УДК 004.65:004.78 (045)

Артамонов Є.Б., к.т.н.

АНАЛІЗ МЕТОДІВ ПРОТИДІЇ АВТОМАТИЧНИМ СИСТЕМАМ ВИЗНАЧЕННЯ ПЛАГІАТУ В ЕЛЕКТРОННИХ ДОКУМЕНТАХ

Національний авіаційний університет

Розглянуто методи протидії автоматичним системам визначення плагіату. За статистикою майже 70% студентських робіт списано з web-сторінок без посилання на джерела. Багато ВУЗів для боротьби з плагіатом використовують автоматичні або автоматизовані системи перевірки робіт. Стаття розкриває основні проблеми, які пов'язані з "нечесним навчанням", і спираючись на досвід західних та вітчизняних дослідників, аналізує методи виявлення плагіату в роботах студентів та проблеми, які виникають під час перевірки робіт з «прихованим» плагіативним матеріалом.

Вступ

Питання боротьби з плагіатом у вищій освіті є досить актуальним. Але майже всі рішення мають лише локальний розвиток і не йдуть далі одного ВУЗу, а іноді навіть далі однієї кафедри. Масштаби "нечесного навчання" просто вражають і не тільки в Україні чи країнах у минулому СНД, тенденції у розповсюдженні списування (навіть у більш масштабних проявах) можна спостерігати у навчальних закладах всього світу.

На основі статистики відомих систем антиплагіату до 70% студентських робіт списано з web-сторінок без посилання на джерела. У мережі існує близько 1500 англомовних і близько 3000 україно- і російськомовних сайтів, які за невелику суму (або безкоштовно) пропонують реферати, курсові роботи та дипломні проекти. На час сесій відвідуваність таких сайтів становить близько 6,5 млн. відвідувань у день [1, 3].

Подібна практика здобула настільки широке розповсюдження в країнах Західної Європи, США і Канаді, що ректори провідних університетів світу були змушені створити спеціальну Асоціацію під гаслом "За чесне навчання", метою якої є боротьба проти вказаного негативного явища.

Так дослідження хорватських викладачів показали [2], що при написанні творчих робіт тільки 9% студентів не займалися плагіатом взагалі і у 34% ступінь плагіативності була менш, ніж 10%. Спостереження проводилось на протязі двох років і його результати було проаналізовано у різних перетинах і класифікаціях.

Так середня норма плагіату (процент плагіативного тексту) склала 19%, при цьому студенти з високим середнім балом займалися плагіатом менше, ніж студенти з низьким середнім балом, а тип джерела і його складність ніяк не впливала на норму плагіату. Також на норму плагіату у роботах студентів не впливали ані їх стать, ані складність доступного тексту.

Складність визначення поняття плагіату

У енциклопедичних словниках дають наступне визначення: *плагіат* (від лат. *plagio* викрадаю), навмисне привласнення авторства на чужий витвір літератури, науки, мистецтва, винахід або раціоналізаторську пропозицію (повністю або частково). Передбачається кримінальна і адміністративна відповідальність за порушення авторських і винахідницьких прав [1].

Але у наш час навчаючись в школі, інституті, працюючи над новим винаходом чи програмою, обговорюючи новини, ми постійно знаходимося у сфері обговорення чиїхось ідей, думок, висловів.

Безумовно, що будь-яка інтелектуальна праця немислима без використовування досвіду попередніх дослідників. Аналіз різних точок зору, посилання на своїх попередників у вивченні того або іншого питання - звична справа не тільки в студентських, а також і в наукових роботах. Але таке використовування повинне знаходитися в рамках етичних правил і закону, тобто необхідно враховувати, що при використанні чужих ідей необхідно посилатися на автора. Отже, якщо спростити визначення слова, то плагіат - це використання чужих ідей і слів без посилання на джерело інформації [4].

Для запобігання плагіату ми завжди повинні відрізняти авторські розробки від тих, що стали загальновідомими. Де загальновідома інформація – це інформація, яка відома всім чи всім очевидна. Але, якщо конкретизувати поняття всім, то загальновідомою можна вважати інформацію, якщо її або посилання на неї можна знайти принаймні в 5 різних джерелах, або, якщо будь-яка людина може вільно знайти цю інформацію із загальних джерел без особливих зусиль.

Основні причини поширення списування

Визнання самого факту присутності плагіату в роботах студентів не вирішує цієї проблеми. Тому необхідно з'ясувати, що штовхає студентів на таке негативне використання сучасних технологій, для того, щоб прийняти міри по боротьбі з причинами, а не з наслідками, що нажаль є більш поширеним явищем.

Як писав Роберт Харріс [5] студенти - це природні економісти (у розумінні шукачів найпростіших шляхів вирішення проблем). Багато з студентів зацікавленні у найкоротшому можливому проходженні через навчальний курс. Тому для них копіювання чужих робіт виглядає, як спрощення виконання завдання, особливо, якщо такі теми вже виконувалися у минулому.

Окрім цього одною з основних причин списування є нехватка часу у студентів, тому необхідно не тільки завчасно видавати завдання і відслідковувати його поетапне виконання, а також і наголошувати студентам, що основною метою є не виконання завдання, а напрацювання навичок вирішення подібних задач.

Також однією з причин списування можна назвати незацікавленість студентів у вивченні предмету, що набуває найбільшого поширення при вивчанні загальних чи неспеціалізованих дисциплін (так для студентів гуманітарних направлень найчастіше використовують чужі роботи при виконанні завдань по технічним дисциплінам, для технічних спеціальностей подібну тенденцію можна помітити при виконанні гуманітарних дисциплін).

Використання комп'ютерних технологій для розповсюдження плагіату

До поширення Інтернету у розповсюдженні матеріалів для "нечесного навчання" можна було звинуватити нечесних викладачів, які "торгували" готовими роботами чи групу інтелектуально розвинутих людей з необмеженим доступом до бібліографічної літератури, то з появою комп'ютерів використання чужих електронних наробіток стало навіть простішим, ніж переписування з книги.

А з постійною популяризацією Інтернет розповсюдження плагіативних матеріалів набуває ефекту лавини. Коли щомісяця з'являється десятки нових сайтів, які пропонують готові роботи, при цьому половина з них на абсолютно безкоштовній основі.

Але і для тих, хто не має доступу до Інтернету, є можливість після придбання компакт-диску отримати десятки тисяч готових рефератів, які згруповано за тематикою, а іноді і за ВНЗами.

Реферати, контрольні, курсові та дипломні роботи, що представлені як в Інтернеті, так і на компакт-дисках, у більшості випадків мають посередню якість чи носять плагіативний характер. Але легкість їх отримання є одним з основних факторів, який штовхає на використання "чужих" робіт, замість написання власних, але більш якісних.

Масове розповсюдження оргтехніки дозволяє за допомогою скануючих пристроїв за лічені хвилини перевести надрукований текст в електронний. А глобальна електронофікація учбової і наукової літератури призводить до розповсюдження електронних версій підручників, що дозволяє без вичитування в текст за допомогою простого копіювання і вставки "зібрати" реферат чи контрольну роботу майже на будь-яку тему, як виняток залишаються роботи, що потребують математичних розрахунків, але і тут на допомогу приходять комп'ютери з блоками файлів, що обраховуються автоматично на основі вхідних параметрів.

За останніми статистичними дослідженнями в Україні лише 10% населення мають комп'ютери вдома, тому, спираючись на досліди західних вчених (до речі, у країнах західної Європи процентне співвідношення населення, що має вдома комп'ютер вище, ніж в Україні) легко побачити, що попит на безкоштовні чи умовно платні реферати буде тільки зростати – і не тільки в Україні, а також і в усьому світі.

Використання комп'ютерних технологій для боротьби з плагіатом

Певною мірою комп'ютери полегшують плагіат, але вони також дозволяють спростити методи знаходження плагіативних ділянок роботи і визначити ступінь плагіативності роботи в цілому. Взагалі, можна виділити два основних направлення протидій "нечесному навчанню" – запобігання плагіату і його визначення [5].

Так до методів запобігання можна віднести такі:

1) необхідно пояснити студентам, що таке "чиста" робота. Це робота, яка є індивідуальною, унікальною, ця робота повинна повністю розкривати задану тему за передбаченим планом. Такі вимоги пояснюються тим, що у багатьох випадках плагіативність чи неякісність контрольних робіт спостерігається у випадках, коли завдання було незрозумілим для студентів;

2) необхідно забезпечити навчальний процес темами, які будуть змінюватися із року в рік,

також потрібно дозволити студентам розробляти нові теми, які можуть утворюватись шляхом поєднання двох заданих. Саме великий вибір тематик і їх вузька специфікація дозволяють помітно зменшити процент плагіативних робіт;

3) бажано вимагати використання у контрольних роботах ряду основних джерел, наприклад, при написанні реферату однією з вимог може бути використання двох джерел з Інтернет (з зазначенням електронної адреси), двох періодичних видань і двох підручників. При цьому бажано, щоб один з підручників був минулорічного видання, замість періодичних видань можна видати власні матеріали, які допомагають зрозуміти, але не повністю розкривають тематику завдання;

4) при розгляді стандартних тем треба висунути вимогу про наявність у рефераті анотованого звіту про досліджену літературу;

5) у разі потреби організуйте поетапну звітність, тому що однією з причин списування, є неправильно організований час на виконання завдання;

6) у разі можливості (якщо є заплановані практичні заняття) проведіть захист контрольних робіт у вигляді виступу перед аудиторією чи на основі завдань провести тестовий контроль знань, які були підготовлені під час виконання контрольного завдання.

Виявити плагіативність у контрольних роботах чи наукових статях можна наступними методами:

1) при перегляді контрольної роботи зверніть увагу на такі ключові признаки можливої плагіативності документу: різний стиль оформлення цитат чи взагалі відсутність посилань на використані першоджерела, незвичне оформлення документу (різня поля та режими вирівнювання тексту на сторінці), невідповідність тематиці чи її неповне розкриття, посилання на застарілі данні (більше п’яти років), використання застарілих термінів, написання тексту у різних “жаргонних” стилях (коли використані данні з різних джерел, які були підготовлені під різну аудиторію читачів), “грубі помилки” при списуванні (залишені чужі примітки, прізвища авторів, посилання на джерело розповсюдження плагіативних матеріалів – наприклад, “Реферат взято з *www.referat.ru*”). При цьому дані методи перевірки не займають багато часу і не потребують використання першоджерел для оцінки плагіативності;

2) для прискорення аналізу робіт на плагіативність необхідно мати під рукою чи знати місцерозташування основних підборок матеріалів у електронному вигляді – сайти Інтернет (безкоштовні ресурси – *www.5ballov.ru*, *www.referat.kulichki.net* та ін., платні ресурси – *www.referat.mpv.ru*, *www.ronl.ru/zakaz.htm* та ін.), сайти бібліотек чи з доступом до електронних копій видань (*www.baraxolka.ru*, *skachat1.com.ru*, *http://bizkit.land.ru* та ін.), ресурси *CD-ROM* (енциклопедії та бази рефератів на оптичних носіях);

3) для проведення пошуку першоджерела роботи з підозрою на плагіат в Інтернеті можна використовувати стандартні пошукові сервери (www.yandex.ru, google.ru та ін.), а починати пошук з вказування теми чи ключових слів, для звуження кола пошуку можна додавати в кінці запиту слово реферат, завдяки якому будуть відсікатися більшість сайтів, які не містять готових навчальних матеріалів. У разі негативного результату пошуку виділити з тексту роботи “оригінальні” фрази по 5-6 слів і задати пошук по ним. Якщо робота була завантажена із Інтернету, то один із серверів вкаже на першоджерело;

4) останнім методом (а іноді і достатнім) є використання одного з детекторів плагіату. Зараз в Інтернет у вільному доступі є близько десятка програм, які перевіряють текст на плагіат, при цьому деякі працюють тільки з англомовними текстами (*Turnitin* – з можливістю пошуку в Інтернеті, *WCopyfind* – порівняння двох файлів) та інваріантні по відношенню до мови (*TheGlatt Plagiarism Service* – технологія контрольної перевірки, коли студенту необхідно заповнити кожне п’яте слово у свої роботі), до того ж в Інтернеті є ряд сайтів, діяльність яких направлена на боротьбу з плагіатом (*Plagiarism.org* на *http://www.plagiarism.org* – оперативний сервіс, який перевіряє представлені студентські роботи на основі своєї бази даних і формує звіт, а також контролює сайти по розповсюдженню готових контрольних робіт; *Plagiarism Finder* на *http://www.m4-software.com* – пошуковий сервер матеріалів в Інтернеті; *Eve* на *http://www.canexus. com/eve/* – недорогий агент програмного забезпечення, який зрівнює підозрілу роботи з Інтернет-джерелами, а також показує сайт і ступінь співпадання).

Але, як виявилось, “слов’янський” сектор залишився майже сам на сам з проблемою плагіату, тому що якісні продукти виявлення плагіату налаштовані на англомовні тексти і проводять пошук спочатку в англомовному секторі, а іноді і в загалі не розуміють кирилицю. Але і налаштування програм під кирилицю не призведе до гарного результату, тому що у програмах порі-

вняння основним є словник термінів і база даних готових матеріалів, яка накопичується роками.

І якщо у Росії змогли розгорнути проект, який отримав державну підтримку [6], то в Україні подібного глобального проекту нема до сих пір.

Тому групою науковців у порядку особистої ініціативи було створено комп'ютерну систему порівняльного аналізу електронних текстів, яка зараз проходить апробацію в Національному авіаційному університеті.

Ця система призначена для виявлення збігів та протиріч у текстах, написаних українською, російською, англійською, німецькою, французькою мовами. Система не реагує на зміну порядку чергування слів у реченні і вживання синонімів.

Можливими галузями використання системи є:

– освіта (виявлення плагіату і компіляції у рефератах, контрольних, курсових та дипломних проектах і роботах, підручниках та навчальних посібниках);

– наука (виявлення плагіату і компіляції у дисертаціях);

– законотворчість (виявлення збігів та протиріч у законодавчих та нормативних актах України, а також між законодавчими актами України та інших держав);

– діяльність видавництв, інформаційних агентств, засобів масової інформації;

– патентування, інноваційна діяльність, захист інтелектуальної власності.

Але дієвість цієї програми можлива лише у разі розповсюдженням системи серед усіх навчальних закладів України. Що є відповідальним кроком на шляху до "чесного навчання" в Україні.

Протидія роботі системам виявлення плагіату

Поширення автоматичних систем виявлення плагіату сприяло створенню нового виду комп'ютерних програм, які забезпечують генерування тексту, який з високим відсотком достовірності зможе пройти перевірку на плагіативність. Як окрема гілка розвинулись програми автоматичного генерування псевдонаукового тексту [7].

Також популярність даних програм обумовлена використанням пошуковими системами алгоритмів, що подібні виявленню плагіату у тексті. Дана функція застовується в пошукових системах для виявлення унікальності наповнення сайтів.

Одним з направлень у розробці програм по перетворенню текстів стали так звані "синонімайзери". Хоча основною їх функцією є розмноження статей для просування сайту, дані програми з успіхом використовують студенти для обману систем перевірки на плагіат.

Приклад роботи онлайн версії програми представлено на рис. 1.

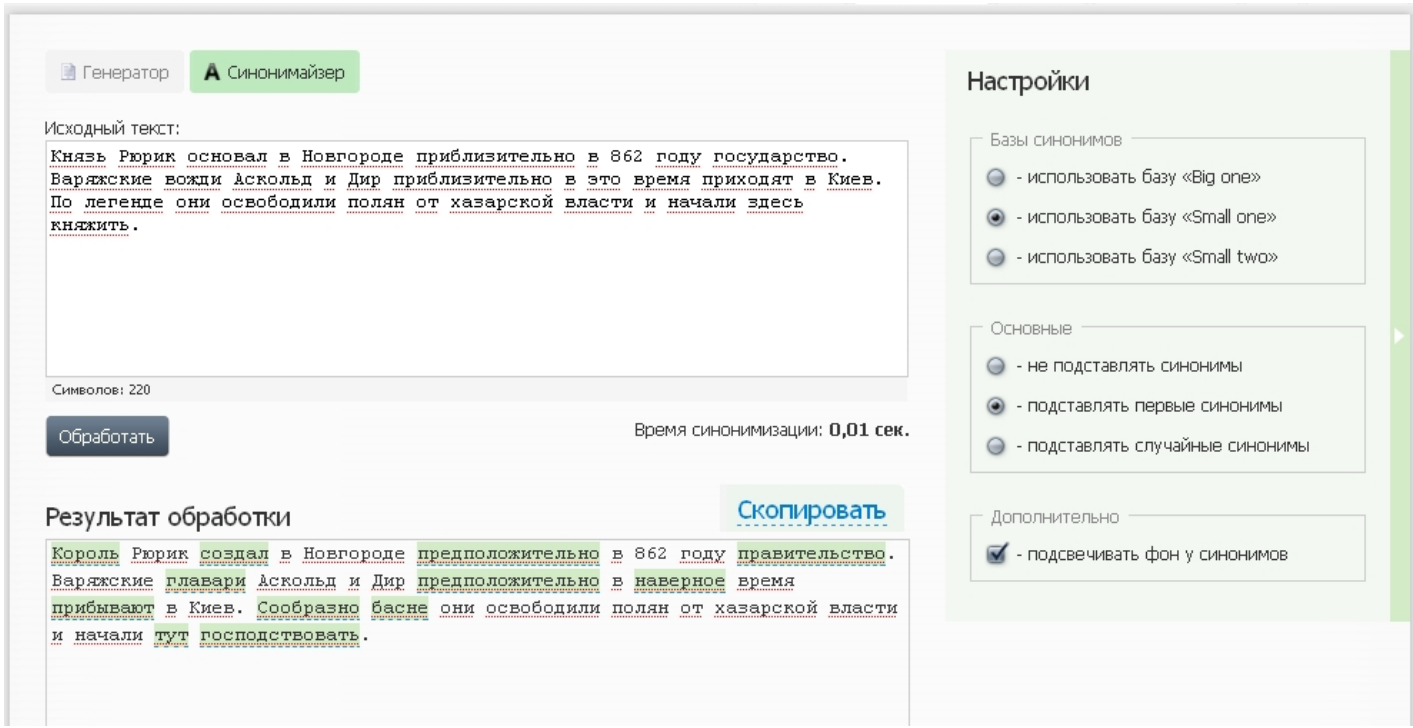

Рис. 1. Робоче вікно і основні налаштування онлайн версії синонімайзеру

Необхідно звернути увагу на те, що при аналізі отриманого тексту можна виявити нелогізми, невідповідність відмінків та інші помилки. Але при автоматичній перевірці даний фрагмент більшістю систем перевірки на плагіат буде помічено, як авторський.

Інша група програм дозволяє сформувати текст на основі шаблону (рис. 2).

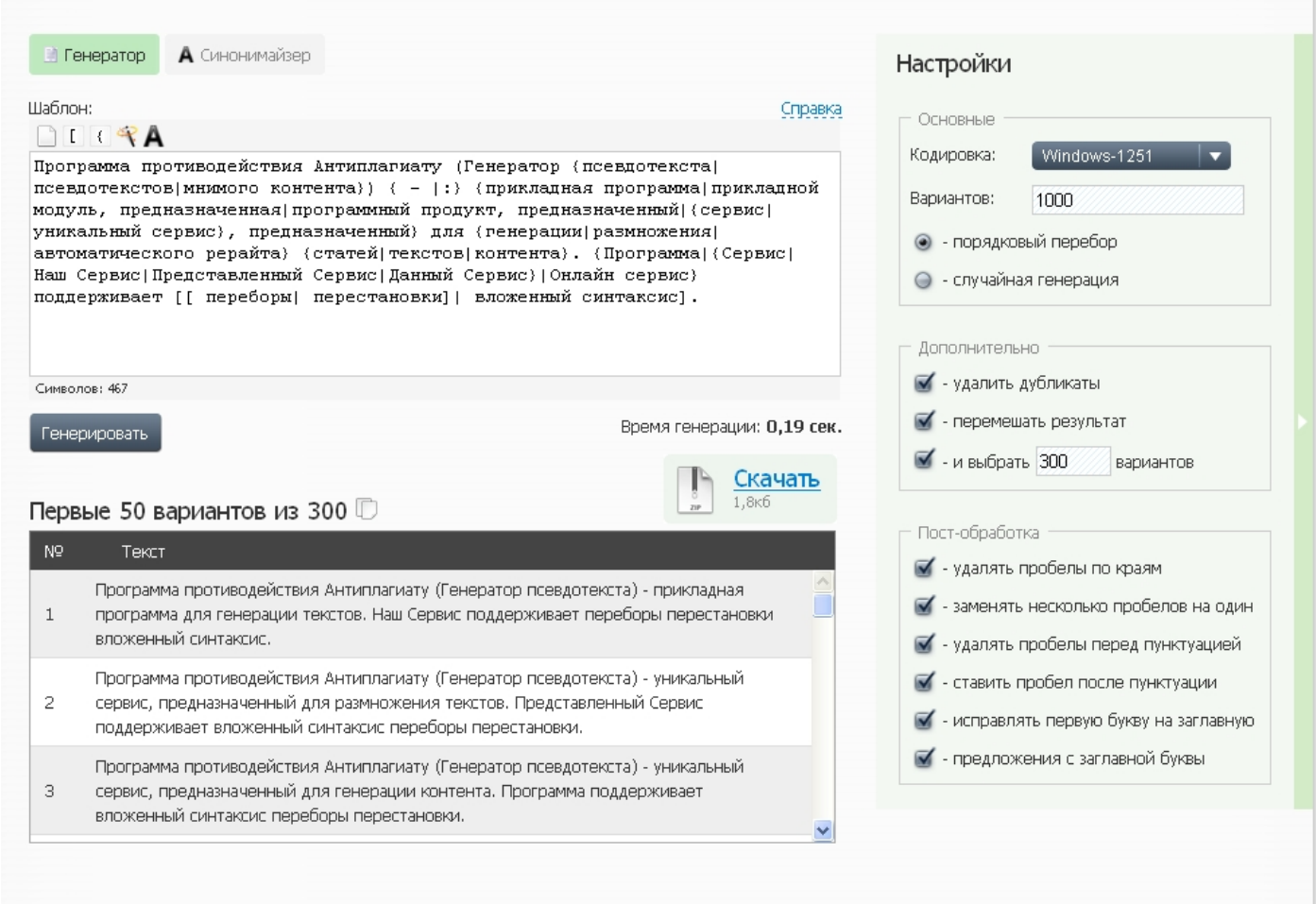

Рис. 2. Робоче вікно програми з автоматичним формуванням тексту на сонові заданого шаблону

Дана програма дозволяє на основі шаблону створювати різноманітний текст за рахунок перестановки, заміни, вибору зі списку вказаних фраз та окремих слів. Після задавання декількох десятків позицій слів, з варіантами заміни по кожній позиції, програма може створити більше сотні варіації відібраного тексту. Даний метод можна визначити, як «направлену варіативну синонімізацію».

Наприклад з конструкції російською мовою (на даний момент в широкому доступі відсутні програми для української мови):

Программа противодействия Антиплагиату (Генератор {псевдотекста|псевдотекстов|мнимого контента}) { - |:} {прикладная программа|прикладной модуль, предназначенная|программный продукт, предназначенный|{сервис|уникальный сервис}, предназначенный} для {генерации|размножения|автоматического рерайта} {статей|текстов|контента}. {Программа|{Сервис|Наш Сервис|Представленный Сервис|Данный Сервис}|Онлайн сервис} поддерживает [[переборы| перестановки]| вложенный синтаксис].

Можна отримати близько 500 унікальних фраз, які при цьому еквівалентні за смислом:

1. Программа противодействия Антиплагиату (Генератор псевдотекста) - прикладная программа для генерации контента. Наш Сервис поддерживает вложенный синтаксис переборы перестановки.

2. Программа противодействия Антиплагиату (Генератор псевдотекста) - уникальный сервис, предназначенный для размножения статей. Программа поддерживает вложенный синтаксис перестановки переборы.

Даний метод вимагає не тільки більше часу але і розуміння автором тематики, для якої

створюється шаблон, тому широкого використання для великих робіт він не набув.

На основі аналізу роботи автоматичних систем перевірки тексту на плагіат були виявлені наступні методи протидії їх роботі (табл. 1). Також до цієї таблиці увійшли: наявність захисту від перелічених методів у розвинених системах [6] (позначка "+" - відповідає наявності захисту у більшості систем, "+/-" – захист використовується у меншості систем, "-" – захист практично відсутній) та можливі рішення для розробників систем виявлення плагіату щодо принципів реалізації даних методів захисту.

Таблиця 1.

Аналіз методів протидії антиплагіат-системам

п/п	Метод протидії	Наявність захисту	Метод захисту
1.	Заміна кириличних літер схожими по виду латинськими: е → e, до → k, м → m, у → y, а → a, х → x, р → p, с → c	+	При підготовці тексту до перевірки всі літери з переліченого ряду приводяться до однієї системи кодування (мови)
2.	Перестановка абзаців	+	Перевірка тексту проводиться за абзацами
3.	Заміна точок комами (розбиття і злиття рядків)	+/-	При перевірці рівня плагіативності необхідно паралельно оцінювати абзацами та речення
4.	Перегрупування рядків в абзацах	+/-	
5.	Перестановка слів у реченнях	+/-	
6.	Заміна пробілів точками	-	Без додаткового аналізу роботи системи дану протидію виявити неможливо, тому що кожен з наведених методів породжує нові слова, які відсутні у базі системи. Для реалізації захисту необхідно вести облік нових слів, які були додані до словника під час роботи з текстом, і у разі перевищення визначеного ліміту зупиняти обробку і передавати роботу експерту-людині
7	Розбиття слів недрукованими або невидимими символами	-	
8.	Заміна слів синонімами	-	Деякі системи використовують заміну всіх слів їх кодовими значеннями з словників, словники можуть використовувати синонімічні ряди та переклади слова іншими мовами. Але дані алгоритми занадто часто видають ознаку плагіату на унікальних текстах, тому в автоматичних системах використовуються дуже рідко

Це аналіз найбільш популярних методів протидії тексту, тому більшість систем виявлення плагіату використовуються у режимі консультанта, тобто визначають підозрілі роботи, але тільки фахівець-людина може дати висновок, щодо плагіативності того або іншого тексту.

Висновки

Впровадження систем виявлення плагіату не може в один момент знищити явище списування. Але постійне використання даних систем змушує змінити точку зору на навчання у багатьох студентів і викладачів. Саме використання автоматичних систем аналізу і перевірки студентських робіт дозволить поширити зону «чесного навчання» на весь світ.

Будь-яка дія викликаэ протидію: так створення антиплагіативних систем запустило роботу над системами, які їм протидіють. І якщо раніше всі маніпуляції для приховування плагіату робились в ручному режимі, то на сьогодні існує сотні програм з власними алгоритмами обробки тексту саме для приховування викраденого тексту від антіплагіативних систем.

В статті було проаналізовано основні методи приховування викраденого тексту:

1) заміна кириличних літер схожими по виду латинськими;

2) перестановка абзаців;

3) заміна точок комами (розбиття і злиття рядків);

4) перегрупування рядків в абзацах;

5) перестановка слів у реченнях;

6) заміна пробілів точками;

7) розбиття слів недрукованими або невидимими символами;

8) заміна слів синонімами.

На основі алгоритмів обробки текстів можна запропонувати наступні методи виявлення замаскованого викраденого матеріалу:

1) при підготовці тексту до перевірки всі літери з переліченого ряду приводяться до однієї системи кодування (мови);

2) при перевірці рівня плагіативності паралельно оцінюються абзаци та речення;

3) проводиться додатковий аналіз роботи системи для ведення обліку нових слів, які були додані до словника під час роботи з текстом;

4) заміна всіх слів їх кодовими значеннями з словників, словники можуть використовувати синонімічні ряди та переклади слова іншими мовами.

Кожен з описаних методів вже має розроблені алгоритми, тому у найближчому часі можна очікувати загострення боротьби між системами виявлення і приховування плагіативних фрагментів текстів.

Список літератури

1. Артамонов Є.Б., Прадідом Ю.Ф. Деякі проблеми використання сучасних комп'ютерних технологій// Вісник Університету внутрішніх справ, 2001. – №13. – С. 212-215.

2) Biliæ-Zulle L., Frkoviæ V., Turk T., Azman J., Petroveèki M. Prevalence of Plagiarism among Medical Students, Croat Med, J 2005, 46 (1), 126-131.

3) Bloomfield L. The importance of writing. Originally published on the Commentary Page of the *Philadelphia Inquirer* on Sunday, April 4, 2004, edited by John Timpane.

4) Жданова Р. Понятие плагиата, IATP Казахстан, http://www.iatp.kz/?id=news&sendnews=1868&lang=2, 02.04.2004

5) Harris R. Anti-Plagiarism Strategies for Research Papers from <http://www.virtualsalt.com/antiplag.htm>

6) Кошиль А. В наших вузах запустили "охотника за плагиатом". Посилання в Інтернеті: http://gorod.dp.ua/news/11156.

6) http://www.antiplagiat.ru – онлайн система виявлення плагіату в електронних текстах.

7) http://pdos.csail.mit.edu/scigen/ - англомовний сайт з можливістю автоматичної генерування псевдонаукових робіт.

8) http://www.seogenerator.ru/tools/ - сайт з автоматичним синонімайзером та генератором варіацій тексту на основі шаблону.