DOI: 10.13140/RG.2.1.3286.2169
УДК 008.5

Білощицький Андрій Олександрович
Доктор технічних наук, завідувач кафедри інформаційних технологій, *ORCID: 0000-0001-9548-1959*
Київський національний університет будівництва і архітектури, Київ
Криштоф Світлана Дмитрівна
Кандидат технічних наук, директор департаменту атестації кадрів вищої кваліфікації МОНУ
Міністерство освіти і науки України, Київ
Білощицька Світлана Василівна
Кандидат технічних наук, доцент кафедри інформаційних технологій проектування та прикладної математики *ORCID: 0000-0002-0856-5474*
Київський національний університет будівництва і архітектури, Київ
Діхтяренко Олександр Васильович
Аспірант кафедри основ інформатики
Київський національний університет будівництва і архітектури, Київ

МЕТОД ВИЛУЧЕННЯ ПОМИЛКОВИХ ЗБІГІВ ТЕКСТІВ В ЕЛЕКТРОННИХ ДОКУМЕНТАХ

Анотація. *Розглянуто модель збігу та метод визначення нечітких збігів у тексті, на їх основі запропоновано метод вилучення помилкових збігів текстів у документах, що перевіряються. Показано, що за рахунок використання методу локально-чутливої хеш-функції знаходження нечітких збігів можна отримати кращий результат, ніж при використанні криптографічної хеш-функції. Оскільки зі збільшенням повноти охоплення точок страждає точність методу, було розроблено метод фільтрації помилкових збігів, який базується на припущенні, що справжні збіги між елементами індексу обов'язково будуть з'являтися на незначній відстані один від одного (відстань – різниця номерів елементів індексу), причому одна група збігів повинна мати незначні відстані як в документі, що перевіряється, так і в документі, з яким перевіряється. Розроблений метод використовує Декартову площину та оптимізований спосіб розрахунку відстаней між елементами для вилучення помилкових результатів і визначення нечітких збігів.*

Ключові слова: хеш-функції; хешування; шингли; перевірка збігів; плагіат

Вступ

З появою Інтернету передача інформації стала простою і загальнодоступною. Однак однією з особливостей інформації, розміщеної в Інтернеті, є те, що її неможливо захистити від копіювання. Доступна в електронному вигляді інформація може використовуватися як база для подальших досліджень, бути відредагована та видана іншим автором за свою. Звісно, проблема плагіату існувала ще задовго до появи Інтернету або електронних документів, але саме зараз займатися плагіатом чужих робіт стало легко як ніколи раніше.

Всі ці фактори негативно впливають на рівень підготовки фахівців, люди все менше намагаються створити щось нове, якщо простіше адаптувати або видати за свою роботу чиюсь іншу. В окремих випадках людина навіть не читає зміст роботи, яку намагається захистити або опублікувати, просто відкориговує назву і змінює автора. Визначити справжнього автора – задача не проста і майже неможлива навіть для фахівця в своїй сфері, адже ніхто не зможе запам'ятати всі наявні роботи в своїй галузі і постійно читати і запам'ятовувати нові. Незважаючи на те, що ця задача непідйомна для звичайної людини, вона може бути вирішена за допомогою комп'ютерної програми. Тому актуальною є розробка ефективних методів та моделей для виявлення нечітких дублікатів в електронних документах і запобігання плагіату, що в свою чергу приведе до підвищення рівня наукових робіт та освіти в цілому. Існує також проблема, коли створені інформаційні системи не можуть знайти збіги в переопрацьованому документі, або знаходять помилкові збіги. Тому не менш важливим є створення ефективних методів вилучення помилкових збігів текстів, і як наслідок, хибних сигналів.

Мета статті

Мета статті – розробка методу пошуку та виявлення помилкових збігів текстів в електронних документах.

Виклад основного матеріалу

Для забезпечення ефективного пошуку збігів в електронних документах, що проходять перевірку на унікальність, необхідно розробити моделі та методи, які б змогли в якісному рерайті, який все одно є плагіатом, знайти збіги з оригінальним текстом.

Модель збігу та метод визначення нечітких збігів в тексті

Нехай T – вхідний текст; T_1, T_2, …,T_r – тексти, що знаходяться в базі даних документів, де r – кількість текстів в базі. Необхідно визначити збіги тексту вхідного документу в текстах кожного з наявних документів T_i, де $i = \overline{1,r}$.

Позначимо через $I(E_d)$ – елементи індексу документу, що являють собою бітові рядки кожного з шинглів E_d, $d = \overline{1, l-h+1}$ вхідного тексту Т. Через $I(E_{d_w}^w)$ – бітові рядки кожного з шинглів $E_{d_w}^w$, текстів $w = \overline{1,r}$, де $d_w = \overline{1, l_w - h + 1}$. l_w – позиція останнього слова документу T_w, $w = \overline{1,r}$. Якщо кожен з бітів позначити через δ, то $I\left(E_{d_w}^w\right) = \left\{\delta_{d_w 1}^w, \delta_{d_w 2}^w, \dots, \delta_{d_w c}^w\right\}$, $d_w = \overline{1, l_w - h + 1}$, де δ – біт; w – номер документу T_w з бази даних; d_w – номер шинглу документу T_w; с – кількість бітів (однакова для всіх), а $I(E_d) = \{\delta_{d1}, \delta_{d2}, \dots, \delta_{dc}\}$, $d = \overline{1, l-h+1}$, де d – номер шинглу документу Т.

Індекс документу представлений множиною $\left\{I(E_1^w), I(E_2^w), \dots, I(E_{l_w-h+1}^w)\right\}$, де l_w – кількість канонізованих слів отриманих з тексту T_w; h – кількість слів у шинглі.

Для пошуку збігів в текстах документів T_w для вхідного документу T необхідно створити індекси цих текстів та обрахувати відстані Хеммінга між усіма елементами:

$$H\left(I(E_d), I(E_{d_w}^w)\right) = \frac{1}{c},$$

$d = \overline{1, l-h+1}$, $d_w = \overline{1, l_w - h + 1}$, $w = \overline{1,r}$.

Після цього треба побудувати множину пар G^w, яка складається з усіх пар елементів індексу, відстань Хеммінга між якими не перевищує порогового значення λ: $G^w = \left\{(d, d_w) | H\left(E_d, E_{d_w}^w\right) < \lambda\right\}$, де λ – поріг, максимально допустиме значення відстані між двома елементами, $i = \overline{1,d}$, $j = \overline{1,d_w}$, $w = \overline{1,r}$, r – кількість документів у базі даних.

Модель збігу показана на рис. 1. Для побудови збігу необхідно визначити його позиції в двох текстах, що містять цей збіг.

Всі позиції вимірюються у кількості символів від початку тексту.

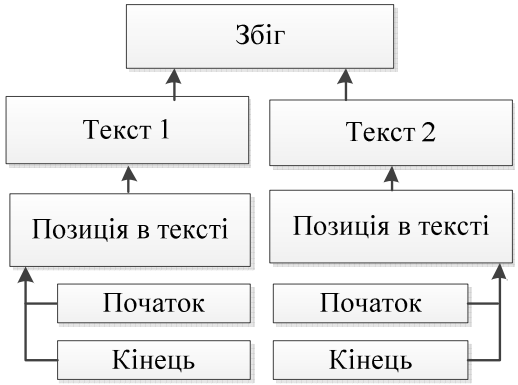


Рисунок 1 – Структурна модель збігу

Метод вилучення помилкових збігів

Метод відкидання помилок в інформаційній технології визначення нечітких збігів ґрунтується на гіпотезі, що справжні збіги повинні знаходитися поряд один з одним. Під словом «поряд» мається на увазі, що відповідні текстові фрагменти шинглів, які збіглися, мають знаходитися на невеликій відстані один від одного в початковому тексті. Відстань можна визначати в словах чи символах. Аналізуючи отримані після перевірки пари збігів, слід також врахувати особливості методу фрагментації тексту, а саме – перекриття частин індексу. З табл. 1 видно, що за заданими параметрами фрагментації достатньо вибирати кожен третій шингл (1, 3, 6 тощо), щоб отримати вихідний текст, що був фрагментований. Це обумовлено тим, що елементи йдуть з перекриттям, наприклад, слова, що увійшли в шингли E_2 та E_4, повністю повторюються в E_1 і E_3.

Отримані шингли E_{l-h+1} являють собою послідовність, що характеризує текст документу, їх вже можна використовувати для пошуку.

Шингли містять інформацію про позицію в тексті (початок і довжину) власного текстового фрагменту, тому для порівняння позицій можна використовувати ці дані, а кожен збіг пар шинглів представити як відрізок на шкалі документу, що має свій початок і кінець відповідно до позицій тексту в початковому варіанті (до попередньої обробки). Необхідно врахувати два фактори: по-перше, фрагментація тексту відбувається по словах, а кількість слів, що входить в кожен шингл, однакова; по-друге, шингли не включають в себе стоп-слова, різного роду розділові знаки, пробіли та невидимі символи.

Таблиця 1 – ***Приклад побудови шинглів***

Номер послідовності	**процес**	**заміна**	**слово**	**синонім**	**випадок**	**канонічний**	**форма**
E_1	процес	заміна	слово				
E_2		заміна	слово	синонім			
E_3			слово	синонім	випадок		
E_4				синонім	випадок	канонічний	
E_5					випадок	канонічний	форма

Тому навіть в сусідніх елементах, відстань від кінця попереднього елемента (обрахована як сума його початку та довжини) і початком наступного може перевищувати розмір самого елемента. Отже, доцільніше і простіше використовувати як характеристику позиції саме номери шинглів, оскільки їх номери відповідають порядку розташування текстових фрагментів, що до них входять.

Наприклад, візьмемо номери пар збігів, що наведені в табл. 2. Дані пари створені автоматично, але їх достатньо для демонстрації роботи.

Таблиця 2 – ***Вигляд пар збігів, отримані після застосування методу пошуку***

d	3	3	4	5	6	11	14	16	17	17	19
d_w	9	15	5	3	7	9	19	10	11	5	8

Отримані пари (d, d_w) розміщуємо на Декартовій площині, значення d_w розмістимо на вертикальній осі, а d – на горизонтальній (рис. 2).

Пройдемо по графіку: по осі абсцис від 0 до d відкидаємо кожну точку d_e, $e \in [0, l-h+1]$, для якої $d_e - d_{e-1} > s$ та $d_{e+1} - d_e > s$, де s – максимальна допустима відстань між точками, для прикладу візьмемо $s = 2$. На рис. 3, *а* показані проекції точок на вісь d та позначені відстані, що не задовольняють умові. Площина розбивається на зони, ознакою межі зони є відстані між точками.

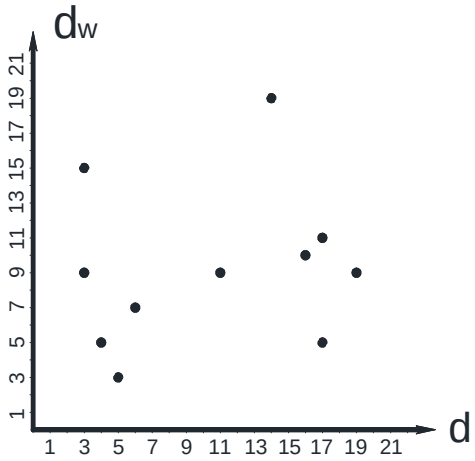
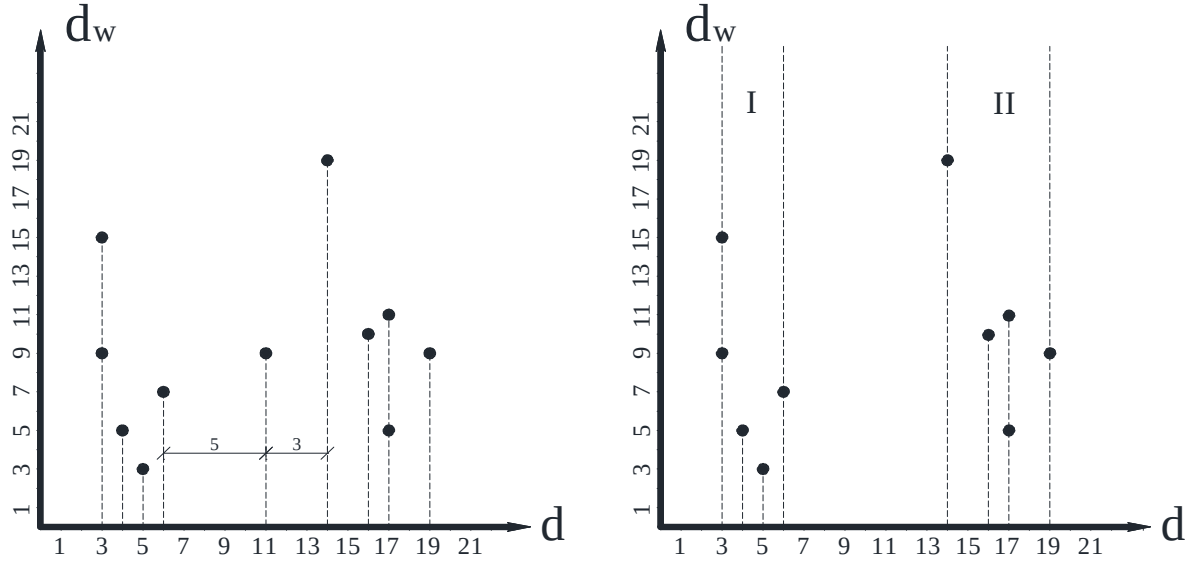


Рисунок 2 – Пари збігів у шинглах розміщені на Декартовій площині

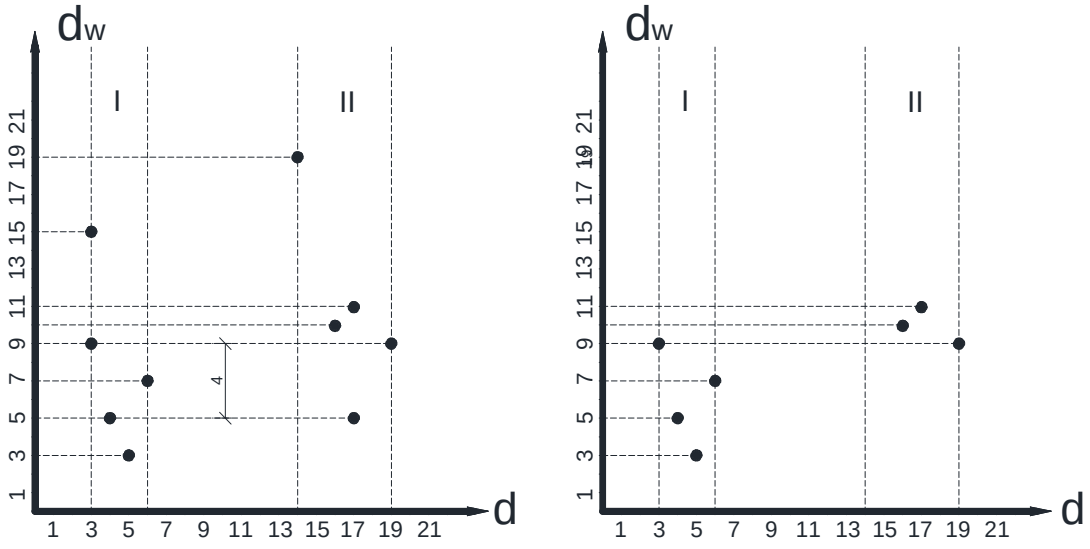
В межах однієї зони відстань між двома сусідніми точками не перевищує s (рис. 3, *б)*. При наступному проходженні дані зони будуть використовуватися як обмеження по відбору сусідніх точок. Тобто, для того щоб дві і більше точки вважалися поряд вони повинні обов'язково знаходитися поряд. Якщо між точками мінімальна відстань, а вони все-одно знаходяться в різних зонах – це свідчить, що між точками значна відстань по осі d_{w0}.

Аналогічно, йдучи по осі d_w від 0 і до останньої точки, відкидаємо кожну точку d_{we}, $e_w \in [0, l_w - h + 1]$, для якої $d_{w,e_w} - d_{w,e_w-1} > s$ та $d_{we_w+1} - d_{we_w} > s$ (рис. 3, *а*, *б)*.



а *б*

Рисунок 3 – Визначення відстаней між точками та зонування по осі d



а *б*

Рисунок 4 – Визначення відстаней між точками d_w

Істотна відмінність від попереднього вилучення це те, що для порівняння відстаней беремо лише ті точки, які знаходяться в одній і тій же зоні (рис. 4, *а*). Основна мета такого вилучення – залишити скупчення точок, що знаходяться поряд як по осі *d* так і по d_w. Точки, що знаходяться в різних зонах, можуть бути поряд відносно осі d_w, але на великій відстані при проекції на вісь *d*.

Інший варіант виконання подібної фільтрації – знаходження відстаней між кожною парою точок та вилучення точок, для яких всі відстані перевищують максимальне значення. Описаний спосіб вирішує ту саму задачу, але процес дещо оптимізований. Якщо для розрахунку всіх відстаней необхідно вирахувати $\frac{p^2-p}{2}$ відстаней, де p – кількість точок на площині, то за наведеним способом буде лише $2p-2$ відстані. Тому наведений спосіб ефективніший при застосуванні на площинах з великою кількістю точок.

Об’єднавши скупчення точок у групи, отримаємо 2 збіги та їх проекції на документи T, і T_w (рис. 5).

Наведені групи мають такі характеристики: розмір групи (визначається кількістю точок, що входять до групи), проекція на вісь *d*, проекція на вісь d_w. Для першої групи розмір дорівнює 4 (точок), проекція на вісь *d*: 3-6, на вісь d_w: 3-9. Для другої розмір – 3, проекція на d – 16-19, проекція на d_w – 9-11. Фактично, це означає, що знайдено два збіги: перший збіг показує, що фрагмент документу d від 3 по 6-й шингл збігається з фрагментом документу d_w від 3 по 9-й шингл, аналогічно другий збіг показує інший фрагмент.

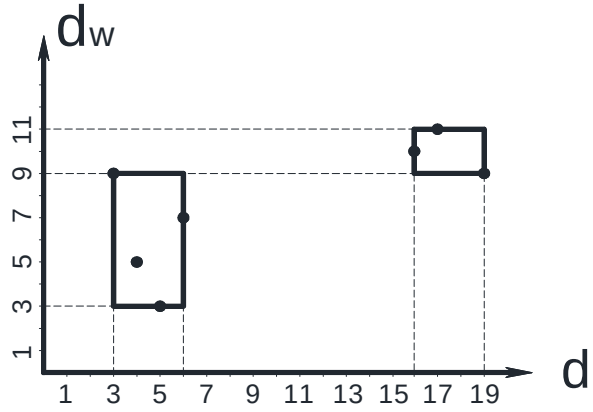


Рисунок 5 – Об’єднані групи збігів

На даному етапі необхідно також відкинути групи малого розміру. Мінімальний розмір групи вибираємо залежно від мети пошуку: знайти всі, навіть незначні збіги або лише вагомі фрагменти.

Координати кожної точки являють собою номер шинглу у відповідному документі. Розглянемо перший збіг відносно вхідного документу *T*. Початок збігу відповідає шинглу E_3, кінець – E_6. У свою чергу, за умови, що крок побудови шинглів дорівнює одиниці, а довжина шинглу – 3, ці шингли мають вигляд:

$$E_3 = \left\{\bar{S}_{k_3}^{\bar{v}_{k_3}}, \bar{S}_{k_4}^{\bar{v}_{k_4}}, \bar{S}_{k_5}^{\bar{v}_{k_5}}\right\},\ E_6 = \left\{\bar{S}_{k_6}^{\bar{v}_{k_6}}, \bar{S}_{k_7}^{\bar{v}_{k_7}}, \bar{S}_{k_8}^{\bar{v}_{k_8}}\right\}.$$

Нас цікавить перший елемент першого шинглу і останній елемент другого шинглу, це елементи $\bar{S}_{k_3}^{\bar{v}_{k_3}}$ і $\bar{S}_{k_8}^{\bar{v}_{k_8}}$. З цих двох елементів ми отримуємо координати слів у тексті $\bar{v}_{k_3}$та $\bar{v}_{k_8}$ відповідно.

Дані координати дорівнюють $\bar{v}_{k_i} = (\alpha_{k_i}, \beta_{k_i})$, де α – позиція слова в тексті; β – довжина слова (кількість символів). Відповідно позиція першого слова буде відповідати позиції початку збігу. Сума позиції та довжини останнього слова $\alpha + \beta$ буде відповідати кінцю збігу. Таким чином було визначено конкретне місцезнаходження збігу у документі T і побудовано сам збіг. Для документу T_w позиція збігу вираховується аналогічно.

Висновки

Якісна реалізація операції канонізації тексту (морфологічного розбору та приведення до одного синоніма) разом із застосуванням модифікованого алгоритму шинглів для пошуку збігів у текстах повинна давати хороші результати, подібні до справжньої перевірки текстів за змістом. З огляду на те, що морфологічний аналіз текстів є складною процедурою, яка вимагає значних зусиль для реалізації систем пошуку збігів, доцільніше використовувати простіші технології, комбінуючи їх з методами нечіткого пошуку.

Обов'язковим пунктом модифікованого методу є перевірка помилок. Крім видимих помилок, помилки можуть бути непомітні для людини. До «невидимих» помилок належить заміна літер кирилиці на аналогічні латиниці та розрив слів на частини за допомогою невидимих символів. Такі елементи не грають ніякої ролі для людини, хоча насправді вносять значні структурні зміни.

Для боротьби з явищами синонімізації та переписування тексту своїми словами використовується один з видів морфологічного аналізу, визначення роду, числа, частини мови. Після визначення даних параметрів стає можливим зведення слова до єдиної форми (канонізація) та визначення його синонімів. Відбувається визначення антонімів для слів, яким передує частинка «не» або які починаються з «не». Припускається, що автор тексту може використати антоніми для приховування збігів, тому наведені слова замінюються на рівнозначні за змістом антоніми. Якщо в оригінальному тексті використовуються слова з частинками «не» і вони без змін були скопійовані в інший документ – такий збіг буде все-одно знайдено, тому що попередню обробку проходять обидва документи і для обох вона однакова.

Список літератури

1. *Закон України «Про авторське право і суміжні права» № 3729-12 від 05.12.2012, підстава 5460-17*
2. *Білощицький, А.О. Ефективність методів пошуку збігів у текстах / А.О. Білощицький, О.В. Діхтяренко // Управління розвитком складних систем. – 2013. – № 14.– С. 144 – 147.*
3. *Ke, Y., Sukthankar, R., Huston, L., Ke, Y., & Sukthankar, R. (2004, October). Efficient near-duplicatedetectionand sub-imageretrieval. In ACM Multimedia (Vol. 4, No. 1, p. 5).*
4. *Lv, X., & Wang, Z. J. (2012). Perceptual image hashing based on shape contexts and local feature points. Information Forensics and Security, IEEE Transactions on, 7(3), 1081-1093.*
5. *Hamming, R. W. (1950). Error detecting and error correcting codes. Bell System technical journal, 29(2), 147-160.*
6. *Chum, O., Philbin, J., &Zisserman, A. (2008, September). NearDuplicateImageDetection: min-Hash and tf-idfWeighting. In BMVC (Vol. 810, pp. 812-815).*
7. *Leskovec, J., Rajaraman, A., &Ullman, J. D. (2014). Miningofmassivedatasets. CambridgeUniversityPress.*
8. *Platter W., Phashion, (2014), GitHubrepository, https://github.com/westonplatter/phashion*
9. *Білощицький А.О. Оптимізація системи пошуку збігів за допомогою використання алгоритмів локально чутливого хешування наборів текстових даних/ А.О. Білощицький, О.В. Діхтяренко // Управління розвитком складних систем. – 2014. – № 19.– С. 113 – 117.*
10. *Богунский, В.Д. Обоснование закона о конкурентных свойствах проектов / В.Д. Богунский, С.В. Руденко, П.А. Тесленко // Управління розвитком складних систем. –2011. – № 8. – С. 14 – 16.*
11. *Оборський, Г.О. Стандартизація і сертифікація процесів управління якістю освіти у вищому навчальному закладі / Г.О. Оборський, В.Д. Богунський, О.С. Савельєва // Тр. Одес. политехн. ун-та. –2011. – № 1(35). – С. 251 – 255.*
12. *Колесникова, Е.В. Моделирование слабо структурированных систем проектного управления / Е.В. Колесникова // Тр.Одес. политехн. ун-та. – 2013. - № 3 (42). – С. 127 – 131.*
13. *Колесникова, Е.В. Трансформация когнитивных карт в модели марковских процессов для проектов создания программного обеспечения / Е.В. Колесникова, А.А. Негри // Управління розвитком складних систем. - 2013. – №15. – С. 30 – 35.*
14. *Vaysman, V. A. The planar graphs closed cycles determination method / V. A. Vaysman, D. V. Lukianov, K. V. Kolesnikova // Тр. Одес. политехн. ун-та. – 2012. – № 1(38). – С. 222 – 227.*
15. *Burkov, V. N., Biloshchytskyi, A. A., & Gogunsky, V. D. (2013). Options citation of scientific publications in scientometric databases. Management of development of difficult systems. Kyiv, Ukraine: KNUCA, 15, 134 - 139.*
16. *Gogunsky, V. D., Kolyada, A. S., & Iakovenko, V. O. (2014). Scientometric data scientific publication "Management of development of difficult systems. Management of development of difficult systems. Kyiv, Ukraine: KNUCA, 19, 6 – 11.*

17. Vlasenko, O. V., Lebed' V. V., & Gogunsky, V. D. (2012). Markov model of communication processes in international projects. Management of development of difficult systems. Kyiv, Ukraine: KNUCA: 12, 35 - 39.

18. Gogunsky, V. D., Iakovenko, V. O., & Kolyada, A. S. (2014). Application of Latent Dirichlet allocation for the analysis of scientometric publications database. Proc. of Odes. Polytechnic. Univ. Odessa, Ukraine, ONPU: 1 (43), 186 – 191.

Стаття надійшла до редколегії 15.04.2015

Рецензент: д-р техн. наук, проф. С.Д. Бушуєв, Київський національний університет будівництва і архітектури, Київ.

Белощицкий Андрей Александрович
Доктор технических наук, профессор кафедры информационных технологий, *ORCID: 0000-0001-9548-1959*
Киевский национальный университет строительства и архитектуры, Киев
Крыштоф Светлана Дмитриевна
Кандидат технических наук, директор департамента аттестации кадров высшей квалификации МОНУ
Министерства образования и науки Украины, Киев
Белощицкая Светлана Васильевна
Кандидат технических наук, доцент кафедры информационных технологий проектирования и прикладной математики, *ORCID: 0000-0002-0856-5474*
Киевский национальный университет строительства и архитектуры, Киев
Дихтяренко Александр Васильевич
Аспирант кафедры основ информатики
Киевский национальный университет строительства и архитектуры, Киев

МЕТОД ИСКЛЮЧЕНИЯ ОШИБОЧНЫХ СОВПАДЕНИЙ ТЕКСТОВ В ЭЛЕКТРОННЫХ ДОКУМЕНТАХ

***Аннотация**. Рассмотрена модель совпадения и метод определения нечетких совпадений в тексте, на их основе предложен метод исключения ложных совпадений текстов в проверяемых документах. Показано что за счет использования метода локально-чувствительной хэш-функции нахождения нечетких совпадений можно получить лучший результат, чем при использовании криптографической хэш-функции. Поскольку с увеличением полноты страдает точность метода, был разработан метод фильтрации ложных совпадений, который базируется на предположении, что настоящие совпадения между элементами индекса обязательно будут появляться на незначительном расстоянии друг от друга (расстояние - разность номеров элементов индекса), причем одна группа совпадений должна иметь незначительные расстояния как в документе, который проверяется, так и в документе, с которым проверяется. Разработанный метод использует Декартову плоскость и оптимизированный способ подсчета расстояний между элементами для отбрасывания ложных результатов и определения нечетких совпадений.*

Ключевые слова: хэш-функции; хеширование; шинглы; проверка совпадений; плагиат

Biloshchytskyi Andrii
Doctor of Technical Sciences, Head of information technology, *ORCID: 0000-0001-9548-1959*
Kyiv National University of Construction and Architecture, Kiev
Kristof Svitlana
Ph.D., director of the department of certification personnel of higher qualification
The Ministry of Education and Science of Ukraine, Kiev
Biloshchytska Svitlana
Ph.D., assistant professor of information technology designing and applied mathematics, *ORCID: 0000-0002-0856-5474*
Kyiv National University of Construction and Architecture, Kiev
Dikhtiarenko Oleksandr
Postgraduate at the Department of computer science fundamentals
Kyiv National University of Construction and Architecture, Kiev

THE METHOD OF ELIMINATION OF ERRONEOUS COINCIDENCES TEXT IN ELECTRONIC DOCUMENTS

***Abstract**. The article describes a model and matching method for determining fuzzy matches in the text on the basis of their proposed method of extracting false matches text in scanned documents. It is shown that by using the method of locally sensitive hash finding fuzzy matches will get better results than using a cryptographic hash function. But with the increasing suffering of completeness accuracy of the method. Therefore, we developed a method of filtration of false matches, which is based on the assumption that these coincidences between the elements of the index is required to appear at a slight distance from each other (the distance - the difference between the numbers of elements of the index), and one group matches should have a slight distance in the document that checked, and the document, which is checked. The developed method uses a Cartesian plane and optimized method of calculating the distance between the elements to discard false positives and identify fuzzy matches.*

The method of extraction of false matches text in a document scanned in information technology determining fuzzy matches based on the hypothesis that these matches have to be next to each other. The word "near" is meant that the relevant text fragments shingles that match, should be at a small distance from each other in the source code.

Keywords: a hash-function; hashing; shingles; test matches; plagiarism

References

1. The law of Ukraine on copyright and related rights № 3729-12 on 05.12.2012

2. Biloshchytskyi, A., & Dikhtiarenko, O. (2013). The effectiveness of methods for finding matches in texts. Management of complex systems, 14, pp. 144 – 147.

3. Ke, Y., Sukthankar, R., Huston, L., Ke, Y., & Sukthankar, R. (2004, October). Efficient near-duplicate detection and sub-image retrieval. In ACM Multimedia (Vol. 4, No. 1, p. 5).

4. Lv, X., & Wang, Z. J. (2012). Perceptual image hashing based on shape contexts and local feature points. Information Forensics and Security, IEEE Transactions on, 7(3), 1081-1093.

5. Hamming, R. W. (1950). Error detecting and error correcting codes. Bell System technical journal, 29(2), 147-160.

6. Chum, O., Philbin, J., & Zisserman, A. (2008, September). Near Duplicate Image Detection: min-Hash and tf-idf Weighting. In BMVC (Vol. 810, pp. 812-815).

7. Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of massive datasets. Cambridge University Press.

8. Platter W., & Phashion, (2014), GitHub repository, https://github.com/westonplatter/phashion.

9. Biloshchytskyi, A., & Dikhtiarenko, O. (2014). Optimization of Matching algorithms by using local-sensitive hash sets of text data. Management of complex systems, 19, pp. 113 – 117.

10. Gogunsky, V. D., Rudenko, S. V., & Teslenko, P. A. (2012). Justification law on competitive properties of projects. Management of development of difficult systems. Kyiv, Ukraine, KNUCA: 8, 14 - 16.

11. Oborsky, G. A., Gogunsky, V. D., & Saveleva O. S. (2011). Standardization and certification processes of the quality management education in higher education. Proceedings of Odes. Polytechnic. Univ, 1 (35), 251 – 255.

12. Kolesnikova, K. V. (2013). Modeling weakly structured project management systems. Proceedings of Odes. Polytechnic. Univ, 3 (42), 127 – 131.

13. Kolesnikova, K. V., & Negri, A. A. (2013). Transformation of cognitive maps in the model of Markov processes for projects creating software. Management of development of difficult systems. Kyiv, Ukraine: KNUCA, 15, 30 – 35.

14. Vaysman, V. A. Lukianov, D. V. & Kolesnikova, K. V. (2012). The planar graphs closed cycles determination method. Proceedings of Odes. Polytechnic. Univ, 1(38), 222 – 227.

15. Burkov, V. N., Biloshchytskyi, A. A., & Gogunsky, V. D. (2013). Options citation of scientific publications in scientometric databases. Management of development of difficult systems. Kyiv, Ukraine: KNUCA, 15, 134 – 139.

16. Gogunsky, V. D., Kolyada, A. S., & Iakovenko, V. O. (2014). Scientometric data scientific publication "Management of development of difficult systems. Management of development of difficult systems. Kyiv, Ukraine: KNUCA, 19, 6 – 11.

17. Vlasenko, O. V., Lebed' V. V., & Gogunsky, V. D (2012). Markov model of communication processes in international projects. Management of development of difficult systems. Kyiv, Ukraine: KNUCA: 12, 35 - 39.

18. Gogunsky, V. D., Iakovenko, V. O., & Kolyada, A. S. (2014). Application of Latent Dirichlet allocation for the analysis of scientometric publications database. Proc. of Odes. Polytechnic. Univ. Odessa, Ukraine, ONPU: 1 (43), 186 – 191.

Посилання на публікацію

APA *Biloshchytskyi, A., Kristof, S., Biloshchytska, S., & Dikhtiarenko, O. (2015). The method of elimination of erroneous coincidences text in electronic documents. Management of Development of Complex Systems, Issue 22 (1), P. 144 – 150. dx.doi.org\10.13140/RG.2.1.3286.2169*

ГОСТ *Білощицький А.О. Метод вилучення помилкових збігів текстів в електронних документах [Текст] / А.О. Білощицький, С.Д. Криштоф, С.В. Білощицька, О.В. Діхтяренко // Управління розвитком складних систем. – 2015. - № 22(1). – С. 144 - 150. dx.doi.org\10.13140/RG.2.1.3286.2169*