УДК 004.012.82:378.16

д.т.н., проф. **Шостак І.В.**
(НАУ ім. М.Є. Жуковського, "ХАІ", Харків)
Груздо І.В.
(НАУ ім. М.Є. Жуковського, "ХАІ", Харків)

КОМП'ЮТЕРИЗАЦІЯ ПРОЦЕСУ ВИЯВЛЕННЯ ПЛАГІАТУ У СТУДЕНТСЬКИХ РОБОТАХ

У статті викладена інформаційна технологія аналізу робіт студентів технічних ВНЗ на предмет виявлення в них текстових запозичень. Здійснено огляд і критичний аналіз існуючих програмно-інструментальних засобів аналізу природно-мовних текстів (на російській і українській мовах) в аспекті їхньої придатності до виявлення плагіату в студентських роботах. Запропоновано підхід до оцінювання якості рішень, які сформовано програмним засобом у випадку виявлення плагіату. Обґрунтовано комплексний критерій якості студентської роботи в аспекті наявності в ній текстових запозичень.

Ключові слова: плагіат, студентська робота, програмний засіб, інформаційна технологія, якість рішень.

Вступ. Результати критичного аналізу існуючого програмного забезпечення [1, 2], призначеного для виявлення текстових запозичень у роботах студентів дають можливість стверджувати, що на цей момент на ринку програмних продуктів країн СНД не представлено жодної подібної системи, яка б мала прийнятний рівень ефективності. Проте, найбільше поширення в Україні на цей момент мають програми Антиплагіат і WCopyfind. Однак, дані програмні засоби з самого початку були призначені для аналізу природно-мовних текстів, тому вони не можуть бути використані для виявлення запозичень у роботах студентів технічних ВНЗ, оскільки ці роботи в більшості випадків містять у собі фрагменти текстів з різних семіотичних систем (математичні структури, вихідні коди програм і т.ін.).

Принципова вада більшості систем щодо виявлення запозичень у студентських роботах полягає в їхньому функціонуванні в Web-середовищі, що приводить до низької захищеності баз даних таких робіт. Іншим істотним недоліком таких систем є відсутність у них засобів кількісної оцінки припустимих обсягів цитування з урахуванням значущості цитуємого фрагмента як структурної частини тексту [1].

Зазначені вище обставини обумовлюють актуальність задачі розробки ефективного інформаційного інструментарію для контролю якості виконання академічних робіт на предмет виявлення плагіату й, як наслідок, розширення можливостей моніторингу навчального процесу.

Мета статті полягає у викладенні основних етапів спеціалізованої інформаційної технології виявлення плагіату в студентських роботах, а так само принципів функціонування програмного засобу, який реалізує зазначену технологію.

Постановка задачі створення інформаційної технології виявлення текстових запозичень. Змістовна постановка загальної задачі створення інформаційної технології містить в якості вихідних даних множину студентських робіт різних типів, які виконано на поточний момент у технічному ВНЗ, текст яких занесено до спеціалізованої бази даних .

У ході реалізації інформаційної технології необхідно отримати для кожного типу студентської роботи набір засобів автоматизації процесів текстологічного аналізу для встановлення факту запозичення й фрагменту запозиченого тексту з множини наявних студентських робіт. При цьому зазначені засоби повинні забезпечувати роботу з усіма типами мовних об'єктів, які мають місце в роботах студентів технічного ВНЗ. При цьому під мовним об'єктом розуміється текст або його фрагмент із встановленим типом структури.

Для рішення зазначеної задачі необхідно: провести класифікацію навчальних робіт студентів технічних ВНЗ, що навчаються на різних спеціальностях (даний етап припускає формування номенклатури робіт і створення класифікатора по обсягу й контенту); конкретизувати особливості різних типів студентських робіт; обґрунтувати значення

граничних параметрів допустимості текстологічних запозичень стосовно структурних частин основних типів студентських робіт; розробити програмний засіб (ПЗ) для виявлення текстологічних запозичень у роботах студентів технічних ВНЗ, реалізувати його на ПЕВМ для підтримки прийняття нечітких ідентифікаційних і прогнозних рішень в умовах невизначеності з метою апробації підтвердження ефективності етапів спеціалізованої інформаційної технології; сформулювати й вирішити за допомогою розробленого ПЗ тестові задачі підтримки прийняття рішень при аналізі й оцінці студентських робіт; провести експериментальні дослідження адекватності й ефективності інформаційної технології виявлення текстових запозичень у роботах студентів технічних ВНЗ.

Особливості класу практичних задач прийняття рішень при оцінюванні робіт студентів. Процес аналізу й оцінювання студентської роботи у вищих навчальних закладах являє типовий приклад недостатньо формалізованої задачі прийняття рішень, для відсутності надійні кількісні моделі й закономірності, що описують зв'язок конкретної навчальної роботи зі структурними частинами й можливими запозиченнями в них, що спричиняє неточність при прийнятті рішень. У таких випадках лише інтуїція, досвід і вміння викладача як експерта у своїй галузі дозволяють прийняти адекватне рішення про оцінку роботи в аспекті наявності в ній запозичень.

Набір умінь викладача, заснований на його набутому досвіді й «чуття», на явних і неявних знаннях, надає йому можливість ефективно вирішувати недостатньо формалізовані задачі прийняття рішень щодо оцінювання студентських робіт в аспекті наявності в них текстових запозичень.

При аналізі й оцінюванні студентських робіт на предмет наявності в них плагіату необхідно враховувати таку множину факторів: специфіку студентської роботи, причини запозичень, значимість структурних частин роботи, у яких були визначені запозичення, показники припустимості текстологічних запозичень стосовно структурних частин. При цьому необхідно враховувати, що якість роботи в цілому вимірюється у різних шкалах.

На практиці аналіз і оцінка студентських робіт багато в чому здійснюється суб'єктивно, тому що викладач не в змозі оглянути й урахувати при ухваленні рішення усі наведені вище чинники. Це обумовлює необхідність збереження знань і вмінь досвідчених викладачів у БД для наступного їхнього використання для підтримки прийняття рішень менш досвідченими викладачами.

Аналіз досліджень і існуючих рішень щодо комп'ютерізації процесу виявлення плагіату у студентських роботах. Існує широкий спектр пошукових систем для визначення текстологічних збігів у різного роду документах. З урахуванням специфіки таких систем кожну з них можна віднести до одного з наступних класів: утиліти статистичного аналізу тексту; утиліти лінгвістичного аналізу тексту (морфологія, синтаксис) (системи обробки природної мови; Системи обробки машинної мови).

Виходячи з особливостей об'єкта нашого аналізу, а саме студентських робіт, надалі будимо розглядати системи обробки природної мови. Типовим програмним засобом цього типу є TurnItIn.

TurnItIn - це комерційна розробка американської компанії iParadigms [3]. Система, перевіряє документи на наявність у них некоректних запозичень із різноманітних текстів які подано природною мовою. Пошукова база даних TurnItIn містить близько 8 млрд веб-сторінок і більше 4,5 млрд рефератів і курсових, десятки тисяч наукових праць і аналітичних статей з відкритих і закритих джерел, а так само всі студентські роботи, які коли-небудь проходили перевірку. Час обробки запиту на перевірку однієї роботи може становити кілька днів. По закінченні перевірки користувачеві надається звіт про оригінальність роботи із вказівкою у відсотках частки текстологічних запозичень.

У цей же час правомірність використання служби TurnItIn піддається сумніву [4]. Система включає всі академічні роботи, що перевіряються, у свою базу даних, одержуючи при цьому економічну вигоду без виплати компенсації авторам. Служба, призначена для виявлення випадків присвоєння авторства, сама стає інструментом порушення прав

інтелектуальної власності.

Оскільки система TurnItIn оперує винятково електронними текстами, набраними латиницею, проблематична її адаптація до аналізу робіт українській і російській мовах.

"Детектор плагіату" [5] - комп'ютерна система пошуку порушення авторських прав, орієнтованих на пошук плагіату в Інтернеті. Після проведення порівняння користувачеві надається можливість переглянути докладну інформацію про знайдені фрагменти збігу. Дана система орієнтована на роботу з російськомовними текстами, однак головним її недоліком є не ергономічний інтерфейс.

Комп'ютерна система Плагіат-Інформ [6], у порівнянні з іншими комп'ютерними системами розглянутого класу має безсумнівну перевагу, перед іншими засобами виявлення плагіату, оскільки спеціально розроблена для визначення наявності факту плагіату в студентських роботах, крім того інформаційна база поєднує інформаційний простір безлічі ВНЗ і, відповідно, забезпечує незалежність від територіального чинника, забезпечуючи при цьому швидкий і зручний доступ до робіт, що одночасно захищенні в межах цього простору. Одним з недоліків PlagiatInform є не можливість роботи з українськомовними текстами.

Система «Антиплагіат» [7] орієнтована, головним чином, на пошук плагіату в студентських роботах, для найбільш ефективного рішення цієї задачі розроблена спеціальна версія системи «Антиплагіат.ВУЗ». Безумовною перевагою даної системи є можливість надання двох режимів - швидкого й детального. Аналіз документа в системі «Антиплагіат» завершується видачею докладного звіту, у якому є можливість переглянути детальну інформацію про знайдені фрагменти збігу [8]. Підтримуваний формат файлів ASCII.

Недостатня ефективність розглянутої системи проявляється по перше в порівняно низькій швидкості аналізу текстів (судячи з інформації, представленої на офіційному сайті [7], швидкість перевірки, даної програми, становить 800 сторінок (формату А4) у секунду (при незазначеному ступені заповнення бази даних)), і по друге, в істотному ризику нерозпізнання плагіату, якщо аналізований текст являє собою повну копію, одне з документів в інформаційній базі.

Advego Plagiatus [8] - комп'ютерна система в якій реалізовано два методи аналізу тексту - простий і глибокий. Перший метод використовується програмою за замовчуванням і працює трохи швидше. Другий спосіб більш точний, але вимагає більше часу на пошук схожих фраз і словосполучень в Інтернеті.

Для розглянутої системи характерна висока швидкість обробки текстів залежно від застосовуваного методу (від однієї до декількох хвилин). Advego Plagiatus у більшості випадків точно ідентифікує плагіат [9]. Недоліком системи є те, що результат її роботи обчислюється у відсотках унікальності контента. Цей показник лише побічно вказує на текстологічні запозичення оскільки не несе в собі інформації про джерело плагіату.

Лінгвоаналізатор [10] - комп'ютерна система, спеціально розроблена для визначення наявності плагіату в тексті, що належить певному авторові. Обраний текст система порівнює за його стильовими характеристиками з текстами усіх авторів у кожної подвиборці, після чого видається аргументований висновок про близькість тексту тому або іншому авторові, а також до текстів обраних авторів. Після проведення порівняння користувачеві надається можливість переглянути докладну інформацію про знайдені фрагменти збігу.

Недостатня ефективність розглянутої системи проявляється по перше в її порівняно низької надійності (самі розроблювачі вказують [11], що пошук плагіату за допомогою аналізу стилю документа може дати негативний результат при безумовній наявності плагіату), по друге, відсутня можливість роботи з текстами, обсяг яких менш 500 аркушів, у третіх, неможливістю роботи з українськомовними текстами.

Програма EVE2 є комерційним додатком [12] і являє собою інтерфейс до пошукової машини Google. Відповідно до огляду [13] в EVE2 реалізована можливість обробки документів різних форматів, у яких зберігається значна частина матеріалу в Інтернеті: HTML, ASCII. До недоліків EVE2 слід віднести, те що пошук виконується тільки по Web-контенту у форматі HTML і ASCII, а більша частина матеріалу всесвітній павутині

зберігається в інших форматах [14]. Оскільки EVE2 оперує винятково з електронними текстами, набраними латиницею, проблематична її адаптація до аналізу робіт на українській і російській мовах.

Plagiarism-Finder - дозволяє перевірити текст на предмет наявності текстових збігів з документами, що зберігаються в Інтернеті [15]. У цій системі процедура пошуку достатньо швидка й не перевищує декількох хвилин. Після перевірки документа, користувач одержує докладний HTML звіт, у якому представляються результати, і виділені в тексті підозрілі абзаци так само є посилання на джерела запозичень.

Одним з недоліків Plagiarism-Finder є можливість роботи тільки з точними збігами, що складаються із семи й більше слів, але при цьому він не працює, з меншими фрагментами тексту. Крім того, істотним недоліком розглянутої системи є необхідність підключення до Інтернету.

CopyCatch Gold розроблений програмістом британської фірми CFL Software Developments Д. Вулсом [16], підтримує багато популярних форматів, однак, обробка російськомовних документів у форматі MS Word ведеться не коректно. Про це свідчать результати випробувань демоверсії продукту, яка доступна для завантаження на офіційній Web-сторінці [17].

Спеціальна інтернет-служба Tumitin є лідером на ринку по виявленню плагіату [14]. Як джерела Tumitin використовує Інтернет-ресурси. У випадку виявлення текстових збігів система видає попередження. При цьому програма не готує висновки, чи є перевірена робота плагіатом, а лише виділяє кольором в аналізованому тексті ті місця, які є повтореннями з інших джерел. При цьому в системі ведеться облік відсотка таких повторень стосовно загального обсягу аналізованої роботи.

Принциповими недоліками Tumitin є те, що вона, будучи здатна аналізувати документи, складені на кожній з більшості європейських мов, по перше допускає при цьому значні погрішності й, по друге, не адаптована для російської й української мов.

Автономний засіб (desktop application) WordCHECK [18] орієнтовано, головним чином, на пошук плагіату в студентських роботах. Всі документи, що, завантажуються в базу даних, де вони перевіряються з метою виявлення копіювання в межах навчальної групи. Аналіз документа в WordCHECK завершується видачею докладного звіту, у якому є можливість переглянути докладну інформацію про знайдені фрагменти збігу. Дана система не адаптована для російської й української мов, однак головним її недоліком є неергономічний інтерфейс.

WCopyFind - продукт, розроблений професором Л. Блумфілдом в університеті штату Віргінія (США), є одним із самих перших інструментів для виявлення фактів списування студентами [19], має дружній графічний інтерфейс і просту функціональність. Аналіз документа завершується видачею докладного звіту, у форматі HTML з інформацією про знайдені фрагменти збігу й інтернет посиланнями на них. До значних недоліків даного продукту варто віднести відсутність можливості пошуку збігів в Інтернеті й не здатність роботи з українськомовними текстами.

Plagiarism Advisory Service є національною комп'ютерною системою щодо визначення плагіату у Великобританії. Важлива особливість детектора полягає в тім, що він дає інформацію про посилання на Web-Сайти, де перебувають оригінали використаних робіт [9]. Оскільки Plagiarism Advisory Service оперує винятково з електронними англомовними текстами, проблематична її адаптація до аналізу робіт на українській і російській мовах. Крім того, істотним недоліком розглянутої системи є необхідність підключення до Інтернету.

У всіх системах, для яких відомі використовувані в них методи аналізу, ці методи не є прямим відтворенням одного із класичних методів аналізу текстів (Кука, рядків Фібоначчі, Батога - Морріса, Бойєра-Мура, Бойєра-Мура-Хорспула, Бейза-Ейтс і т.інш. [21]), а являють собою спеціально розроблені процедури, що становлять елементи «ноу-хау» розроблювача.

Опис інформаційної технології виявлення текстових запозичень у роботах студентів технічних ВНЗ. Однією з головних задач вищої школи є підвищення ефективності

навчального процесу за рахунок скорочення часу на аналіз і оцінку студентських робіт, зменшення числа помилок при ухваленні рішення й зниження основного ризику - невиявлення запозичень. Викладачеві в процесі аналізу й оцінювання студентських робіт необхідно приймати відповідальні рішення, спираючись на свій досвід і інтуїцію. Для цього йому необхідно володіти інформацією, що надходить із різних структурних одиниць навчального закладу: ректорату. факультетів, кафедр і т.д., а також володіти інформацією про поточний стан БД робіт студентів і ін.

Задача виявлення плагіату в освітньому процесі зводиться до наступних етапів: класифікація видів плагіату, діагностики причин плагіату, створення технологій виявлення плагіату, обґрунтування критеріїв плагіату, організації контролю й відповідальності за плагіат і заходів по його запобіганню.

Технологія аналізу студентських робіт з використанням повинна дозволяти використання документів, які подано у форматах TXT, DOC, DOCX і RTF, та передбачати реалізацію таких етапів:

1. Завантаження користувачем студентської роботи для аналізу
2. Настроювання додаткових умов пошуку, таких як метод аналізу й релевантних умов.
3. Автоматична перевірка документа на наявність плагіату.
4. Формування звіту за результатами пошуку плагіату в аналізованій роботі. У випадку, якщо обсяг запозичень знайдений у документі перевищує граничне значення.
5. Формування системою рішення про наявність в аналізованому тексті плагіату із вказівкою джерел запозичень.
6. Проведення, якщо буде потреба, аналізу при змінених параметрах.
7. Роздрукування звіту за результатами аналізу.

Авторами разом зі співробітниками кафедри «Програмного забезпечення комп'ютерних систем» Національного аерокосмічного університету ім. М.Є.Жуковського «ХАІ» створено ПЗ «Plagiarism» з метою забезпечення можливості акумуляції й накопичування досвід викладачів і нормоконтролерів, а також підтримки прийняття рішень при на аналізі й оцінюванні студентських робіт.

Основне призначення ПЗ «Plagiarism» полягає у автоматизації процесу аналізу в студентських роботах на предмет виявлення запозичень у тексті, що дозволить підвищити якість перевірки робіт за рахунок спрощення діяльності експертів при аналізі результатів виконання студентами індивідуальних завдань, так само прискорити даний процес і знизити при цьому ризик невиявлення запозичень. Це особливо актуально в умовах зростаючих вимог до якості підготовки фахівця, необхідністю частої зміни вимог до студентських робіт і підвищення якості навчального процесу. Крім того, впровадження системи стане «стимулом» для студентів до самостійного написання текстів, а не створенню їх, наприклад, шляхом компіляції зі знайдених в Інтернеті різних документів, що стосуються заданої тематики.

На основі виділення мінімальних необхідних функцій системи і їх складових, а так само аналізу необхідної функціональності до системи, була розроблена схема ПЗ «Plagiarism», яка наведена на малюнку 1.

Розроблений прототип ПЗ «Plagiarism» призначений для:

- використання викладачем в учбово-методичному відділі університету (УВУ);
- використання викладачами й нормоконтролерами для одержання інформації про індивідуальність виконання студентом виконаної роботи; завідувачами кафедр й деканами факультетів;
- виконання аналітичних розрахунків за результатами аналізу роботи студента;
- формування звіту про ступінь самостійності виконання студентом певного виду роботи, що забезпечує своєчасне підведення підсумків контрольно-оцінної діяльності викладача.

Структура ПЗ «Plagiarism» включає наступні блоки: інтеграції; зберігання документів; адміністрування БД; пошуку; аналізу документів; формування підсумкових результатів

аналізу; надання сервісних можливостей; настроювання інтерфейсу; формування звітів.

Режими роботи ПЗ «Plagiarism» відповідає номенклатурі користувачів і варіантам застосування.

У варіанті використання на кафедрі користувачем ПС «Plagiarism» є викладач, що здійснює перевірку навчальних робіт у рамках навчального навантаження, а так само нормоконтролер випускаючої кафедри, що перевіряє кваліфікаційні роботи.

На рівні факультету користувачем ПЗ «Plagiarism» є нормоконтролер в обов'язки якого входить контроль якості кваліфікаційних робіт, і виявлення запозичень у них з робіт студентів різних спеціальностей факультету.

На рівні інституту користувачем ПЗ «Plagiarism» є нормоконтролер ВНЗ, який покликаний забезпечувати контроль якості кваліфікаційних робіт на основі порівняння робіт студентів різних кафедр і факультетів.

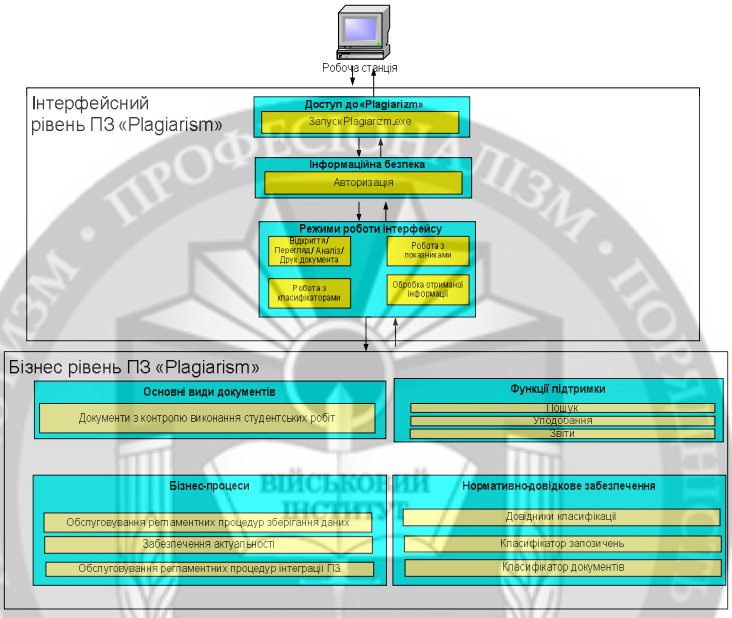

Рис 1. Функціональна схема ПЗ «Plagiarism»

Кожний користувач ПЗ «Plagiarism» може мати одну із двох функціональних ролей, які визначають набір доступних можливостей і операцій над файлами.

До складу ПЗ «Plagiarism» входять наступні підсистеми (рис. 2): підсистема керування базою даних; підсистема настроювання процесу аналізу; модель предметної галузі; підсистема зв'язку із зовнішніми додатками; підсистема підтримки прийняття рішень; підсистема аналізу студентської роботи; підсистема видачі рекомендацій; графічна підсистема.

Підсистема керування базою даних призначена для витягу з бази певних видів робіт, а також для керування ресурсами банку даних або окремих баз даних. База даних тестових завдань складається з наборів різних робіт, об'єднаних по наступним ознакам: група в якій учився студент, рік навчання, спеціальність.

Підсистема настроювання процесу аналізу призначена для того, щоб користувач мав можливість указувати ті складові аналізу, які для нього важливі при аналізі студентських робіт, і які графічні результати він хотів би одержати в результаті аналізу.

Модель предметної галузі являє собою об'єктну модель предметної галузі, що є ядром системи.

Підсистема зв'язку із зовнішніми додатками призначена для забезпечення роботи із зовнішніми об'єктами, такими, наприклад, як Microsoft Word 2007.

Підсистема підтримки прийняття рішень призначена для оперативного аналізу й обробки даних студентської роботи залежно від обраних настроювань користувача.

Підсистема аналізу студентської роботи необхідна для проведення аналізу текстової інформації на предмет наявності в ній запозичень на основі методів нечіткої логіки, а також попереднього аналізу результатів, отриманих підсистемою. Підсистема видачі рекомендацій і висновків є системою підтримки прийняття рішень, що дозволяє формувати набір рекомендацій, щодо поліпшення процесу навчання й надавати його експертові.

Графічна підсистема спеціально орієнтована на потреби користувача ПЗ «Plagiarism» і складається з достатнього числа візуальних компонентів, призначених для зручного подання даних.

Ядром розробленої інформаційної технології є інформаційна технологія виявлення текстових запозичень у роботах студентів технічних ВНЗ.

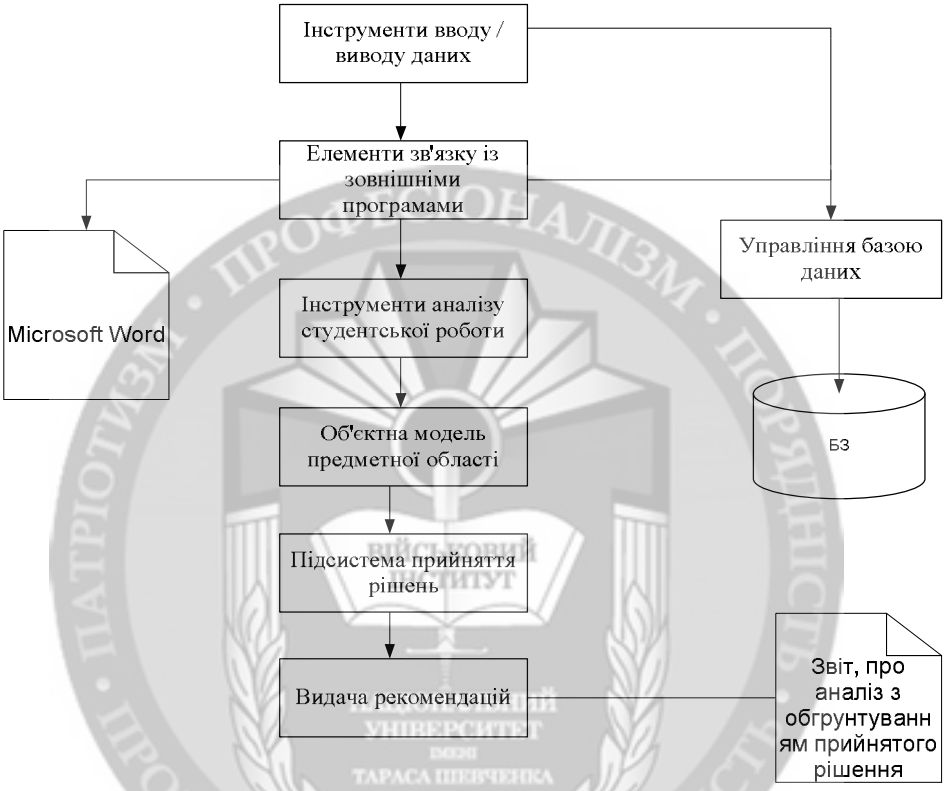

Рис. 2. Схема взаємодії підсистем ПЗ «Plagiarism»

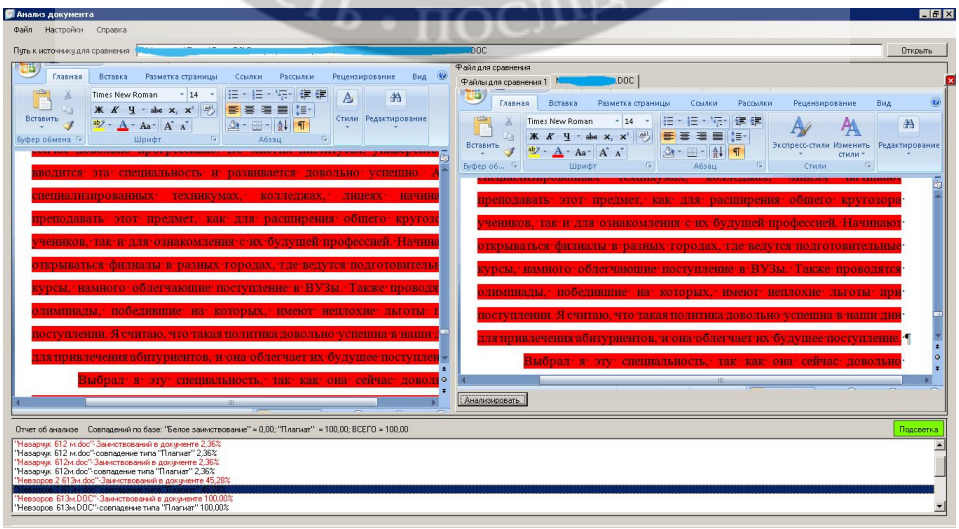

Рис. 3. Вид головного вікна з підсвічуванням типу «плагіат» ПЗ «Plagiarism»

Розробка ПЗ «Plagiarism» здійснювалася шляхом реалізації синтезованих алгоритмів засобів об'єктно-орієнтованого програмування з використанням спеціально розроблених моделей і методів підтримки прийняття рішень в умовах невизначеності.

Авторами розроблений програмний засіб «Plagiarism» (рис. 3), в основу функціонування якого покладена сукупність як традиційних, так і оригінальних методів аналізу текстів із застосуванням технології нечіткого пошуку, що забезпечують ефективність процесу виявлення запозичень при аналізі гетерогенних текстів, характерних для робіт студентів технічного ВНЗ [22]. Ця особливість визначає головну перевагу програмного засобу «Plagiarism» перед іншими спеціалізованими засобами текстологічного аналізу студентських робіт. Крім того, ПЗ «Plagiarism» має розширену функціональність.

Оцінювання якості рішень, які формовано ПЗ «Plagiarism». Задача виявлення по наявності плагіату передбачає експериментальне дослідження й оцінку адекватності моделі виявлення плагіату, а так самого рішення про нього залежно від структурних частин у які він знайдений.

Нехай загальна задача прийняття рішень S визначаються формально [23] четвіркою:

$$S = (\Theta, U, L, P). \quad (1)$$

де Θ – множина можливих значень невідомої ознаки ОПР, тобто множина можливих ситуацій по результатам спостережень $X = \{X_1, X_2, \ldots, X_n\}$; $U = \{u\}$ – множина можливих рішень; $L : \Theta \times U \rightarrow R$ – обмежена дійсна функція втрат $L(\Theta, U)$; P – деяка статистична закономірність на Θ, тобто замкнуте у відповідній топології непусте сімейство кінцево-аддитивних імовірних мір на 2^{Θ}, або заданий клас функціональних закономірностей.

Потрібно вибрати таке $u \in U$, котре мінімізувало б втрати при невідомому $\theta \in \Theta$. Помітимо, що в нашім випадку є можливість набувати знання за допомогою експертів, навчатися на сценарних прикладах і формувати функціональні закономірності на Θ.

Це дозволяє задавати схему прийняття рішень (СПР) Z у вигляді впорядкованої трійки:

$$Z = (\Theta, U, L) \in \mathbf{Z}. \quad (2)$$

із класу Z припустимих СПР і реалізувати алгоритмічний пошук системи P функціональних закономірностей у формі для висновку шуканих рішень.

Отже, задача (2) уточнимо так: заданий СПР $Z = (\Theta, U, L) \in \mathbf{Z}$ і клас закономірностей $P(\Theta)$, що характеризують $\{\theta_n\} \in \Theta$. Необхідно так вибрати послідовність рішень $\{u_n\} \in U$, щоб середні втрати (або ризик R(S)) були мінімальними при заданому класі закономірностей $P(\Theta)$.

Отже, у цілому адекватність і ефективність запропонованого методу оцінки факту запозичення в студентській роботі будемо обчислювати мінімаксним критерієм ризику R. Найкращим рішенням будемо вважати рішення u^{**}, для якого виконується умова

$$\sup_{\theta} R(\theta, u^{**}) = \inf_{u} \sup_{\theta} R(\theta, u). \quad (3)$$

На практиці ризик R оцінюється величиною помилки, що робить знайдене вирішальне правило на контрольних ситуаціях $\{\theta_k, k \in K\} \subseteq \Theta$. Величину помилки охарактеризуємо відношенням числа не правильно прийнятих правилом рішень до загальної кількості пропонованих. Наведена вище методика оцінювання використовувалася в машинному експерименті.

Експериментальне оцінювання адекватності й ефективності спеціально розробленого методу оцінки факту запозичення в студентській роботі проведено на основі отриманих результатів вирішення тестових і реальних задач прийняття рішень із використанням ПЗ «Plagiarism» наочно зображено на рис. 4.

Рішення 2400 контрольних значень дало наступний результат: кількість правильних рішень – 2272; кількість неправильних рішень – 128 разів. Таким чином, ризик прийняття неправильних рішень за допомогою ПЗ «Plagiarism» становить 5,3%. Рішення цих же

контрольних завдань без використання ПЗ «Plagiarism» за допомогою іншого програмного інструментарію, а зокрема WCopyfind дало наступні результати: кількість правильних рішень – 2224; кількість неправильний рішень – 176. Таким чином, у цьому випадку ризик становить близько 7,5%. За допомогою ж Антиплагіат аналіз студентських робіт дав наступні результати: кількість правильних рішень - 2160; кількість неправильних рішень - 240. Таким чином, у цьому випадку ризик становить 10%.

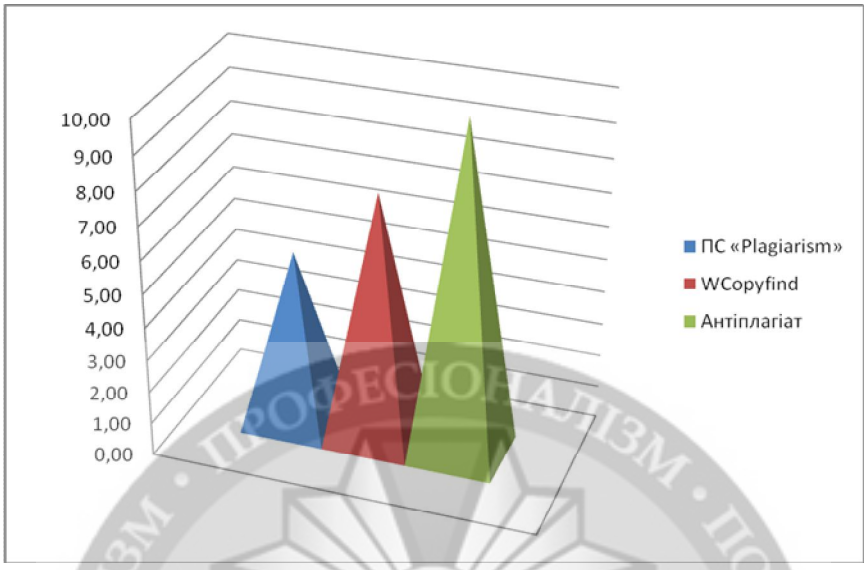

Рис. 4. Оцінка якості прийнятих рішень щодо наявності плагіату в роботах студентів

Отже, використання ПЗ «Plagiarism» призначеного для виявлення текстових запозичень (плагіату) у роботах студентів надає змогу знизити ризик прийняття невірних рішень на 12% за рахунок використання при прийнятті знаниєоріентованих рішень на основі нових моделей і методів текстового аналізу, реалізовані в даному дослідженні.

Висновки. Розроблена інформаційна технологія виявлення текстових запозичень у роботах студентів технічних ВНЗ, втілена у вигляді ПЗ «Plagiarism», що існує у формі дослідницького прототипу, успішно випробувана при рішенні тестових і реальних задач аналізу й оцінювання студентських робіт на предмет виявлення запозичень у тексті.

Результати досліджень показали, що створена інформаційна технологія виявлення текстових запозичень у роботах студентів технічних ВНЗ, і розроблена на її основі ПЗ «Plagiarism» мають більш високу ефективність у порівнянні з аналогічними по призначенню системами Антиплагіат і WCopyfind. Так, застосування ПЗ «Plagiarism» дає можливість формувати рішення, імовірність помилки в яких знаходиться на рівні 16%. Цей показник менше, ніж в аналогічних системах.

ЛІТЕРАТУРА:

1. Груздо, І.В. Анализ программного обеспечения для обнаружения плагиата в научных работах [Текст] / І.В. Груздо // Перша Науково-Технічна конференція Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління2010., - г. Київ, 13-14 грудня 2010г. – С 264-264.

2. Груздо И.В. Проблемы анализа естественно-языковых текстов для обнаружения плагиата в учебных работах / И.В. Груздо // Радіоелектронні і комп'ютерні системи, 2011, №1 (49), С. 130- 138.

3. TurnItIn. Plagiarism prevention service TurnItIn. [Электронный ресурс]. – Режим доступа к документу: http://www.turnitin.com. - Заголовок с экрана.

4. Technology & Teaching. Turnitin.com, a Pedagogic Placebo for Plagiarism [Электронный ресурс]. Emailed 6/5/01; Archived on the Web 6/13/01 – Режим доступа к документу: http://www.bedfordstmartins.com/technotes/techtiparchive/ttip060501.htm. - Заголовок с экрана.

5. Wendy Sutherland-Smith, Rodney Carr. Turnitin.com: Teachers' Perspectives of Anti-Plagiarism Software in Raising Issuesof Educational. Journal of Univ ersity Teaching and Learning Practice - December 2004 - P. 95-101.

6. Детектор плагиата – поиск плагиата и нарушение авторских прав в Интернете. Купить программу. [Электронный ресурс]. – Режим доступа к документу: http://www.detector-plagiata.ru/buy.html. Заголовок с экрана.

7. PlagiatInform Проблемы плагиата в вузах. - [Электронный ресурс]. – Режим доступа к документу: http://www.searchinform.com/search-site/ru/main/full-text-search-case-studies-plagiatinform.html. Заголовок с экрана.

8. Интернет-сервис AntiPlagiat.ru. [Электронный ресурс]. – Режим доступа к документу: http://www.antiplagiat.ru/

9. Википедия. Антиплагиат. [Электронный ресурс]. – Режим доступа к документу: http://ru.wikipedia.org/wiki/Антиплагиат. Заголовок с экрана.

10. Advego Plagiatus. [Электронный ресурс]. – Режим доступа к документу: http://advego.ru/blog/read/plagiatus_news/. Заголовок с экрана.

11. Проверка контента на плагиат. [Электронный ресурс]. – Режим доступа к документу: – Сервисы, программы, алгоритмы http://blog.negotiant.org/proverka-kontenta-na-plagiat-servisy-programmy-algoritmy/. Заголовок с экрана.

12. ЛингвоАнализатор. [Электронный ресурс]. – Режим доступа к документу: http://www.rusf.ru/books/analysis/. Заголовок с экрана.

13. Хмелёв Д.В. О Лингвоанализаторе 3-эпсилон. [Электронный ресурс]./ Электронный журнал "Самиздат" 01/2009. – Режим доступа к документу: http://samlib.ru/h/hmelew_d_w/descrwin.shtml Заголовок с экрана.

14. CaNexus. Plagiarism detection system EVE2. [Электронный ресурс]. – Режим доступа к документу: http://www.canexus.com. Заголовок с экрана.

15. Renoir Gaither. Plagiarism Detection Services, Renoir Gaither, Shapiro Undergraduate Library, University of Michigan (11/04) [Электронный ресурс] Режим доступа к документу: http://www.lib.umich.edu/acadintegrity/instructors/violations/detection.htm

16. Colin J. Neill, Ganesh Shanmuganthan, A Web-Enabled Plagiarism Detection /Tool, IEEE IT Pro, Volume 6 Issue 5, September 2004, - P 19-23.

17. Plagiarism-Finder. [Электронный ресурс]. – Режим доступа к документу: http://www.m4-software.com/en-index.htm. - Заголовок с экрана.

18. CFL Software Limited. Collusion and plagiarism detection program. [Электронный ресурс]. – Режим доступа к документу: http://www.copycatchgold.com. Заголовок с экрана.

19. Turnitin. The global leader in addressing plagiarism and delivering rich feedback. [Электронный ресурс]. – Режим доступа к документу: www.turinitin.com. Заголовок с экрана.

20. Williams J.B. Plagiarism: deterrence, detection and prevention. . [Электронный ресурс]. – Режим доступа к документу: http://www.economicsnetwork.ac.uk/handbook/printable/plagiarism.pdf. Заголовок из файла.

21. WordCHECK keyword software. [Электронный ресурс]. – Режим доступа к документу:http://www.wordchecksystems.com. Заголовок с экрана.

22. Груздо, И.В. Информационная технология выявления заимствований в работах студентов технических вузов на основе нечеткого поиска [Текст] / И.В. Груздо, И.В. Шостак // Международная научно-техническая конференция "Информационные системы и технологии" посвященную 75-летию В.В. Свиридова ИСТ-2012, г. Харьков, 2012. – С. 31.

23. Конорев Б.М., Харченко В.С., Чертков Г.Н. Инструментальная система для поддержки экспертизы и независимой верификации критического ПО: принципы построения и применения //Информационные технологии и безопасность. – Киев: НАНУ, Институт проблем регистрации информации, 2003. – №4. – С. 85 – 91.

Рецензент: д.т.н., проф. Божко В.П.

д.т.н., проф. **Шостак І.В., Груздо І.В.**

КОМП'ЮТЕРИЗАЦІЯ ПРОЦЕСУ ВИЯВЛЕННЯ ПЛАГІАТУ У СТУДЕНТСЬКИХ РОБОТАХ

В статье изложена информационная технология анализа работ студентов технических вузов на предмет выявления в них текстовых заимствований. Осуществлен обзор и критический анализ существующих программно-инструментальных средств анализа естественно-языковых текстов (на русском и украинском языках) в аспекте их пригодности к выявлению плагиата в них. Предложен подход к оценке качества решений, которые сформированы программным средством в случае выявления плагиата. Обоснованно комплексный критерий качества студенческой работы в аспекте наличия в ней текстовых заимствований.

Ключевые слова: плагиат, студенческая работа, программное средство, информационная технология, качество решений.

D.Sc., prof. **Shostak I., Груздо І.**

THE COMPUTERIZATION PROCESS OF IDENTIFYING PLAGIARISMIN THE WORKS OF STUDENTS

In the article the analysis of information technology work students of technology to detect them text loans. Made the survey and critical analysis of existing software tools analysis of natural language texts (in Russian and Ukrainian languages) in terms of their suitability to detect plagiarism in student papers. An approach to evaluating the quality of decisions that formed the software in case of plagiarism. Grounded complex criterion of quality student work in terms of presence in her text loans.

Keywords: plagiarism, the student's work, software, information technology, quality solutions.