Інформаційні технології

УДК 004.912 **doi: 10.26906/SUNZ.2018.6.083**

Ю. М. Главчева, О. В. Каніщева

Національний технічний університет «Харківський політехнічний інститут», Харків, Україна

ІДЕНТИФИКАЦІЯ ТЕКСТОВОГО ПЛАГІАТУ В АКАДЕМІЧНИХ ДОКУМЕНТАХ

Метою статті є аналіз найбільш розповсюджених технологій виявлення текстового плагіату. **Результати.** Про проблему плагіату у академічних документах свідчать дослідження, які проводяться в різних країнах, у тому числі в Україні. Для виявлення плагіату у наш час активно використовуються інформаційні технології. Складність задачі виявлення плагіату обумовлено тим, що існують різні види плагіату. У статті проаналізовано основні види академічного плагіату. Розглянуто найбільш відоме програмне забезпечення, яке дозволяє виявити фрагменти потенційного плагіату. Системи виявлення плагіату в академічних текстах постійно удосконалюються. Але не зважаючи на це, є деякі аспекти цих систем, які потребують доопрацювання або пошуку нових рішень. Основне завдання –створення системи, яка буде з достатнім відсотком вірогідності орієнтовна на виявлення усіх видів плагіату, тобто буде більш універсальною. Досліджено ознаки визначення авторського стилю для визначення плагіату. Проаналізовано два основні підходи до визначення текстового плагіату, визначені основні переваги і недоліки. **Висновок.** Основна увага на даному етапі приділяється пошуку текстового плагіату. Програмні засоби, які працюють на основі пошуку подібностей дають достатньо точні результати, але виділено ряд факторів та обмежень підходу з пошуку подібностей. З цієї причини активно проводяться дослідження інших підходів. Один з них – ідентифікація плагіату на основі авторського стилю написання роботи. Він ще практично не реалізований, але має потенційні можливості використання. Використання ефективних програмних засобів виявлення можливого плагіату сприятиме імплементації принципів академічної доброчесності в цілому, сприятиме підвищенню рівня якості освітнього та наукового процесів як в Україні, так і у всьому світі.

Ключові слова: плагіат, авторський стиль, академічний плагіат, самоплагіат, академічна доброчесність.

Вступ

Академічна доброчесність є однією з основних засад діяльності закладів освіти. Забезпечити імплементацію принципів академічної доброчесності можна шляхом упровадження комплексу заходів. Основні з них: нормативно-правове регулювання, навчання принципам академічної доброчесності, використання програмних засобів для виявлення плагіату. Великі обсяги інформації, доступні в мережі Інтернет, легко можуть бути використані будь-ким у своїх академічних текстах з посиланням на першоджерело. При цьому автори можуть навмисно або ненавмисно не вказати посилання на першоджерело. Програмні засоби допомагають ідентифікувати фрагменти тексту, які вже присутні в інших документах та можуть бути визнані плагіатом.

Проблема плагіату є актуальною для багатьох країн світу. Про проблему плагіату у академічних документах свідчать дослідження, які проводяться в різних країнах, у тому числі в Україні [1, 2].

На базі Національного технічного університету України «Київський Політехнічний інститут ім. Ігоря Сікорського», протягом лютого-травня 2018 року, було проведено опитування [1]. В опитуванні взяли участь 907 респондентів. Відповіді респондентів на питання з анкети «Наскільки поширені серед Ваших однокурсників/колег наступні прояви наступної поведінки?», вказані в табл. 1.

Відсоток недоброчесних практик є значущим не зважаючи на те, що 77,2% опитуваних вказали, що їм відомо про перевірку студентських академічних текстів на плагіат.

Опитування, проведене у 2018 р. серед магістрантів спеціальності «Дошкільна освіта» в Полтавському національному університеті імені Ю. Кондратюка показало, що на питання «Чому студенти не виконують роботи самостійно?», – 78%, респондентів відповіли: «Через вільний доступ їх у INTERNET» [3]. Величезна кількість легко доступних електронних текстів, інформаційні програми та інструменти дозволяють легко і швидко модифікувати інформацію.

Таблиця 1 – **Відповіді респондентів на питання: «Наскільки поширені серед Ваших однокурсників/колег наступні прояви наступної поведінки?»**

Питання анкети	A, %	B, %	C, %	D, %
Самоплагіат – оприлюднення (частково або повністю)власних раніше опублікованих наукових результатів як нових	13,12	20,29	35,17	31,42
Компіляція – створення значного масиву тексту без поглибленого вивчення проблеми шляхом копіювання тексту	29,66	31,09	28,00	11,25
Парафраза – переказ своїми словами чужих думок, ідей або тексту; сутність парафрази полягає в заміні слів (знаків)	29,17	35,17	26,57	8,49
Представлення суміші власних і запозичених аргументів без належного цитування	27,67	37,82	26,24	8,7
Внесення незначних правок у скопійований матеріал (перефразування речень, зміна порядку слів у них тощо) та без належного оформлення цитування	23,48	41,02	27,01	8,49

Примітка: A – дуже поширено; B – швидше поширено; C – швидше не поширено; D – взагалі не поширено

Експерту важко визначити можливий плагіат без спеціальних інструментів. Для виявлення плагіату у наш час активно використовуються інформаційні технології. Складність задачі виявлення плагіату обумовлено тим, що існують різні види плагіату. Програмне забезпечення має визначати плагіат будь-якого виду в різних умовах. Саме тому, пошук ефективних підходів до виявлення академічного плагіату є актуальним завданням.

Академічний плагіат, його види

Визначення поняття «Академічний плагіат» можна знайти у двох типах документів: нормативно-правові документи; публікації науковців та розробників програмних засобів для виявлення плагіату.

У документі [4] вказані такі види академічного плагіату в наукових роботах:

1. Відтворення в тексті наукової роботи без змін, з незначними змінами, або в перекладі тексту іншого автора (інших авторів), Відтворення в тексті наукової роботи без змін, з незначними обсягом від речення і більше, без посилання на автора (авторів) відтвореного тексту.

2. Відтворення в тексті наукової роботи, повністю або частково, тексту іншого автора (інших авторів) через його перефразування чи довільний переказ без посилання на автора (авторів) відтвореного тексту.

3. Відтворення в тексті наукової роботи наведених в іншому джерелі цитат з третіх джерел без вказування, за яким саме безпосереднім джерелом наведена цитата.

4. Відтворення в тексті наукової роботи наведеної в іншому джерелі науково-технічної інформації (крім загальновідомої) без вказування на те, з якого джерела взята ця інформація.

5. Відтворення в тексті наукової роботи оприлюднених творів мистецтва без зазначення авторства цих творів мистецтва.

Неоднозначність поняття «плагіат» також проявляється у визначенні та різноманітті його видів, описаних в дослідницькій літературі [5]. Плагіат може проявлятися у різних формах. Існує два основних типи плагіату: текстовий плагіат, плагіат вихідного коду [5]. Текстовий плагіат зазвичай спостерігається в освіті та наукових дослідженнях. Авторами публікації [5] визначається сім його видів (табл. 2).

Для текстового плагіату можна використовувати програмне забезпечення для обробки природньої мови та спеціальні алгоритми для визначення ознак кожного виду плагіату. Окремо треба визначити, що плагіат ідей є зараз невирішеною задачею.

Таблиця 2 – **Види плагіату та їх опис**

Назва	Опис
Навмисне копіювання (створення клонів)	копіювання інших творів та їх представлення у якості свої власної роботу без посилання на оригінальне джерело
Перефразування	перефразування або мозаїка з різних фраз без посилання на оригінальне джерело
Метафора	використання метафор для представлення іншої ідеї без посилання на оригінальне джерело
Ідея	ідея або рішення запозичені з інших джерел і претендують як власний у дослідницькій роботі
Самоплагіат	у цій формі автор використовує свої попередні публікації
Помилка 404	некоректне використання джерел: неправильних або неіснуючих
Повторення	автор наводить посилання на правильне джерело, але його текст дуже схожий на оригінал

Програмне забезпечення для виявлення академічного плагіату

Для виявлення можливого плагіату створена значна кількість програмних засобів, найвідоміші з яких: Turnitin, Strike Plagiarism, Unicheck, Urkund, Advego Plagiatus, eTXT, Антиплагиат тощо [6]. Короткі характеристики найбільш відомих і розповсюджених програмних продуктів щодо виявлення у текстах плагіату та запозичень розглядаються у інформаційному огляді [7].

Сервіси для перевірки текстів на унікальність працюють, переважно, за однаковим алгоритмом. Поточні дослідження в області автоматичного виявлення плагіату для текстових документів зосереджені на алгоритмах, які порівнюють досліджувані документи з потенційними оригінальними документами. Вони демонструють високу точність виявлення текстових запозичень за умови, коли оцифрований оригінальний документ присутній у колекції, за якою проводиться перевірка. Усі перелічені програми основані на підході, який використовує алгоритм порівняння на подібності. Підхід має обмеження. Він потребує великої бази даних текстів певною мовою для пошуку подібностей. До кінця не вирішеною є завдання точного порівняння зображень, схем, креслень.

Паралельно ведеться пошук підходів, які не мають вказаних обмежень. Наприклад, пошук плагіату без використання електронної колекції. У роботі [8] автори називають клас завдань виявлення потенційного плагіату лише на основі аналізу стилю письма без використання зовнішніх текстів, класом виявлення внутрішнього плагіату (class intrinsic plagiarism). Таким чином, проблема визначення плагіату без електронної колекції досліджується багатьма дослідниками.

Порівняння підходів до виявлення плагіату

Визначення авторського стилю при написанні текстів може бути використане при вирішенні великої кількості практичних завдань у різних сферах діяльності: лінгвістичні дослідження, визначення оригінальності тексту, робота служб безпеки, криміналістиці та інше. Порівняємо деякі характеристики двох підходів: пошуку подібностей та визначення авторського стилю написання роботи (табл. 3).

Таблиця 3 – **Порівняння деяких характеристик двох підходів**

Підхід	пошук подібностей	визначення авторського стилю написання роботи
База даних	формування та зберігання великих колекцій електронних текстів	необхідно декілька оригінальних документів автора
Правовий аспект	система завантажує тексти, що перевіряє	система зберігає формальні показники, які характеризують авторський стиль написання роботи
Обмеження	фрагмент, який скопійовано, відсутній в колекції електронних текстів	відсутній оригінальний текст автора

Опишемо основні відмінності між підходами для виявлення можливого плагіату.

По-перше для пошуку подібностей необхідно мати колекцію електронних видань мовою, за якою проводиться порівняння. Для зберігання колекцій електронних текстів або індексної бази даних необхідно мати достатній власний фізичний простір для зберігання та забезпечувати швидкі комунікації з іншими зовнішніми ресурсами. Наприклад, Turnitin працює порівнюючи академічні тексти студентів з матеріалами, що містяться в його великих базах даних, які включають понад 45 мільярдів веб-сторінок, 130 мільйонів академічних статей і 337 мільйонів студентських робіт (Turnitin, 2014 р.). В основному це англійська мова [9]. У 2017 році Міністерство освіти та науки України прийняло рішення про початок робіт з формування Національного репозитарію академічних текстів. Пункт 3 «Положення про Національний репозитарій академічних текстів» основну мету ресурсу: «Основною метою Національного репозитарію є сприяння розвитку освітньої, наукової, науково-технічної та інноваційної діяльності шляхом поліпшення доступу до академічних текстів та сприяння академічній доброчесності» [10]. Ресурси Національного репозитарію мають стати допоміжними засобами проведення експертизи академічних текстів на плагіат. А якщо документ, з якого проведено незаконне запозичення не має електронної форми представлення та не представлений в колекції, то пошук плагіату буде неуспішним. Тому ведуться дослідження методів, які будуть менше залежати від складу за розмірів БД для пошуку подібностей.

По-друге, робота, завантажена для перевірки може зберігатися у системі, яка виконує пошук можливого плагіату. Наприклад, за замовчуванням Turnitin зберігає студентські роботи в своїх бази даних. Саме тому, інтелектуальна власність студентів зберігається в програмному забезпеченні Turnitin і це є окремою етичною проблемою. [9].

Інакше, якщо пошук можливого плагіату проводиться на основі методу визначення авторського стилю написання роботи, то для подальшої роботи системі необхідно зберігати формальні показники, що характеризують авторський стиль написання тексту. Ці ж показники розраховуються для будь-якого тексту та порівнюються з еталоном (характеристиками певного автора).

По-третє, стрімко зростає кількість та об’єм електронних текстів за якими необхідно проводити пошук подібностей. Вже зараз перевірка може зайняти від декількох хвилин до годин. Для збереження такої швидкості при проведенні перевірки необхідно підтримувати робочі характеристики системи у відповідному стані. Підвищуються вимоги і до клієнтських робочих місць. У порівнянні з цим розрахунок показників та порівняння показників є менш ресурс ємким процесом.

Дослідження письмового стилю автора документу проводять за основними стилістичними ознаки, за якими можна визначити характеристики тексту [11]: лексичний аналіз; символьний аналіз; синтаксичний аналіз; семантичний аналіз; особливості застосування. Для визначення авторського стилю використовують комбінацію показників за різними стилістичними ознаками. Слід зазначити, що одні й ті ж самі стилістичні ознаки для різних мов демонструють різні відсотки вірогідності авторства. Це зумовлено особливостями певної мови.

Щодо використання стилю письма для визначення плагіату проведено багато досліджень, але створені вченими програмні засоби в основному не доступні для загального використання. За даними розробників, вірогідність правильного визначення програмою автора для текстів англійською складає 85%. Для даного підходу також є завдання, які ще необхідно вирішити: зміна авторського стилю написання з часом; документи з кількома авторами; навмисна зміна стилю написання автором; особливості письмового стилю у текстах різних тематичних напрямів; визначення характеристик лінгвістичного корпусу для проведення досліджень (загальна кількість та об’єм документів у наборі тренувань).

Висновки

Системи виявлення плагіату в академічних текстах постійно удосконалюються. Але не зважаючи на це, є деякі аспекти цих систем, які потребують доопрацювання або пошуку нових рішень. Основне завдання – створення системи, яка буде з достатнім відсотком вірогідності орієнтовна на виявлення усіх видів плагіату, тобто буде більш універсальною [12].

Основна увага на даному етапі приділяється пошуку текстового плагіату. Програмні засоби, які працюють на основі пошуку подібностей демонструють достатньо точні результати. Але прослідковується ряд факторів та обмежень підходу з пошуку подібностей. З цієї причини активно проводяться дослідження інших підходів. Один з них – ідентифікація плагіату на основі авторського стилю написання роботи. Він ще практично не реалізований, але має потенційні можливості використання.

Використання ефективних програмних засобів виявлення можливого плагіату сприятиме імплементації принципів академічної доброчесності в цілому, сприятиме підвищенню рівня якості освітнього та наукового процесів як в Україні, так і у всьому світі.

СПИСОК ЛІТЕРАТУРИ

1. Академічна доброчесність, результати онлайн-опитування студентів та співробітників КПІ ім. Ігоря Сікорського [Электронный ресурс]. – Режим доступа: http://ela.kpi.ua/bitstream/123456789/23076/1/Akademichna_dobrochesnist.pdf.
2. Епідемія академічного плагіату в цифрах [Электронный ресурс]. — Режим доступа: http://studway.com.ua/plagiat-2.
3. Вашак О.О. Академічна доброчесність: виклики сучасності / О. О. Вашак, Ю. Гринь Ю. // Академічна доброчесність: виклики сучасності, Республіка Польща, Варшава, 1–13.10.2018). — Варшава, 2018. – С. 93-98.
4. Про Рекомендації щодо запобігання академічному плагіату та його виявлення в наукових роботах [Электронный ресурс]. — Режим доступа: http://osvita.ua/legislation/Vishya_osvita/61647.
5. Chowdhury H. A., Bhattacharyya D. K. Plagiarism: Taxonomy, Tools and Detection Techniques //arXiv preprint arXiv:1801.06323. – 2018.
6. Академічна чесність як основа сталого розвитку університету / Міжнарод. благод. Фонд "Міжнарод. фонд. дослідж. освіт. політики" ; за заг. ред. Т. В. Фінікова, А. Є. Артюхова – К.; Таксон, 2016. – 234 с.
7. Програмне забезпечення для перевірки наукових текстів на плагіат : інформаційний огляд / автори-укладачі: А. Р. Вергун, Л. В. Савенкова, С. О. Чуканова ; редколегія: В. С. Пашкова, О. В. Воскобойнікова-Гузєва, Я. Є. Сошинська; Українська бібл. асоціація. – Київ: УБА, 2016. – Електрон. вид. – 1 електрон. опт. диск (CDROM). – 36 с.
8. Eissen, SMZ (Eissen, Sven Meyer zu); Stein, B (Stein, Benno) Intrinsic plagiarism detection Серия книг: LECTURE NOTES IN COMPUTER SCIENCE Том: 3936 Стр.: 565-569 Опубликовано: 2006
9. Samuel Bruton & Dan Childers (2016) The ethics and politics of policing plagiarism: a qualitative study of faculty views on student plagiarism and Turnitin®, Assessment & Evaluation in Higher Education, 41:2, 316-330, DOI: 10.1080/02602938.2015.1008981
10. ПОЛОЖЕННЯ про Національний репозитарій академічних текстів [Электронный ресурс]. — Режим доступа: http://zakon2.rada.gov.ua/laws/show/541-2017-п
11. Stamatatos, E. A survey of modern authorship attribution methods [Огляд сучасних авторських методів атрибуції]. Journal of the Association for Information Science and Technology. 2009. 60(3). pp.538-556.
12. Křížková, Š., Tomášková, H., & Gavalec, M. (2016, July). Preference comparison for plagiarism detection systems. In Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on (pp. 1760-1767). IEEE.

Рецензент: д-р техн. наук, проф. К. С. Козелкова,
Державний університет телекомунікацій, Київ
Received (Надійшла) 17.10.2018
Accepted for publication (Прийнята до друку) 05.12.2018

Идентификация текстового плагиата в академических документах

Ю. Н. Главчева, О. В. Канищева

Целью статьи является анализ наиболее распространенных технологий обнаружения текстового плагиата. **Результаты**. О проблеме плагиата в академических документах показывают исследования, проводимые в различных странах, в том числе в Украине. Для выявления плагиата в наше время активно используются информационные технологии. Сложность задачи выявления плагиата обусловлено тем, что существуют различные виды плагиата. В статье проанализированы основные виды академического плагиата. Рассмотрены наиболее известное программное обеспечение, которое позволяет обнаружить фрагменты потенциального плагиата. Системы обнаружения плагиата в академических текстах постоянно совершенствуются. Несмотря на это, есть некоторые аспекты этих систем, требующих доработки или поиска новых решений. Основная задача – создание системы, которая будет с достаточной вероятностью ориентировона на выявление всех видов плагиата, то есть будет более универсальной. Исследованы признаки определения авторского стиля для определения плагиата. Проанализированы два основных подхода к определению текстового плагиата, определены основные преимущества и недостатки. **Вывод.** Основное внимание на данном этапе уделяется поиску текстового плагиата. Программные средства, которые работают на основе поиска сходств, дают достаточно точные результаты, но выделен ряд факторов и ограничений подхода по поиску сходств. По этой причине активно проводятся исследования других подходов. Один из них – идентификация плагиата на основе авторского стиля написания работы. Он еще практически не реализован, но имеет потенциальные возможности использования. Использование эффективных программных средств выявления возможного плагиата способствует имплементации принципов академической добропорядочности в целом, будет способствовать повышению уровня качества образовательного и научного процессов как в Украине, так и во всем мире.

Ключевые слова: плагиат, авторский стиль, академический плагиат, самоплагиат, академическая добродетель.

Identification of text plagiates in academic documents

Yu. Glavcheva, O. Kanishcheva

The purpose of the article is to analyze the most common technologies for detecting text plagiarism. **Results** Research on the problem of plagiarism in academic documents is being conducted in different countries, including Ukraine. Today, information technology is actively used for detecting plagiarism. The complexity of detecting plagiarism is due to the fact that there are different types of plagiarism. The article analyzes the main types of academic plagiarism. The most well-known software, which allows to identify fragments of potential plagiarism, is considered. Plagiarism detection systems in academic texts are constantly being improved. But despite this, there are some aspects of these systems that need to be improved. The main task - the creation of a system, which will identify all typed of plagiarism with a sufficient percentage, and will be more multiporpose. The signs of determining the author's style for determination of plagiarism are investigated. Two main approaches to the definition of text plagiarism are analyzed, the main advantages and disadvantages are defined. **Conclusion.** At this stage, the focus is on finding text plagiarism. Software tools that are based on the search for similarities give fairly accurate results, paid attention to a number of factors and limitations in the search for similarities. For this reason, other approaches are being actively pursued. One of them is the identification of plagiarism based on the author's writing style. It is still practically not implemented, but has potential uses. The use of effective software tools to detect possible plagiarism will contribute to the implementation of the principles of academic integrity , will contribute to increasing the quality of educational and scientific processes both in Ukraine and throughout the world.

Keywords: plagiarism, writing style, academic plagiarism, self-plagiarism, academic integrity.