УДК 004.912:82-995(045)

Ланде Д.В., д-р техн. наук
Жигало В.В.

ПІДХІД ДО РІШЕННЯ ПРОБЛЕМИ ПОШУКУ РІЗНОМОВНОГО ПЛАГІАТУ

Інформаційний центр «Електронні вісті»

Описано підхід до пошуку дублікатів документів (або їх частин), наведених різними мовами. В основу підходу покладено використання частотних морфологічних словників, а також словників перекладів. Пошук дублікатів проводиться за допомогою виділени опорних слів, які витягаються за допомогою емпірико-статистичних правил, використання частотних словників та їх перекладів. Даний підхід реалізовано в системі контент-моніторингу InfoStream. Також у результаті виконання процедури пошуку дублікатів було створено двомовний паралельний корпус документів

Вступ

На сьогоднішній день мільйони документів, наведених в *Internet*, багатократно дублюються, це стосується не тільки новин, але й статей, рефератів, і навіть дисертацій та наукових звітів.

Передрук новин не забороняється законами – у цьому випадку ніхто не несе відповідальності за порушення авторських прав. Але це не стосується використання документів інших авторів без належного зазначення відповідного посилання на ту чи іншу використану роботу іншого автора. У відповідності із законодавством України плагіат – це оприлюднення (опублікування), повністю або частково, чужого твору під іменем особи, яка не є автором цього твору [1].

На сьогоднішній день стоїть гостра проблема виявлення випадків плагіату. Так як з появою мережі Internet все більше документів стають доступними, а отже прослідити використання кожного документа не завжди можливо.

Вже є досить багато сервісів які допомагають виявити плагіат, але вони найчастіше обмежені однією мовою, якою був створений документ. Існують лише деякі позитивні приклади рішення цієї проблеми [2], в рамках цієї статті пропонується ще один підхід, який вже успішно втілено при рішенні проблеми пошуку дублікатів у новинних повідомленнях [3].

Автоматизоване виявлення випадків плагіату

Автоматизоване виявлення випадків плагіату здійснюється з використанням спеціальних програмних засобів – текстових аналізаторів. Текстовий аналіз у цих випадках проводиться на основі суб'єктивних або кількісних методів [4]. Суб'єкти вні методи аналізу текстів опираються на використання спеціальних термінів, словосполучень, язикових штампів тощо. Кількісні ж методи будуються на базі лексичних, морфологічних, синтаксичних конструкцій мови та містять у собі певні статистичні характеристики тексту, що підлягає аналізу.

В даній статті описано саме кількісний підхід, за допомогою якого можливо виявити дублікати документів, наведених різними мовами (реалізовано пошук дублікатів статей, наведених українською та російською мовами). Процедура виявлення дублікатів побудована на використанні методів витягу опорних ключових слів на основі емпірико-статистичних властивостей тексту за допомогою частотного морфологічного словника, а також перекладу цих слів на іншу мову.

Зазвичай в основу пошуку плагіату покладено порівняння текстів [5, 6]. Спочатку проводиться порівняння текстів в цілому, а далі відбувається розбивка на абзаци і потім пошук конкретних фрагментів тексту в інших документах.

В інших же випадках використовується пошук за ключовим словами або ж словосполученнями.

Підраховується загальна кількість знайдених слів і словосполучень, а також знайдених фрагментів тексту. Враховуючи кількість малочастотних термінів знайдених при перевірці, як результат отримують підтвердження про те що документ чи фрагмент плагіат, або навпаки.

Чисельні системи, виконують пошук не тільки фрагментів а й навіть проводять достатньо складний аналіз тексту який включає використання такого метода, як метод Флеша – який дозволяє вичислити індекс «легкості» тексту.

Аналізуючи показники індексів абзаців у роботі, що перевіряється, можна ідентифікувати найбільш ймовірніші абзаци плагіату, пошук яких потім проводиться на внутрішній базі даних, або ж в мережі *Internet.*

Вибір моделі представлення тексту

Для виділення ключових слів з тексту в рамках підходу, що розглядається, використовується векторна модель документа, у відповідності до якої, кожному слову документа привласнюється його ваговий коефіцієнт. Чим більше вага слова, тим більше це слово характеризує документ. Авторами було перевірено два підходи обчислення вагових коефіцієнтів слів.

TFIDF (від англ. *TF – erm frequency, IDF –inverse document frequency*) – статистична міра, використовувана для оцінки важливості слів у контексті документа, що входить до складу деякого текстового масиву. Вага окремого слова пропорційний кількості його вживання в документі та зворотно пропорційний частоті вживання в інших документах масиву.

Міра *TF-IDF* часто використовується в завданнях аналізу текстів та інформаційного пошуку, наприклад, як один із критеріїв релевантності документа пошуковому запиту, для розрахунку міри близькості документів при кластеризації.

$$TF = \frac{n_i}{\sum_k n_k}, \qquad (1)$$

де *nk* – число вживань слова, а в знаменнику – загальне число слововживань.

IDF (*inverse document frequency* – зворотна частота документа) – нверсна частота, з якою деяке слово зустрічається в документах масиву. *IDF* зменшує вагу часто вживаних слів:

$$IDF = \frac{|D|}{N_i}, \qquad (2)$$

де $|D|$ – кількість документів у масиві, N_i – кількість документів, у яких зустрічається термін.

Таким чином, міра *TFIDF* є добутком двох співмножників: *TF* і *IDF*.

При тестуванні процедури на базі стандартного підходу *TFIDF* алгоритм поводився не зовсім коректно на великих текстових масивах. Як альтернативу стандартному методу *TFIDF* було взято його модифікацію – *Okapi BM*25:

$$TF \cdot IDF = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 (1 - b + b \cdot \frac{|D|}{avgdl})}, \qquad (3)$$

де *f* (*qi*, *D*) – частота терміна *qi* у документі *D*; $|D|$ – довжина документа *D* (число слів); *avgdl* – середня довжина документа в колекції; *k*1 і *b* – вільні параметри, зазвичай обрані як *k*1=2.0 і *b*=0.75.

IDF (q_i) – *IDF* інверсна частота документа, що обчислюється за формулою:

$$IDF(q_i) = \log \frac{N - n(qi) + 0.5}{n(qi) + 0.5}, \qquad (4)$$

де *N* – загальна кількість документів у масиві, *n*(*qi*) – кількість документів, що містять термін *qi*.

Суть даного підходу полягає у тому, що на відміну від загально прийнятого підходу в *Okapi BM*25 береться до уваги довжина документа.

Алгоритм виявлення дублікатів

Процедуру виявлення дублікатів можна розкласти на декілька етапів:

- створення морфологічних словників;

- створення частотних словників – навчання системи;
- створення словників перекладів;
- побудова програмами пошуку ключових слів;
- створення процедури пошуку дублікатів на різних мовах.

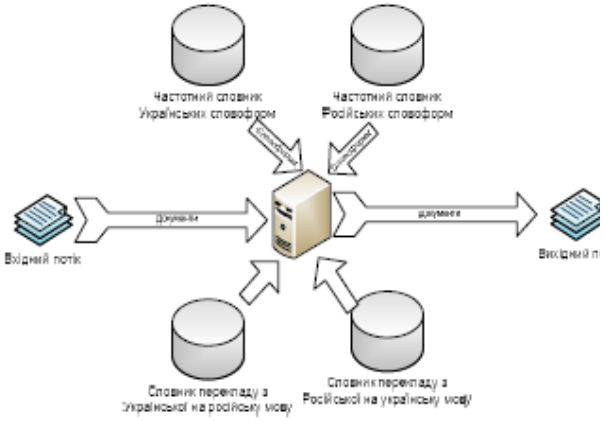

Рис. 1. Функціональна схема пошуку та перекладу ключових слів

Спочатку створюються морфологічні словники які для кожної словоформи містять її нормальну форму. Це потрібно для того, щоб в подальшому можна було привести всі знайдені словоформи до нормальної форми.

Далі створюється частотний словник на базі морфологічного словника, в якому записується частота кожної словоформи, знайденої в процесі «навчання» частотного словника на тестовому масиві документів.

Для побудови електронних морфологічних словників було взято наявну електронну версію словника Залізняка, який налічує близько 93 тис. слів у нормальній формі, для російської мови та безкоштовний словник *ispell*, який налічує близько 1 мільйона українських словоформ, відповідно, для української мови.

Морфологічні словники були доповнені відомими прізвищами та назвами установ і організацій, яких не було в морфологічному словнику.

Для виявлення опорних слів документів побудовано частотний словник, у якому для кожного слова записана кількість його появ у деякому великому масиві документів, а також кількість документів, у яких знайшлося це слово.

Для створення частотного словника був взятий корпус документів за 2007 рік, які скануються з Інтернет системою контент-моніторингу *InfoStream*. Корпус складається з текстів веб-публікацій на українській (1 344 086 документів) і російській мові (2 399 367 документів). При машинному навчанні частотного словника з кожного документа в корпусі витягалися словоформи, які (з певною ймовірнісною похибкою) були приведені до нормальної форми. При цьому підраховувалася кількість, як словоформ, так і нормальних форм у документах, а також підраховувалася кількість документів, у яких зустрілася словоформа і/або нормальна форма.

Для ефективності пошуку опорних слів у результуючі словники входили тільки ті слова, що зустрілися у документальному корпусі більше ніж два рази. Також було вирішено використовувати тільки іменники.

Навчання словника, проходить у три етапи. Перший етап полягає в поділі документів на словоформи й запис отриманих словоформ із інформацією про номер документа, у тимчасовий файл. На вхід програми подається документ, програма розділяє документ на словоформи, і записує все у файл. На другому етапі, створений файл сортується за словоформою та номером документа. Далі підраховується кількість входжень однієї словоформи та кількість документів, у яких вона зустрілася. Знайдені частоти записуються у частотний словник, після чого відбувається пошук нормальної форми. У новому файлі зберігається нормальна форма і номери документів.

При виявленні омонімії у вихідний файл записуються всі нормальні форми відповідній словоформі. Тобто якщо одній словоформі відразу відповідає декілька нормальних форм, зберігаються підраховані частоти з усіма знайденими нормальними формами. На наступному етапі відбувається підрахунок кількості нормальних форм у документах, і збереження результатів у частотний словник.

У рамках даних досліджень використовувалися словники перекладів з російської на українську, і з української на російську мову. Вихідні дані для побудови словників перекладів були отримані шляхом перекладу іменників в нормальній формі існуючими програмами перекладу текстів. Якщо одному слову відповідало декілька перекладів, то вибиралось найбільш уживане значення у відповідності з частотним словником.

Формування опорних слів та їх перекладів

Програма формування опорних слів та їх перекладів завантажує стопсловники для кожної мови. Завантаження стоп-словників проходить у два етапи.

Перший здійснюється при старті програми, а другий коли вибираються опорні слова документа. Перший етап відбувається при завантаженні морфонологічних словників, при цьому відсіваються всі нормальні форми які відповідають даним словам у словниках, що знаходяться в стоп-словнику. Після цього відбувається завантаження словників перекладів мов.

Для кожного документа, який зчитується із вхідного потоку, відбувається його розподіл за словоформами. Після цього відбувається пошук нормальної форми для кожної словоформи. У випадку омонімії, вибирається та нормальна форма, що є найбільш частотною в словнику. Далі відбувається підрахунок кількості словоформ. Опорні слова витягаються за допомогою формули *Okapi BM*25. Після обчислення вагових коефіцієнтів відбувається ранжирування нормальних слів і вибираються найперші дванадцять.

Отримані дванадцять опорних слів перекладаються з однієї мови на іншу за допомогою словників перекладів. Всі опорні слова й слова переклади приписуються до документа.

Експертні оцінки показали, що вдалося домогтися 99% якості при перекладі опорних слів.

Одним з методів оцінки якості витягу опорних слів було виявлення кількості небажаних для перекладу слів (омонімів), що потрапляли до їх складу при перегляді документів різної довжини, створених за допомогою двох різних алгоритмів.

Для цього бралися 1000 довільно обраних документів, які мали різну довжину. Відбувалося обчислення опорних слів відразу за двома алгоритмами та за допомогою експертних оцінок вираховувалась загальна кількість омонімів.

В результаті було виявлено що класичний підхід *TFIDF* поводиться стабільно на відносно малих документах, що містять 100 – 150 слів. При перевищенні даної межі в опорні слова попадали небажані омоніми. На відміну від *TFIDF*, *Okapi BM*25 не витягала небажаних слів практично на всіх документах.

Алгоритм пошуку дублікатів документів

Пошук дублікатів здійснюється у два етапи. На першому етапі проводиться пошук дублів документів різними мовами за допомогою системи *InfoStream*.

Системі подавалося п'ять опорних слів з українських документів, що являли собою, перекладені опорні слова з української на російську мову (рис. 2).

Далі проводиться порівняння поданих опорних слів із дванадцятьма опорними словами документів російською мовою.

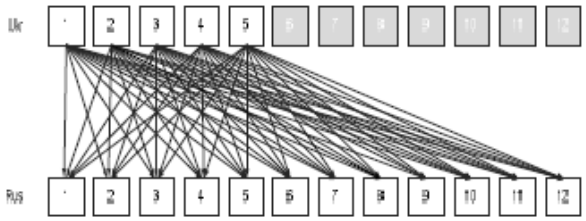

Рис. 2. Порівняння опорних слів

Після цього проводилася фільтрація небажаних, «неповних» дублікатів документів. Для цього були використані такі додаткові критерії відсіювання не повних дублікатів:

- загальна кількість слів у переведеному варіанті не повинна відрізнятися більше ніж на 10%; - кількість слів які починаються з великої букви

(не на початку рядка) неповинно відрізнятися більше чим на 3 слова;

- кількість чисел у документах не повинна вирізнятися більше чим на два;

- знайдені числа в документах не повинні відрізнятися більш ніж на 15 %.

У результаті пошуку дублів новинних документів було створено паралельний двомовний корпус документів [2]. В результаті було побудовано паралельний двомовний корпус веб-публікацій об'ємом близько 30 тис. документів (рис. 3).

З отриманого корпуса документів було вибрано 1000 випадкових документів, які піддалися вивченню експертами.

Аналіз показав, що у середньому 98% змісту кожного документа мають різні доповнення та зміни: наприклад, посилання на інше видавництво, або ж інший заголовок. Також аналіз показав, що з 1000 обраних документів знайшовся один документ, що не зовсім відповідав документу на іншій мові. Відмінність складалася лише у тому, що в документі перекладу були більш докладно описані подробиці первинної статті, а довжина первинної статті була дуже малою – близько 40 слів.

Рис. 3. Інтерфейс системи пошуку у двомовному корпусі

Висновки

Наведений алгоритм дозволяє проводити пошук дублікатів, представлених не тільки мовою, якою був написаний первинний документ, але й іншою мовою. Тобто використовуючи механізм підключення інших мов до системи, можливий пошук дублікатів, представлених відразу на декількома мовами.

Як приклад пошуку дублікатів, авторами було створено двомовний паралельний корпус документів, який налічує близько 30 тис. пар документів, до якого надається вільний доступ.

На практиці не завжди можливо виявити дублікат документу, якщо він був створений на основі об'єднання декількох текстів, ця ситуація найчастіше виникає при пошуку плагіатів. У таких випадках розглянутий алгоритм, якщо його застосовувати без модифікацій, буде малоефективним, але ситуацію можливо поліпшити шляхом ведення пошуку опорних слів не на рівні всього тексту, а на рівні декількох абзаців. Після цього можна без обмежень застосовувати алгоритм, наведений у даній статті.

Список літератури

1. Закон України від 23 грудня 1993 р. № 3792-XII «Про авторське право і суміжні права».

2. *Литвиненко О.Є.* Методи визначення плагіату в електронних документах/ О.Є. Литвиненко, А.В. Шевченко, Є.Б. Артамонов // Зв`язок. – К.: – 2006. – С. 24 –26.

3. *D.V. Lande, V.V. Zhygalo*: About the creation of a parallel bilingual corpora of web-publications, Publication: eprint arXiv: 0807.0311v1.

4. *Чернокозинкий С.А.* Использование текстовых анализаторов для защиты информации в образовательной сфере // Журнал «Информационное противодействие угрозам терроризма», 2005. – № 4, – С. 239–241.

5. *Аушра А.В.* Научная Электронная Библиотека как средство борьбы с плагиатом // "SCIENCE ONLINE: электронные информационные ресурсы для науки и образования", 2006.

6. *W.R. Stone*. Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong! // IEEE Antennas and Propagation, Aug. 2003. –Vol. 45. – №4.