УДК [347.78:004.738.5]:378

Т. Е. ЛІХАЧОВА,
І. О. ПОБІЖЕНКО,
В. В. ПОБІЖЕНКО

ІНФОРМАЦІЙНІ СИСТЕМИ ПОШУКУ ПЛАГІАТУ В СТУДЕНТСЬКИХ НАУКОВИХ ПРАЦЯХ В УМОВАХ ІНТЕРНЕТ-ПРОСТОРУ: ТЕХНОЛОГІЇ, МЕТОДИ, МОДЕЛІ

Розглядаються проблеми плагіату у вищій школі в умовах Інтернет-мережевого простору. Характеризуються методи та технології, що використовуються для запобігання плагіату в наукових працях студентів. Пропонується узагальнена модель автоматизованої системи перевірки текстів на плагіат і її пошукові можливості.

Ключові слова: *наукові праці, вищі навчальні заклади, студенти, технології, алгоритми, методи, автоматизована система пошуку плагіату (АСПП), плагіат.*

Рассматривается проблема плагиата в высшей школе в условиях Интернет-сетевого пространства. Характеризуются методы и технологии, которые используются для предотвращения плагиата в научных работах студентов. Предлагается общая модель автоматизированной системи проверки текстов на плагиат и ее поисковые возможности.

Ключевые слова: *научные труды, высшие учебные заведения, студенты, технологии, алгоритмы, методы, автоматизированная система поиска плагиата (АСПП), плагиат.*

The plagiarism problem at the higher school in the conditions of Internet and network space is considered. Methods and technologies which are used for plagiarism prevention in scientific works of students are characterized. The general model is offered to the automated system of verification of texts on plagiarism and its search opportunities.

Key words: *scientific works, higher educational institutions, students, technologies, algorithms, methods, automated system of search of plagiarism (ASSP), plagiarism.*

Важливої актуальності в змінах соціальних систем останньої чверті ХХ — початку ХХІ ст. набуває перехід до нової системи суспільних відносин, основаної на цінності кожної людини. Суспільство висуває нові вимоги до особистості, детермінуючи той набір життєвих стратегій та якостей, що дозволить їй максимально реалізувати свій потенціал.

Формування й розвиток творчого потенціалу та підвищення активності студентів пов'язані з функціонуванням соціальних інститутів, особливе місце серед яких посідає освіта. Від неї залежать можливості здобуття знань, саморозвитку, самодіяльності, актуалізації інтелектуального потенціалу майбутнього фахівця.

Основним завданням подальшого розвитку вищої освіти є не стільки надання студентам максимуму наукової інформації, скільки

формування здібностей до творчого мислення. Нова концепція вищої освіти покликана формувати в них уміння самостійно, цілеспрямовано й відповідально навчатися. Створення у вищому навчальному закладі необхідних умов для розвитку в студентів творчих здібностей, виховання особистості, здатної до саморозвитку як під час навчання у вищій школі, так і в подальшій професійній діяльності, є пріоритетним напрямом розвитку освіти.

Значний потенціал у вирішенні означених завдань належить науково-дослідницькій роботі, що можна розглядати як один із засобів розвитку творчого потенціалу особистості студента в сучасних умовах, що в цілому сприятме якісному розвиткові суспільства.

Водночас сучасне інформаційно-комунікаційне середовище містить багато спокус для студентів, серед яких позичити (скопіювати) чужу наукову працю без посилань. Для більшості студентів написання самостійних робіт зводиться до суто технічного пошуку інформації: копіювання чи сканування матеріалу з підручника, посібника, статті (зокрема і з Інтернет) із подальшим його роздрукуванням [5]. При цьому авторів не турбує себе навіть редагування тексту написанням речень, які встановлюють зв'язок між окремими частинами запозиченого тексту, а іноді навіть не читають власного «твору». Жодного аналітичного дослідження з творчим осмисленням не буде в таких роботах. Традиційне навчальне шахрайство — списування в сусіда або з підручника під партою — відійшло в минуле. Нині комп'ютерні технології розширили межі плагіату: потрібну роботу (контрольну, реферат, курсову чи навіть дипломну) можна знайти в Інтернеті. Хто не бажає шукати сам — звертається за допомогою: за певну плату вам зроблять що завгодно, навіть дисертацію, були б гроші. Замовнику залишається лише ознайомитися перед захистом роботи з «власним» твором, аби не потрапити в незручне становище.

Існують різні методи пошуку плагіату в наукових документах, які основуються на загальних методах пошуку інформації. Перший метод пошуку рядка розробили Майкл Рабин та Річард Карп у 1987 р. Цей метод шукає шаблон, який у тексті використовує хешування.

Най відомішим методом обробки дублікатів у веб-пошуці, ретельно викладеним Андрієм Бродером зі співавторами в 1997 р. [1], є метод «шинглів». Щоб підвищити ймовірність того, щоб у результаті невеликої зміни тексту контрольна сума не змінилася, можна спробувати вибрати з тексту декілька підстрок. Шингл (від англійського shingle — черепичка) — це і є підстрока тексту, за якою відбувається обчислення контрольної суми. Оскільки кількість шинглів приблизно дорівнює довжині документа в словах, тобто є достатньо великою, автори запропонували два методи симплювання для отримання підмножин.

Дослідження компаній HP та Microsoft у 2003 р. розвинули дослідження, які викладені Андрієм Бродером [1].

Інший сигнатурний підхід, який оснований не на синтаксичних, а на лексичних принципах, запропонували співробітники Іллінойського інституту у 2002-му та модифікували у 2004 р. [2; 3].

Плагіат має широкий спектр характеристик, тому пошукові методи потребують упровадження різних алгоритмів ідентифікації з метою забезпечення надійності роботи системи. Принципи плагіату зумовлюють необхідність розробки лінгвістичних алгоритмів ідентифікації текстових даних, орієнтованих, відповідно, на пошук еталонних семантичних масивів заданої довжини.

Також необхідні використання складніших алгоритмів пошуку даних за набором ключових слів та подальший інтелектуальний аналіз змісту ідентифікованих ідей. Плагіат полягає в перефразуванні результатів чужої роботи, тому пошукові алгоритми повинні проводити циклічну перевірку лінійних семантичних масивів на предмет виявлення фактів перестановки слів у реченні чи заміни деяких додаткових засобів ідентифікації модифікованих слів, зокрема зміни структури слова й заміни обраних слів їх синонімами. Тому цей метод пошуку плагіату за принципом ідентифікації забезпечує універсальні характеристики пошукових процесів, проте, водночас потребуючи значних часових затрат на виконання ресурсоємних операцій.

Пошук плагіату, за визначенням, ускладнює умову пошукової задачі, тому передбачає використання експертних систем у процесі аналізу текстових масивів даних, а також зумовлює додаткове долучення до оперативної ідентифікаційної бази даних текстів з авторством особи, праці якої перевіряють.

З метою забезпечення швидкодійних режимів роботи пошукової системи розглядаються можливість вибору обмежень ідентифікаційних вимог та реалізація буферного принципу збереження оперативних інформаційних ресурсів, що сприяє оптимізації часових затрат на здійснення пошукових операцій.

Узагальнена модель автоматизованої системи перевірки текстів на плагіат [4] (рис. 1) ілюструє структурну взаємодію базових блоків, що формують програмне середовище АСПП (автоматизованих систем пошуку плагіату).

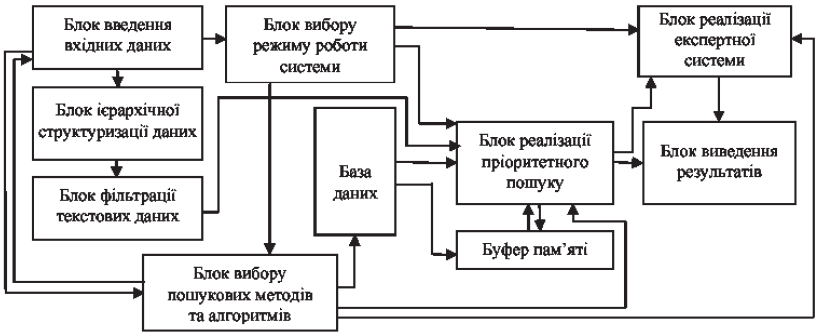

Рис. 1. Узагальнена модель структури АСПП

Блок уведення вхідних даних (рис. 2) реалізує інтерфейсну людино-машинну взаємодію в середовищі АСПП, забезпечуючи початковий та діалоговий режими введення інформації.

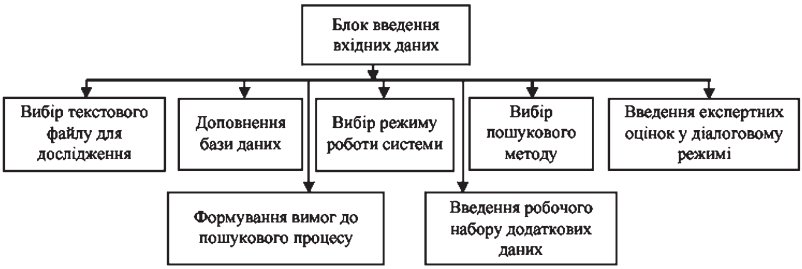

Рис. 2. Структура блоку введення вхідних даних

Блок ієрархічної структуризації даних (рис. 1) формує набір елементарних об'єктів дослідження (рис. 3). Такий підхід дозволяє підвищити швидкодію пошукових процесів та забезпечити функціональну незалежність алгоритмів роботи від вхідних даних.

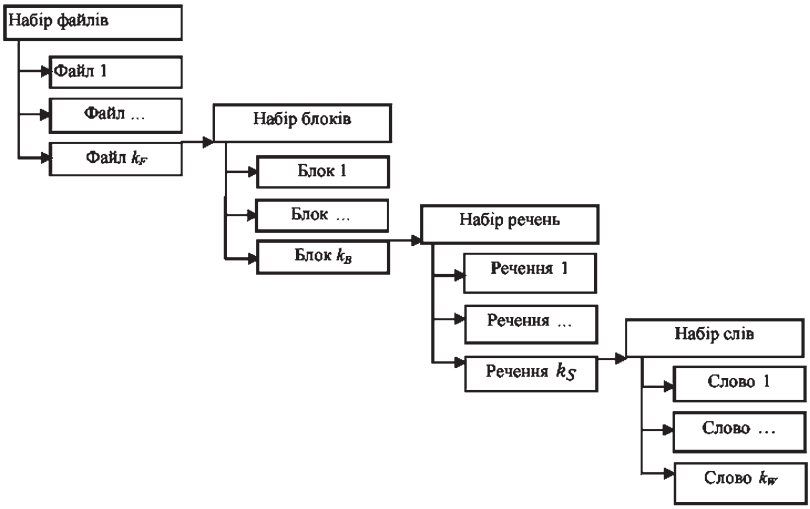

Рис. 3. Структура блоку введення вхідних даних

Ієрархічна структуризація вхідної інформації зумовлює підтримку єдиного підходу до формування інформаційного забезпечення досліджуваного ресурсу, що передбачає можливість оперування даними нетекстового формату без суттєвих змін базового алгоритму.

Блок фільтрації (рис. 1) дозволяє оптимізувати набір вхідних даних текстового файлу, наданого для дослідження, через мінімізацію інформативних ресурсів засобами ієрархічної фільтрації з метою збільшення швидкодії АСПП. Процес фільтрації здійснюється у два етапи. На першому етапі проводиться аналізується вхідний файл на рівні речень. Відфільтровуються речення, які містять менше 30 % символів робочого алфавіту, та речення, довжина яких не перевищує 30 символів. На другому етапі виконується процес зменшення семантичного набору досліджуваного ресурсу через вилучення неінформативних даних, які не впливають на достовірність кінцевого результату пошукового процесу.

Блок вибору режиму роботи АСПП (рис. 4) призначений для встановлення пріоритетності пошукових умов, які впливають на кінцевий вибір методів і алгоритмів перевірки.

Рис. 4. Вибір режиму роботи АСПП

Режим підвищеної швидкодії пошукового процесу передбачає мінімізацію набору вхідних вимог, що дозволяє обмежити кількість обраних алгоритмів текстової ідентифікації в указаному пошуковому методі, та використання буфера пам'яті, куди записуються вибрані з бази даних (БД) робочі інформаційні ресурси з метою їх подальшого дослідження в оперативному режимі.

Блок вибору пошукових методів та алгоритмів (рис. 5) забезпечує вибір набору робочих алгоритмів, зважаючи на вимоги користувача до умов проведення пошукових процесів. У середовищі АСПП реалізовані базові методи перевірки текстів на плагіат, основою яких є принципи визначення сутності плагіату.

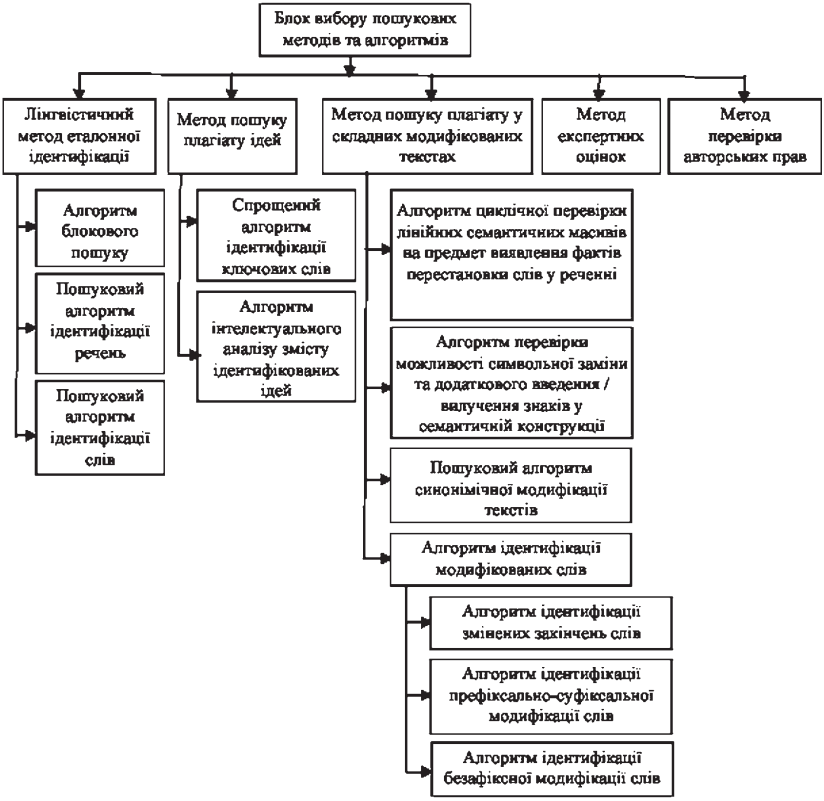

Рис. 5. Структура формування блоку вибору пошукових методів та алгоритмів

Лінгвістичний метод еталонної ідентифікації базується на першому і другому принципах плагіату. Набір алгоритмів методу орієнтований на опрацювання обраних конструкцій структурованих вхідних даних.

Метод пошуку плагіату ідей використовує третій принцип визначення сутності плагіату і реалізує спрощений алгоритм ідентифікації ключових слів за умови підтримки режиму підвищеної швидкодії пошукового процесу й алгоритм інтелектуального аналізу змісту ідентифікованих ідей, який дозволяє проведення ґрунтовніших досліджень.

Метод пошуку плагіату в складних перефразованих текстах розкриває четвертий принцип плагіату. Набір робочих алгоритмів методу дозволяє забезпечити універсальний підхід до перевірки текстових документів за умови їх модифікації.

Метод експертних оцінок орієнтований на використання п'ятого принципу плагіату та передбачає залучення професійного експерта для встановлення ступеня фальсифікації результатів роботи. Основою методу є математичний апарат нечіткої логіки для автоматизованого аналізу результатів експертних оцінок.

Метод перевірки авторських прав базується на шостому принципі плагіату і передбачає дослідження робіт автора, праці якого перевіряються, використовуючи описані вище алгоритми, вибір яких визначається умовами ведення пошукових процесів.

Блок реалізації пріоритетного пошуку (рис. 1) забезпечує підтримку обраних режимів та реалізує використання вказаних методів з ідентифікованим набором робочих алгоритмів.

Блок реалізації експертної системи (рис. 1) містить інтелектуальні методи й алгоритми, має зручний інтерфейс, який дозволяє експертам систематизувати та переглядати результати обробки даних за різними принципами, методами на різних ітераціях і з різною деталізацією. Система автоматично здійснює всі розрахунки, але тільки експерти можуть об'єктивно прийняти остаточне рішення щодо наявності плагіату в тих чи інших частинах інформаційних ресурсів чи ідентифікувати випадковий збіг або використання відомих тез. Адже дві статті на одну достатньо вузьку тему можуть мати до 70% однакових ключових слів та понять, але бути викладенням різних ідей і досягнень. Основою реалізації такої експертної системи має бути математичний апарат нечіткої логіки для мінімізації суб'єктивізму користувача в експертній системі.

Блок виведення результатів (рис. 1) презентує результати перевірки в зручному для користувача вигляді. АСПП визначає коефіцієнт плагіату в досліджуваному документі та уможливлює виведення ідентифікованого збігу семантичних конструкцій в ілюстративному режимі. Крім того, програмно забезпечена можливість друкування результатів та їх збереження в окремому файлі.

Коефіцієнт плагіату визначається відношенням (1)

$$P = 100\frac{k_p}{k_s}\ , \qquad (1)$$

де k_S — загальна кількість структурних елементів у відфільтрованих вхідних даних; k_p — кількість структурних елементів, у яких було виявлено плагіат.

Модель і методи роботи автоматизованої системи перевірки текстів на плагіат реалізують основні принципи визначення сутності плагіату. Система комплексно забезпечує універсальні характеристики пошукових процесів та підтримує режим підвищеної швидкодії завдяки вибору обмежень ідентифікаційних вимог, використання засобів ієрархічної фільтрації вхідних даних та реалізації буферного принципу збереження оперативних інформаційних ресурсів, що сприяє оптимізації часових затрат на прове¬дення пошукових операцій. Результатом роботи показаної системи є визначення коефіцієнту наявності плагіату в досліджуваному документі й ілюстративне виведення ідентифікованого збігу семантичних конструкцій.

Таким чином, АСПП дозволятиме контролювати наукові студентські праці та запобігати плагіату під час їх написання.

У контексті реформування національної освіти відповідно до європейських стандартів в університетах суттєво підвищується роль самостійної роботи студентів. Самостійне здобуття знань під керівництвом викладача розглядається нині як головна ланка в усій системі навчання.

Навчально-дослідна робота в університетах має ставати більше творчо-аналітичною, а студентська молодь — набувати навичок науково-навчальних досліджень: уміти самостійно сформулювати проблему, поставити завдання дослідження, провести інформаційний пошук, проаналізувати, творчо осмислити матеріал. Хто хоч раз у житті пережив муки творчості, провів власне дослідження, зробив своє «відкриття», описав і захистив його, той поважатиме й чужу інтелектуальну власність і не дозволить нікому на неї зазіхати.

У подальших дослідженнях намагатимемося детальніше розкрити питання, пов'язані з поширенням плагіату в студентських працях та подоланням цієї тенденції в майбутньому.

Список літератури

1. Broder A. Min-Wise Independent Permutations / A. Broder, M. Charikar, A. M. Frieze, M. Mitzenmacher // Proc. STOC. — 1998.
2. Chowdhury A. Collection statistics for fast duplicate document detection / A. Chowdhury, O. Frieder, D. Grossman, M. McCabe // ACM Transactions on Information Systems (TOIS), April 2002. — Vol. 20, Issue 2. — 2002. : [Электрон. ресурс]. — Режим доступа: http://ir.iit.edu/~dagr/2002collectionstatisticsfor.pdf

3. Kolcz A. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization / A. Kolcz, A. Chowdhury, J. Alspector // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. — Seattle, WA, USA, 2004. — P.605–610
4. Мокін В. Б. Автоматизована система перевірки текстів на плагіат / В. Б. Мокін, В. В. Войтко, С. В. Бевз, О. В. Гавенко, І. А. Білоус // Вісник Вінницького політехнічного інституту. — №5, 2010. — С. 12–17.
5. Побіженко В. В. Плагіат, як чинник зниження якості освіти / В. В. Побіженко, І. О. Побіженко // Системи обробки інформації: збірник наукових праць. — Х. : Харківський університет повітряних сил ім. Івана Кожедуба, 2011. — №8(98). — С. 310-313.

Надійшла до редколегії 09.08.2013 р.