Список використаних джерел:

1. G. Paschos, E. Basëtugˇ, I. Land, G. Caire, and M. Debbah, «Wireless caching: Technical misconceptions and business barriers,» arXiv preprint arXiv:1602.00173, 2016.
2. X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, «Cache in the Air: Exploiting content caching and delivery techniques for 5G systems,» IEEE Communications Magazine, vol. 52, no. 2, pp. 131–139, February 2014.
3. M. Tao, E. Chen, H. Zhou, and W. Yu, «Content-centric sparse multicast beamforming for cache-enabled cloud RAN,» [Online] arXiv: 1512.06938, 2015.
4. F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, «Fog computing and its role in the internet of things».

Різник Р.К.

студент,

Харківський національний університет радіоелектроніки

РОЗРОБКА ТА ДОСЛІДЖЕННЯ МЕТОДІВ ВИЯВЛЕННЯ ПЛАГІАТУ В ТЕКСТОВИХ ДОКУМЕНТАХ

Швидкий розвиток інформаційних технологій надає змогу сучасним студентам та викладачам застосовувати метод написання робіт, що отримав назву «copy paste» – копіювання матеріалів з інтернету з їх мінімальним редагуванням. Копіювання наукових робіт без їх опрацювання – найбільш поширена форма наукової несумлінності. Дуже швидко йде руйнування наукової думки в академічній спільноті [1]. За час існування інформаційних технологій було запропоновано безліч методів для виявлення плагіату, включаючи методи засновані на простому порівнянні документів, аналізі термінів що зустрічаються в тексті, аналізі цитат, аналізі мовних стилів та інших. Часто для підвищення ефективності ці методи комбінують.

У даній роботі розглядаються підходи до класифікації показників схожості текстових документів а також розглядаються методи виявлення плагіату в текстових документах та їх комбінація [2].

Зазвичай системи виявлення плагіату базуються на порівнянні двох або більше документів. Для того, щоб порівняти документи і визначити ступінь їх схожості, необхідно кожному документу присвоїти відсоткове представлення унікальності. Існує декілька класифікацій метрик схожості документів. Одні з них базуються на загальній кількості документів, що залучені до аналізу, інші на обчислювальній складності методів визначення схожості. Метрики можна класифікувати за кількістю вимірів, що залежать від кількості документів. Також можна виділити поверхневі і структурні метрики. Поверхневою метрикою вимірюється схожість документів, простим порівняннями. Структурна метрика, навпаки, передбачає аналіз і опрацювання мовних особливостей. Класифікація показників схожості документів базується на семантичних та статистичних методах.

В одному з перших дослідів виявлення нечітких дублікатів для побудови вибірки використовувалися послідовності сусідніх літер що називаються дактилограмами [3]. Дактилограма документу включає всі текстові рядки фіксованої довжини. В якості міри схожості двох документів виступає відношення кількості однакових рядків до розміру документу. Існує схожий синтаксичний метод оцінки подібності між документами, оснований на представленні документа у вигляді множини вже не рядків а всіляких послідовностей довжини k, що складаються з сусідніх слів. Такі послідовності називаються «шинглами». Два документи вважаються схожими, якщо їх множини шинглів в значній мірі пересікаються. Цей метод має значний недолік, так як кількість шинглів досить велика та приблизно дорівнює кількості слів в документі, то для порівняння двох документів необхідно попарно порівняти між собою всі шингли, що досить не ефективно з алгоритмічної точки зору.

Для більш ефективного порівняння документів між собою пропонується використати MinHash – метод визначення схожості множин [4]. Він позбавляє від попарного зрівняння всіх шинглів кожного документа. Ідея цього методу полягає у обчисленні та порівнянні сигнатур документів, що обчислюються лише один раз. Сигнатура обчислюються наступним чином: визначається хеш-функція $h_{min}(S)$, за допомогою якої обчислюється мінімуми всіх хешів-функцій для множини шинглів S. Сигнатура має фіксований розмір, тому для порівняння двох документів необхідно виконати фіксовану кількість операцій.

Для підвищення якості та швидкості при пошуку плагіату доречно буде проводити попередню перевірку тексту на однорідність: у випадку якщо якийсь фрагмент тексту явно виділяється від загального авторського стилю, то досить велика ймовірність того, що цей фрагмент запозичений з деякого іншого джерела. І в такому випадку його необхідно порівнювати з іншими документами. Для проведення такого аналізу можна використовувати методи статистичного аналізу і теорії інформації, методи стиснення інформації, перевірку статистичних гіпотез про рівність середніх на основі критерія Стьюдента, а також методи машинного навчання. Можливе запозичення виявляти шляхом навчання і тестування на фрагментах тексту методом перехресної перевірки. Також можна навчити декілька різних класифікаторів, що будуть виявляти стать автора, його вік, освіту, та стиль конкретного автора.

В даній роботі пропонується комбінована методика виявлення плагіату в текстових документах, що включає наступну послідовність дій, представлену на блок-схемі (див. рис. 1):

– перевірка тексту документа на однорідність;
– визначення «шинглів» для документа;
– обчислення MinHash сигнатури документа на основі його «шинглів»;
– порівняння сигнатури з сигнатурами інших документів та визначення коефіцієнта схожості.

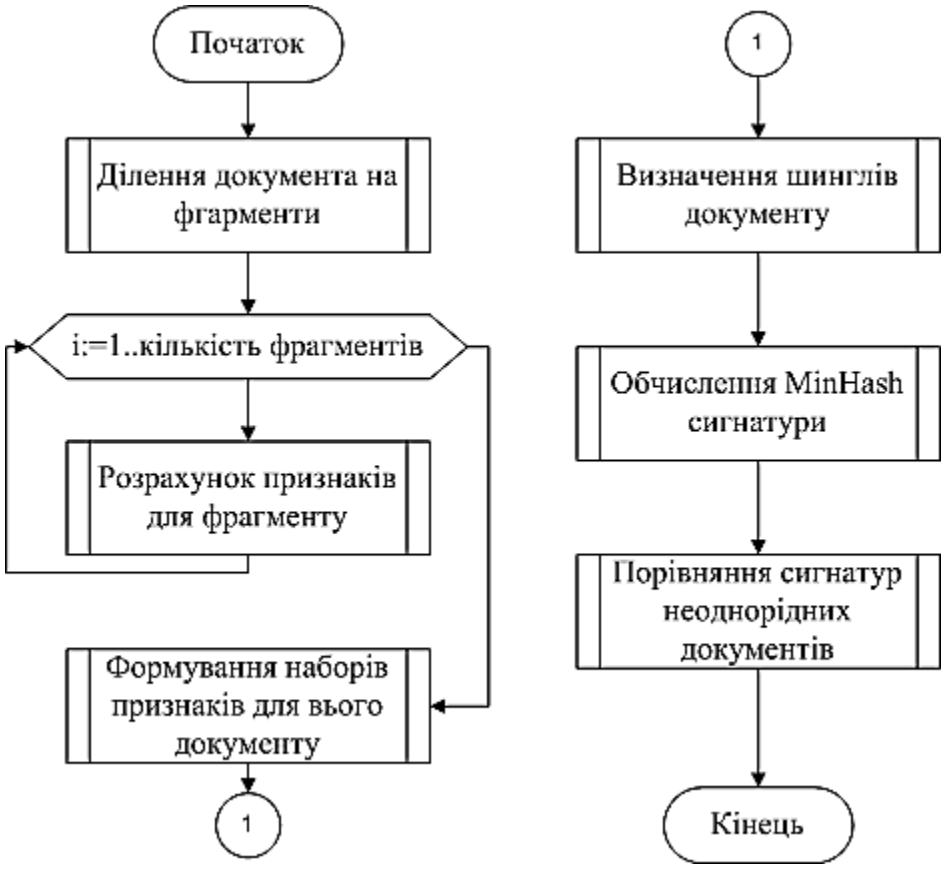

Рис. 1. Блок-схема системи виявлення плагіату

Список використаних джерел

1. Дикань С.А. Плагіат в освіті: походження, причини та шляхи подолання [Текст] / Дикань С. А., Безпека життєдіяльності. 2007. – №5. – Ст. 16-20.

2. Брін С., Девіс Д., Гарсія-Моліна Х. Механізми виявлення копіювання для електронних документів [Текст] / Vine. – 2001.

3. Хейнз Н., Масштабуємі відбитки документі [Текст] / USENIX Семіран по електонній торгівлі, листопад. 1996.

4. Васильвицький С. Робота з Великими Даними (записи з лекцій, університет Колумбії) [Текст] / 2011.

Різник Р.К.

студент,

Харківський національний університет радіоелектроніки

ВИЗНАЧЕННЯ ДОПОМІЖНОГО СЛОВНИКА ДЛЯ АНАЛІЗУ НЕЧІТКИХ ДУБЛІКАТІВ АЛГОРИТМОМ I-MATCH

В останні роки великі динамічні сховища документів стали звичайним явищем, чому сприяє швидкий розвиток інтернету. В силу багатьох факторів, таких як поширення спаму, плагіату та інших постає задача виявлення в