršiti poređenje dva fajla, originalnog i fajla za koji proveravamo da li je plagijat, pomoću UNIX komandi (npr. diff, grep, wc).
4. Iz prethodnog koraka dobijamo procenat karaktera koji su isti. Na osnovu procenta se vidi koliko su dva rada slična, te se ova mera naziva **korelacija karaktera** [3].

Sa obzirom na to da pravljenje većih promena u kodu podrazumeva razumevanje čitavog programa, većina ljudi se odlučuje na menjanje komentara, belina i eventualno imena promenljivih (nivoi 1 i 2 na slici 1), tako da ovaj algoritam detektuje puno slučajeva plagijarizma.

6 Kako izbeći plagijarizam: saveti za studente

Pisanje naučnih i stručnih radova često predstavlja novo iskustvo za studente. S obzirom da pojam plagijarizma nema preciznu definiciju, studenti lako mogu počiniti plagijarizam, a da toga nisu ni svesni. Sledeći saveti [4] mogu pomoći studentima da izbegnu plagijarizam:

- *Posavetujte se sa profesorom*
  Ako postoje neke nejasnoće oko pisanja rada, najbolje je pitati profesore. Profesori imaju dosta iskustva u pisanju radova i lako mogu razrešiti sve nejasnoće.
- *Napravite plan izrade rada*
  Planiranje je prvi i najvažniji korak ako želite da izbegnete plagijarizam. Ako se isplanira pisanje rada, šta će biti naš originalni doprinos, a za šta koristimo literaturu, nećemo zaboraviti da citiramo.
- *Vađenje beleški*
  Najbolji način pripreme za pisanje rada je vađenje beleški iz literature. Na taj način imamo organizovane sve informacije koje su potrebne za samu izradu. Ako se izvlačenje uradi kako treba, pored neophodnih informacija, imaćemo i podatke o tome iz koje literature su preuzeti i na kojim stranama se nalaze. Neki stručnjaci čak preporučuju da se podaci iz različitih izvora označe drugim bojama, fontovima...
- *Ako nisi siguran da li treba da citiraš – citiraj.*
  U slučaju da niste sigurni da li je neka ideja Vaša ili ste samo malo promenili nešto što ste pročitali, treba da citirate izvor. Iako na prvi pogled može delovati da na taj način ne dajete dovoljan doprinos, ovo pokazuje da imate sposobnost obrade informacija, a ne samo kopiranja istih.
- *Naznačite **ko** je **šta** rekao*
  Recimo da pišete rad na temu Petrove diskusije Miloševog mišljenja o Ivu Andriću. Ako u svom radu napišete: ,,...on je lepo prikazao društveni položaj pisca...". Ko je ,,on" u ovoj rečenici? Petar, Miloš ili Ivo? A ko je ,,pisac": Miloš, Ivo ili možda neki lik iz knjige? Uvek vodite računa da razgraničite ko je šta rekao.

- *Naučite da parafrazirate*
  Parafraza je način da se iskaže tuđa ideja, svojim rečima. Studenti često greše kada misle da treba sakriti činjenicu da se oslanjaju na druge izvore. Zapravo, parafraziranje pomaže da se ideje drugih autora fino uklope u rad.
- *Procenite izvore informacija*
  Šta je dobar izvor? Koje uslove rad treba da zadovolji da biste se oslanjali na njega? Niko ne može da garantuje da taj rad nije plagijarizovan. Ono što treba da uradite, je da pogledate ko su autori rada, koje izvore su koristili i kada je rad objavljen. Na taj način možete da procenite da li verujete nekim autorima ili ne.

7 Zaključak

Plagijarizam, koje god forme i nivoa bio, čin je prevare. Ne treba ga izbegavati samo zbog toga što nije po zakonu i što danas postoje brojni algoritmi i programi za automatsko detektovanje plagijarizma. To treba raditi i iz moralnih razloga: zbog poštovanja sebe i drugih. Drugih, jer je ta osoba verovatno uložila mnogo svog truda i vremena u rad. A sebe, jer ono što naučite prilikom istraživanja neke teme i izrade rada mnogo će Vam više koristiti u daljem životu, traženju posla, praksi, a i uživaćete u samom osećaju ponosa jer ste postigli nešto vredno iz čega će drugi moći da uče.

Literatura

[1] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, and Vaclav Snasel. Overview and Comparison of Plagiarism Detection Tools. *Department of Computer Science, VSB-Technical University of Ostrava*, pages 163–171.

[2] Encyclopædia Britannica Inc. Encyclopedia Britannica, 2013. on-line at: http://www.britannica.com/EBchecked/topic/462640/plagiarism.

[3] Alan Parker and James O. Hamblen. Computer Algorithms for Plagiarism Detection. *IEEE Transactions on Education*, 32(2):94, 1989.

[4] plagiarism.org. Student Materials, 2015. on-line at: http://www.plagiarism.org/resources/student-materials/.

[5] plagiarism.org. What is plagiarism?, 2015. on-line at: http://www.plagiarism.org/plagiarism-101/what-is-plagiarism/.

[6] Richard A. Posner. *The little book of plagiarism.* Pantheon Books, New York, 2007.

[7] turnitin.com. Turnitin : Results : Plagiarism Spectrum, 2015. on-line at: http://turnitin.com/assets/en_us/media/plagiarism_spectrum.php.