пошуку є семантична відповідність довільно сформульованому опису об'єкта пошуку чи деякому тексту-зразку. Організація засобів семантичного пошуку навчальної текстової інформації розглядається на основі оцінки тематичної близькості двох документів.

Для кількісного визначення ступеня близькості порівнюваних документів застосовується розроблена методика, що дозволяє врахувати наступні варіанти обчислення тематичної близькості документів:

1) обчислення тематичної близькості без урахування контексту;

2) з урахуванням загального контексту для варіанта, коли множина контекстних елементів не враховує приналежність кожного його елемента ключовому слову тексту документа;

3) з урахуванням індивідуального контексту.

Важливим елементом навчального середовища, призначеного для забезпечення більш ефективного формування чітко структурованої системи знань, вмінь та навичок, орієнтованих на практичне застосування у ході професійної діяльності, у відповідній предметній галузі, згідно з креативною концепцією комп'ютеризованого освітнього процесу, є високоефективні засоби пошуку навчально-методичної інформації.

ПРОБЛЕМА ПЛАГІАТУ У ВИЩИХ НАВЧАЛЬНИХ ЗАКЛАДАХ ТА ШЛЯХИ ЇЇ ВИРІШЕННЯ

Матасар Є.І.,
Київський університет імені Бориса Грінченка, м. Київ

Доступність інформації в сучасному світі створює величезну проблему для освіти. Таке твердження, на перший погляд, викликає величезні сумніви, адже доступність інформації (знань) — це одне з важливих завдань освіти. Але доступна інформація є підґрунтям для виникнення плагіату, який стає нормою у більшості сучасних студентів. Протягом довгого періоду у вищих навчальних закладах України приховували проблему плагіату в студентських роботах. Це пов'язано як з небажанням викладачів витрачати багато часу на аналіз кожної роботи, так і з відсутністю програмних засобів і сервісів для автоматизації перевірки поданих студентами текстів. Сьогодні ця проблема стала

ще більш актуальною, а вищі навчальні заклади знаходяться в пошуках її рішень.

Зараз існує низка актуальних програмних засобів, які шляхом автоматичного порівняння тексту із своїми базами даних і базами даних пошукових систем, допомагають виявити плагіат. Серед цих розробок слід відзначити eTXT Антиплагиат, Copyscape, AntiPlagiat.ru, Pasteit, Turnitin. Усі ці системи мають онлайн-сервіси, які працюють безпосередньо у браузері, а сервіс eTXT Антиплагиат також доступний у вигляді додатку для ОС MS Windows версії XP або новішої.

До головних недоліків цих систем можна віднести обмеженість їх функціональних можливостей в безкоштовних версіях. Так, наприклад, сервіс AntiPlagiat.ru надає розвинуті засоби обробки тексту, але в безкоштовному варіанті можлива перевірка тільки невеликих його фрагментів. Сервіс Turnitin взагалі доступний тільки після придбання платної підписки. Також до недоліків згаданих сервісів можна віднести погану оптимізацію для обробки україномовних робіт, що є серйозною проблемою для українських вищих навчальних закладів. Крім того, деякі сервіси, в тому числі eTXT Антиплагиат, призначені для пошуку плагіату серед робіт, доступних у мереж Інтернет, але не дають змоги аналізувати базу робіт окремих навчальних закладів.

Перелічені недоліки визначають необхідність подальшого вдосконалення цих сервісів, а також створення нових безкоштовних альтернатив, в яких будуть вирішені наявні проблеми та реалізовані додаткові функціональні можливості.

Метою роботи є створення нового безкоштовного сервісу з розширеними функціональними можливостями для перевірки студентських робіт на плагіат.

У ході розробки було досліджено актуальні методи пошуку плагіату, серед яких приділено відповідну увагу алгоритмам для аналізу та порівняння електронних текстів. Для реалізації сервісу обрано «Алгоритм шинглів», який був спеціально розроблений Уді Мамбером для пошуку копій і нечітких дублікатів тексту в мережі Інтернет. Цей алгоритм сьогодні є визнаним, як найбільш оптимальний для пошуку нечітких дублікатів, що використовується в багатьох актуальних програмних засобах для боротьби з плагіатом в мережі Інтернет [1].

Пошук нечітких дублікатів дає змогу з'ясувати, чи є два об'єкти частково або повністю однаковими. Об'єктом може бути текстовий

файл або будь-який інший тип даних. У разі обробки тексту реалізація алгоритму визначає наступні етапи:

- канонізація текстів;
- розбиття тексту на «шингли» (послідовності);
- знаходження контрольних сум;
- пошук однакових послідовностей [2].

Розроблена у ході дослідження версія сервісу успішно порівнює два або більше варіантів тексту між собою та вираховує відсоток унікальності текстів. Особливостями програми є:

- можливість завантаження тексту в форматах MS Office Word (*.doc, *.docx), Portable Document Format (*.pdf) та OpenDocument Format (*.odt);
- наявність бази даних для збереження робіт, результатів їх аналізу, а також необхідних для функціонування програми даних;
- реалізація вибору параметрів аналізу робіт: по окремих студентських групах, курсах, підрозділах навчального закладу тощо.

Можливість завантаження матеріалів з електронних документів спростить опрацювання текстів системою, виключить необхідність ручного перенесення інформації. Створені бази даних дають змогу зберігати оброблені тексти робіт, а також результати проведеного аналізу, що значно спрощує подальше їх опрацювання, зменшує потрібний на це час і навантаження на сервер. Наявні засоби вибору додаткових параметрів аналізу дають змогу сформувати необхідний запит та отримати потрібні результати у найкоротший час.

Розроблений сервіс для аналізу студентських робіт на наявність плагіату має значно спростити практичне виконання перевірки, зберегти час та підвищити ефективність роботи викладачів та студентів.

ДЖЕРЕЛА

1. Зеленков Ю.Г. Сравнительный анализ методов определения нечетких дубликатов для Web-документов [Електронний ресурс] / Ю.Г. Зеленков, И.В. Сегалович. — Режим доступу : http://rcdl2007.pereslavl.ru/papers/paper_65_v1.pdf

2. Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse. Syntactic Clustering of the Web [Електронний ресурс] / Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse. — Режим доступу : http://www.std.org/~msm/common/clustering.html#Common%20shingles