УДК 004.9
С.А. Лупенко, докт. техн. наук., проф., Д.І. Бугальський
Тернопільський національний технічний університет імені Івана Пулюя, Україна

МЕТОДИ АВТОМАТИЗОВАНОГО ПОШУКУ ПЛАГІАТУ В ЕЛЕКТРОННИХ ДОКУМЕНТАХ

S.A. Lupenko, Dr., Prof., D.I. Bugalskyy

METHODS OF THE AUTOMATED SEARCH OF PLAGIARISM IN ELECTRONIC DOCUMENTS

Актуальність автоматизації пошуку плагіату в електронних документах обумовлена проблемою порушення авторських прав. На сьогоднішній день нелегальне поширення творів, що є об'єктом авторського права, - явище буденне. Крім того, проблема незаконного запозичення текстових матеріалів зачіпає і систему вищої освіти.

Якщо говорити про методи виявлення плагіату в довільних текстах, то ці методи можна розділити на два великі класи. Алгоритми, які використовують певні знання про усю колекцію документів, що розглядаються, називають глобальними, інші - локальними.

Основна ідея локальних методів зводиться до синтаксичного аналізу документу.

Простим прикладом може служити алгоритм, який обчислює хеш-функцію (MD5, SHA-2, CRC32) від конкатенації двох щонайдовших речень в документі.

Ефективнішим способом знаходження плагіату може стати метод, заснований на понятті TF (term frequency - частота слова). TF - це відношення числа входжень деякого слова до загальної кількості слів документу.

При використанні семантичної мережі завдання визначення плагіату зводиться до порівняння моделей, що відбивають смислове навантаження текстів.

Велику популярність пошуку плагіату в довільних текстах здобув метод шинглів [1]. Метод шинглів заснований на представленні текстів у вигляді множини послідовностей фіксованої довжини, що складаються з сусідніх слів.

Щодо відомих глобальних методів, то подальшим розвитком методу, що використовує міру TF, став алгоритм, що аналізує документи усієї колекції. У ньому використовуються міра TF - IDF. IDF (inverse document frequency - зворотна частота документу) - інверсія частоти, з якою деяке слово трапляється в документах колекції.

Ще один сигнатурний метод запропонував A. Chowdhury [2]. Ключова ідея цього методу ґрунтується на обчисленні дактилограми I-Match для демонстрації змісту документів.

Метод "опорних" слів, описаний в [3], заснований на сигнатурному підході. Цей метод теж полягає у використанні лексичних принципів, тобто на основі словника.

Як перспективний підхід, спрямований на покращення точності визначення факту плагіату, пропонується застосування взаємодоповнюючих методів аналізу, що забезпечить більш глибокий аналіз вхідних документів.

Література

1. Zweig. Syntactic clustering of the Web. Proc. of the 6th International World Wide Web Conference, April 1997.
2. A. Kolcz, A. Chowdhury, J. Alspector. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization. KDD 2004. http://ir.iit.edu/~abdur/publications/470-kolcz.pdf
3. S. Ilyinsky, M. Kuzmin, A. Melkov, I. Segalovich. An efficient method to detect duplicates of Web documents with the use of inverted index. WWW Conference 2002.