УДК 004.021

КЛАСТЕРИЗАЦІЯ ВЕБ-ДОКУМЕНТІВ НА ОСНОВІ АЛГОРИТМУ ХЕШУВАННЯ

А. О. Осідач
Національний університет «Львівська Політехніка»
вул. Степана Бандери,12, м. Львів, 7900, Україна. E-mail: osidach.publications@gmail.com

Розроблено структурну схему процесу кластеризації веб-документів на основі алгоритму хешування для отримання частих наборів. Розглянуто новий алгоритм – глобальний алгоритм хешування для отримання частих наборів, з метою прискорення процесу видобутку (отримання), а також масштабування документів, незалежно від їх розміру. Проведено оцінку ефективності запропонованого алгоритму, яка представлена результатом порівняння глобального алгоритму хешування для отримання частих наборів з алгоритмом Apriori, підхід заснований на кластеризації веб-документів. Оцінка тексту за алгоритмом кластеризації проведена за допомогою методу k-середніх і FIHC методу. Доведено, що метод кластеризації веб-документів на основі алгоритму хешування для отримання частих наборів більш ефективний, ніж інші підходи кластеризації (FIHC та k-середніх) на основі результатів порівняння їх ефективності. Запропоновано механізм кластеризації веб-документів на основі алгоритму хешування для отримання частих наборів, що забезпечує значне зниження розмірності.

Ключові слова: кластеризація, веб-документ, хешування, алгоритм, індексація, асоціативні правила, інтелектуальний аналіз.

КЛАСТЕРИЗАЦИЯ ВЕБ-ДОКУМЕНТОВ НА ОСНОВЕ АЛГОРИТМА ХЕШИРОВАНИЯ

А. О. Осидач
Национальный университет «Львовская Политехника»
ул. Степана Бандеры, 12, г. Львов, 7900, Украина. E-mail: osidach.publications@gmail.com

Разработана структурная схема процесса кластеризации веб-документов на основе алгоритма хеширования для получения частых наборов. Рассмотрен новый алгоритм – глобальный алгоритм хеширования для получения частых наборов, с целью ускорения процесса добычи (получения), а также масштабирования документов, независимо от их размера. Проведена оценка эффективности предложенного алгоритма, которая представлена результатом сравнения глобального алгоритма хеширования для получения частых наборов и алгоритма Apriori, подход основан на кластеризации веб-документов. Оценка текста по алгоритму кластеризации проведена с помощью метода k-средних и FIHC метода. Доказано, что метод кластеризации веб-документов на основе алгоритма хеширования для получения частых наборов более эффективен, чем другие подходы кластеризации (FIHC и k-средних) на основе результатов сравнения их эффективности. Предложен механизм кластеризации веб-документов на основе алгоритма хеширования для получения частых наборов, обеспечивает значительное снижение размерности.

Ключевые слова: кластеризация, веб-документ, хеширование, алгоритм, индексация, ассоциативные правила, интеллектуальный анализ.

АКТУАЛЬНІСТЬ РОБОТИ. З огляду на стрімке зростання популяризації мережі Інтернет як засобу поширення інформації і, як наслідок, вибухового зростання обсягу контенту, важливо відзначити проблеми, які виникають на рівні користувача (пошук інформації) і контент-провайдера (класифікація та індексація документів).

На сьогодні, онлайн-бібліотеки, пошукові системи та інші великі сховища даних надзвичайно швидко розширюються, що викликає труднощі при класифікації документів вручну. З метою розв'язання цієї проблеми протягом останніх десять років учені досить активно підходять до питань вивчення процесу автоматизації [1–3].

Для вирішення проблеми інформаційного перевантаження в мережі Інтернет, кластеризації та класифікації розглянуто корисні й активні галузі машинного навчання, що обіцяють подолати цю проблему.

У роботі запропоновано підхід до вибору системи захисту електронного документообігу на основі визначення і розрахунку показника її ефективності. Розроблено способи та рекомендації щодо розробки системи захисту електронного документообігу. Досліджено базові механізми захисту електронного документообігу [4].

Процес кластеризації – це організація текстових документів в кластери або групи, іншими словами, документи в одному кластері поділяють ту саму тему, в різних – різні. Це відрізняє кластеризацію документів від класифікації, оскільки відкидає етап підготовки документів за допомогою маркерів.

Висока розмірність ознакового простору вважається основною характеристикою кластеризації документів і є значною проблемою для продуктивності алгоритмів кластеризації. Також труднощі полягають в тому, що не всі функції важливі для кластеризації документів, деякі функції можуть бути зайвими або недоречними, а деякі можуть навіть змінювати результат кластеризації [5].

Знаходження частих наборів є важливим для інтелектуального аналізу документів, і вона бере свій початок від пошуку асоціативних правил. Основний недолік частих наборів в тому, що вони досить великі за обсягом і зберігати їх на комп'ютері неефективно та недоцільно.

На сьогодні розробки в сфері автоматизованої кластеризації документів представлені широким спектром, проте важливо відзначити, що механізм кластеризації веб-документів потребує ретельного дослідження. У період дослідження наукових розробок в сфері автоматизованої кластеризації докумен-

тів слід відзначити високий рівень досягнень сучасних учених.

У роботі [6] був запропонований новий критерій кластеризації транзакцій за допомогою функції відстані. В принципі, цей метод також може бути застосований для кластеризації документів шляхом обробки документа у вигляді транзакцій; однак цей спосіб не приведе до створення ієрархії для перегляду. Новизна такого підходу полягає в тому, що він використовує частий набір елементів (шляхом застосування алгоритму Apriori) для визначення кластера, організацію кластерної ієрархії і зменшення розмірності комплектів документів.

Для вимірювання цілісності того чи іншого кластера безпосередньо за допомогою частих наборів в [1] запропоновано алгоритм ієрархічної кластеризації на основі частих наборів (FIHC). Проте FIHC має низку недоліків в практичному застосуванні.

Алгоритм кластеризації на основі частих наборів запропонований в [7]. Спочатку документи позначаються в векторному просторі моделі і кожен термін відсортований відповідно до його відносної частоти. Потім часті набори визначаються за допомогою закономірності зростання. Нарешті, документи групуються на основі цих частих наборів. Цей підхід ефективний для дуже великих баз даних і дає чітке визначення кластерів по їхніх частих наборах.

Однак, незважаючи на масштабність наукових досліджень щодо автоматизованої кластеризації документів, слід зазначити, що всі розглянуті алгоритми кластеризації ґрунтуються на алгоритмі Apriori, який є витратним за часом реалізації, а час реалізації, як відомо, основний фактор в процесі кластеризації та класифікації документів.

Метою роботи є розробка структурної схеми процесу кластеризації веб-документів на основі алгоритму хешування для отримання частих наборів та оцінка ефективності запропонованого алгоритму.

МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ. Виокремлення етапів процесу кластеризації веб-документів на основі алгоритму хешування. Кластеризації веб-документів на основі алгоритму хешування для отримання частих наборів (КВДХЧН) складається з чотирьох основних етапів:

– попередня обробка документів;
– отримання частих наборів;
– кластеризація документів;
– пост-обробка.

Основною характеристикою цього підходу є те, що в ньому реалізовано новий алгоритм інтелектуального аналізу даних для генерації набору з метою подолати недоліки алгоритму Apriori (алгоритм пошуку асоціативних правил). Крім того, він прискорює процес кластеризації та інтелектуального аналізу.

Структурна схема процесу кластеризації веб-документів на основі алгоритму хешування для отримання частих наборів (КВДХЧН) запропонована на рис. 1.

Рисунок 1 – Структурна схема процесу кластеризації веб-документів на основі алгоритму хешування для отримання частих наборів

Перший етап – етап попередньої підготовки включає в себе декілька кроків попередньої обробки, включаючи видалення стоп-слів, подібних (споріднених) слів й індексації шляхом застосування tf*ідентифікатора:

Видалення стоп-слів: стоп-слова – це слова, які не несуть певної інформації, до них належать (the, a, in, of тощо). Процес видалення стоп-слів необхідний з метою зниження рівня шуму. Однією з основних властивостей стоп-слів є те, що вони надзвичайно поширені. Основними перевагами видалення стоп-слів можна назвати: економію величезної кількості місця, зниження шуму і збереження кістяка слова. В подальшому, це призводить до більш ефективної та дієвої обробки.

Видалення подібних (споріднених, однокореневих) слів. Як правило, процес вилучення коренів відбувається так, що слова перетворюються на їх кореневу форму. Наприклад: підключений і підключення зв'язку, буде перетворено на підключити.

Хороший парадигматичний модуль повинен вміти перетворювати різні синтаксичні форми слова в нормалізованому вигляді, знижувати кількість індексних термінів, з метою економії пам'яті, та, деякою мірою, збільшити продуктивність алгоритмів кластеризації. Портер парадигматичного модуля – метод, який широко застосовується для стовбурових документів [8]. Він компактний, простий і відносно точний та не вимагає створення суфікса списку. В рамках цієї наукової роботи застосовуємо портер парадигматичного модуля для попередньої обробки.

Індексація шляхом застосування tf*ідентифікатора.

$$W_{i,j} = tf_{i,j} \cdot idf_i, \qquad (1)$$

де $W_{i,j}$ – загальна вага термінів, $tf_{i,f}$ – частота терміна, idf_i – зворотна частота документа.

Частота терміна (ЧТ) є функцією кількості входжень конкретного слова в документі, поділена на кількість слів у всьому документі. Слова, які з'являються часто в тексті, вважаються більш важливими для опису контенту, ніж слова, які з'являються рідше. Існує безліч варіантів застосування ЧТ:

$$idf_i = \log\left(N \big/ Nt_j\right), \qquad (2)$$

де Nt_j – кількість документів у колекції N, в якій t_j відбувається принаймні один раз.

Як тільки схему зважування обрано, автоматизоване індексування може бути виконано, просто вибравши кращі K-слова, що задовільняють задану вагу обмеження для кожного документа. Основною перевагою автоматизованої процедури індексування є те, що вона скорочує витрати на індексацію.

Другий етап – це отримання частих наборів. Метою отримання частих наборів інтелектуального аналізу є виявлення наборів, які часто поєднуються в документі. Проблема нетривіальна в текстових документах, тому що документи можуть бути дуже великими, складатися з багатьох елементів і містити набори високої кардинальності. Хоча в алгоритмі Apriori, його ще використовують для генерації частих наборів, такі набори використовуються при кластеризації.

Для того, щоб прискорити процес видобутку (отримання), а також масштабування документів, незалежно від їх розміру, розглянемо новий алгоритм – глобальний алгоритм хешування для отримання частих наборів (ГХЧН). Він докорінно відрізняється від усіх попередніх алгоритмів, оскільки долає недоліки алгоритму Apriori шляхом використання силової структури даних, так званої глобальної хеш-таблиці. Крім того, він використовує нову методологію для формування частих наборів шляхом побудови хеш-таблиці, в ході перевірки документів, тільки один раз, відповідно, кількість операцій сканувань документів зменшується.

Хеш-таблиця – це структура даних, яка прискорює пошук інформації по конкретному аспекту цієї інформації, яка називається ключем. Ідея, що лежить в основі хеш-таблиці, полягає в тому, щоб обробити ключ з функцією, яка повертає хеш-значення; хеш-значення визначає, де в структурі даних буде (або можливо буде) зберігатися. Хеш-таблиці можуть забезпечити постійний час O(1) пошуку в середньому, незалежно від кількості термінів у таблиці.

Складається така таблиця з двох основних компонентів: масиву і хеш-функції.

а) масив – це масив U з розміру R, де під кожна комірка розглядається як набір, а ціле число R визначає ємність масиву;

б) хеш-функція – це друга частина хеш-таблиці, її структура є функцією відносно ключа, розподіленого в діапазоні [0, R-1], де R-ємність масиву для цієї таблиці.

Третій етап – етап кластеризації документів. Кластеризація документів на основі частих наборів вважається фундаментом алгоритму, який підбирає ядро слова з певними критеріями і групи документів на основі цих ключових слів. Цей підхід включає три основних етапи:

- побудова початкових розділів;
- групування розділів на основі документів;
- кластеризації документів за принципом подібності.

Четвертий етап – пост-обробка. Включає в себе основні додатки, наприклад, додаток рекомендацій, який використовує результати кластеризації для рекомендацій користувачам.

Оцінка ефективності алгоритму хешування для частих наборів. Оцінка ефективності запропонованого алгоритму представлена результатом порівняння ГХЧН алгоритму та алгоритму Apriori, підхід ґрунтується на кластеризації веб-документів.

В якості набору даних використана база даних найбільшого агентства України УНІАН [9]. База даних міститься в 22 файлах. Перший файл містить 1000 документів, в той час як останній містить 580 документів. Документи, розподілені за допомогою sgml-тегів. Існує 5 категорій, кожна категорія має підкатегорію, в цілому 670 підкатегорій. З цих документів збираємо допустимі текстові дані кожної категорії шляхом вилучення тексту, який знаходиться між тегами <Body> </ Body> і поміщений в текстовий документ відповідно до його назви.

В якості методів оцінки використана F-міра або міра Ван Різбергена. F-міра – це сукупність точності і повноти концепції інформаційного пошуку [10]. Це відношення кількості релевантних документів, отриманих для запиту до загальної кількості релевантних документів в колекції:

$$\mathrm{Re}call\left(K_i, C_j\right) = n_{ij} / |K_i| \cdot \qquad (3)$$

Точність-відношення кількості релевантних документів до загальної кількості документів, які вимагали для запиту:

$$\mathrm{Pr}ecision\left(K_i, C_j\right) = \frac{n_{ij}}{|C_j|} \cdot \qquad (4)$$

Водночас F-міра для кластера і класу обчислюється як:

$$F\left(K_i, C_j\right) = \frac{2 \cdot \mathrm{Re}call\left(K_i, C_j\right) \cdot \mathrm{Pr}ecision\left(K_i, C_j\right)}{\mathrm{Re}call\left(K_i, C_j\right) \cdot \mathrm{Pr}ecision\left(K_i, C_j\right)}, \qquad (5)$$

де n_{ij} – кількість членів класу K_i в кластері C_j; $|C_j|$ – кількість членів кластера C_j і кількість членів класу K_i.

Зважена сума всіх максимальних F-мір для всіх природних класів використовується для оцінки якості кластеризації результату C. Ця міра називається загальною F-мірою для C, позначається F (C) і обчислюється як:

$$F(C)=\sum_{K_i\in K}\frac{|K_i|}{|D|}\max_{C_j\in C}\{F(K_i,C_j)\}, \qquad (6)$$

де K – всі природні класи, C – кластери на всіх рівнях, $|K_i|$ – кількість документів у природному класі K_i, $|D|$ – загальна кількість документів у наборі даних.

Діапазон становить [0,1]. Більше значення вказує на більш високу точність кластеризації.

Оцінювалися алгоритм ГХЧН і Apriori алгоритм з точки зору їх ефективності і масштабованості. Алгоритм Apriori чутливий до мінімального рівня підтримки і розміру документів. Якщо мінімальна підтримка зменшилася, час роботи Apriori збільшується, оскільки є більш часті набори елементів. Крім того, коли розмір документів стає дуже великим, алгоритм потребує багато часу для сканування документів, виробляючи часті набори з мінімальною підтримкою. У ГХЧН алгоритмі час витрачається на побудову хеш-таблиці тільки один раз. Після збереження в хеш-таблиці даних, немає жодних труднощів у створенні нових різних частих наборів з різним мінімальним порогом підтримки.

На рис. 2 показано порівняння результатів роботи алгоритму ГХЧН і Apriori алгоритму для різних значень мінімальної підтримки порогів, відповідно до виділеного набору даних. Підтримка береться до уваги в якості осі X і час, витрачений на знаходження частих наборів, береться як Y-вісь.

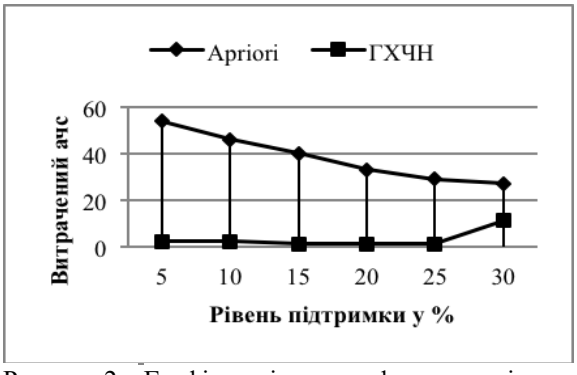

Рисунок 2 – Графік порівняння ефективності двох алгоритмів у співвідношенні витраченого часу до рівня підтримки

З діаграми видно, що при використанні алгоритму ГХЧН витрачений час зменшується пропорційно збільшенню рівня підтримки під час порівняння з алгоритмом Apriori. У ГХЧН алгоритмі, в цілому, час виконання значний тільки при первинному створенні хеш-таблиці. При введенні нового рівня мінімальної підтримки ніякого збільшення витраченого часу не відбувається, час витрачається тільки на пошук в хеш-таблиці. При використанні алгоритму Apriori витрачений час виконання зменшується при збільшенні рівня підтримки. В процесі виконання Apriori алгоритму, щоразу, заходячи на нову мінімальну підтримку, потрібно повторити процес знаходження наборів даних спочатку. Виходячи з проведеного дослідження, робимо висновок, що ГХЧН алгоритм є значно ефективнішим, ніж алгоритм Apriori у всіх випадках щодо великих за обсягом документів, оскільки складність знаходження частих наборів в алгоритму ГХЧН нижча, ніж у Apriori алгоритму.

Для того, щоб визначитися з рівнем масштабованості обох алгоритмів, створюємо новий набір дубльованих файлів, близько 10 000 документів.

На рис. 3 проілюструємо результати застосування алгоритму ГХЧН і Apriori з порогом мінімальної підтримки 15 %, щоб гарантувати, що згенеровані часті набори в обох алгоритмах приблизно однакові. Слід зазначити, що на цьому етапі аналізу, алгоритм ГХЧН приблизно вдвічі-втричі швидше, ніж Apriori і працює краще з великою кількістю документів.

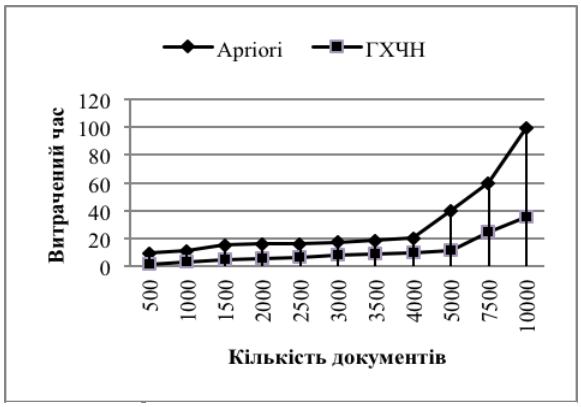

Рисунок 3 – Графік порівняння ефективності двох алгоритмів у співвідношенні витраченого часу до кількості документів за рівнем мінімальної підтримки 15 %

Оцінку тексту за алгоритмом кластеризації проведемо за допомогою методу k-середніх і FIHC методу. Перший – це метод кластерного аналізу, метою якого є розділення m спостережень (з простору R^n) на k кластерів, при цьому кожне спостереження належить до того кластеру, до центру (центроїду) якого воно найближче [11]. Другий – це ієрархічна кластеризація на основі частих наборів. Метод FIHC грунтується на ідеї частих наборів, запропонований R. Agrawal та R. Srikant [12]. Критерієм такої кластеризації є те, що в документі встановлюються деякі часті набори для кластера (теми), і різні кластери мають частку декількох частих наборів.

На рис. 4 представлено порівняння всіх трьох підходів до оцінки тексту за допомогою кластеризації на основі загального F-значення показника з різною кількістю кластерів. Механізм кластеризації вебдокументів на основі алгоритму хешування для отримання частих наборів перевершує всі інші підходи в плані точності.

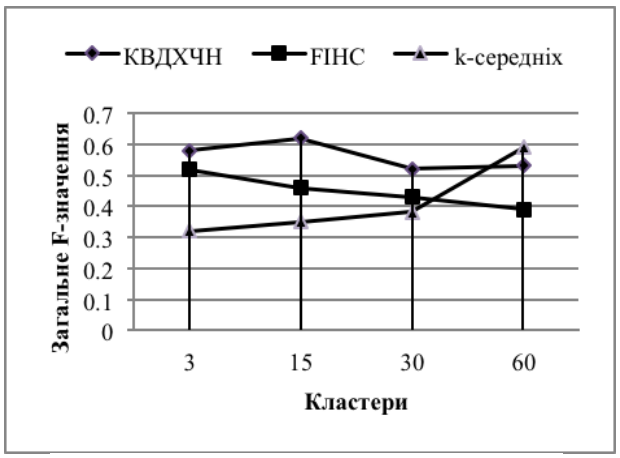

Рисунок 4 – Графік порівняння ефективності підходів кластеризації (КВДХЧН, FIHC та k-середніх) в співвідношенні із загальним F-значенням до кількості кластерів

Рис. 5 представляє порівняльний аналіз продуктивності КВДХЧН підходу з методами FIHC та k-середніх.

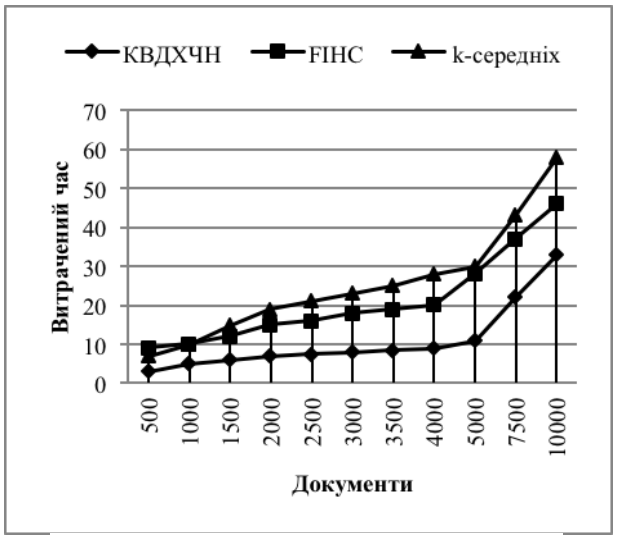

Рисунок 5 – Графік порівняння ефективності підходів кластеризації (КВДХЧН, FIHC та k-середніх) в співвідношенні витраченого часу до кількості документів за рівнем мінімальної підтримки 15 %

Рівень мінімальної підтримки становить до 15%, щоб гарантувати, що точність всіх вироблених підходів кластеризації порівняно однакова. По осі X задається кількість документів, прийнятих в аналізі, по Y-осі – час, необхідний, щоб знайти кластери. КВДХЧН підхід працює приблизно вдвічі швидше, ніж два методи FIHC та k-середніх. Робимо висновок, що метод КВДХЧН більш ефективний, ніж інші підходи.

ВИСНОВКИ. З отриманих результатів видно, що при використанні алгоритму ГХЧН витрачений час зменшується пропорційно збільшенню рівня підтримки під час порівняння з алгоритмом Apriori. У ГХЧН алгоритмі, в цілому, час виконання значний тільки при первинному створенні хеш-таблиці. При введенні нового рівня мінімальної підтримки ніякого збільшення витраченого часу не відбувається, час витрачається тільки на пошук в хеш-таблиці.

При використанні алгоритму Apriori витрачений час виконання зменшується при збільшенні рівня підтримки. В процесі виконання Apriori алгоритму, щоразу, заходячи на нову мінімальну підтримку, потрібно повторити процес знаходження наборів даних спочатку. Виходячи з проведеного дослідження, робимо висновок, що ГХЧН алгоритм є значно ефективнішим, ніж алгоритм Apriori у всіх випадках щодо великих за обсягом документів, оскільки складність знаходження частих наборів в алгоритму ГХЧН нижча, ніж у Apriori алгоритму.

З отриманих результатів можна зробити висновок, що алгоритм ГХЧН приблизно вдвічі-втричі швидше, ніж Apriori і працює краще з великою кількістю документів.

За результатми порівняння ефективності підходів кластеризації (КВДХЧН, FIHC та k-середніх) в співвідношенні витраченого часу до кількості документів за рівнем мінімальної підтримки 15 %, що представлені на рис. 5, робимо висновок, що метод КВДХЧН більш ефективний, ніж інші підходи.

Запропоновано механізм кластеризації веб-документів на основі алгоритму хешування для отримання частих наборів, що забезпечує значне зниження розмірності. Оригінальність КВДХЧН підходу полягає у впровадженні ефективного глобального алгоритму хешування для отримання частих наборів. Алгоритм ГХЧН передбачає створення хеш-таблиці в процесі сканування документів тільки один раз, а також можливість отримання частих наборів з різним мінімальним порогом підтримки без необхідності повторного сканування документів. Це є основним фактором прискорення процесу кластеризації.

Перспективи подальших досліджень у цьому напрямку будуть спрямовані на розробку нової технології кластеризації веб-документів на основі асоціативних правил.

ЛІТЕРАТУРА

1. Свириденко Я.С. Байєсівська класифікація текстових документів. Ієрархічна кластеризація на основі частих наборів // Восточно-Европейский журнал передовых технологий. – Вып. № 4 (45). – 2010. – С. 45–49.

2. Al Qady M. and Kandil A. (2014), Automatic clustering of construction project documents based on textual similarity, Automation in Construction Journal. – Vol. 42. – pp. 36–49.

3. Wang L., Tian L., Jia Y. and Han W. A Hybrid algorithm for web document clustering based on frequent term sets and k-means // Lecture Notes in Computer Science, Springer Berlin, 2010. – Vol. 4537. – pp. 198–203.

4. Розломій І. О. Організація і оцінка ефективності системи захищеного документообігу // Вісник Кременчуцького національного університету імені Михайла Остроградського. – Кременчук: КрНУ, 2015. – Вип. 6/2015(1). – С. 119–124.

5. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терми-

нов, и его применение к анализу тематической структуры новостного потока, а также ее динамики / М.В. Киселев, В.С. Пивоваров, М.М. Шмулевич // Интернет-математика 2005. Автоматическая обработка веб-данных. – С. 412–435.

6. Fung B., Wang K. and Ester M. Hierarchical document clustering using frequent itemsets // International Conference on Data Mining, 2003. – Vol. 30. – pp. 59–70.

7. Anaya H., Pons A. and Berlanga R.A Document clustering algorithm for discovering and describing topics // Pattern Recognition Letters, 2010. – Vol. 31. – pp. 502–510.

8. Портер М.Е. Конкуренция. Пер. с англ. – М.: Изд. Дом «Вильямс». – 2000. – 608 с.

9. Інформаційне агентство УНИАН [Електронний ресурс]. – Режим доступу: http://www.unian.net/ (25.07.16).

10. Баженов Д. Оценка классификатора (точность, полнота, F-мера). [Електронний ресурс]. – Режим доступу: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html (25.07.16).

11. Технологии анализа данных. / А.А. Барсегян, М.С. Куприянов, В.В Степаненко, И.И. Холод. – Санкт-Петербург: БХВ-Петербург, 2007. – 512 с.

12. Baghel R. and Dhir Dr. R. A Frequent concept based document clustering algorithm, International Journal of Computer Applications, 2010. – Vol. 4. – pp. 875–887.

CLUSTERING WEB DOCUMENTS BASED ON HASHING ALGORITHM

A. Osidach

National University «Lviv Polytechnic»

vul. Stepana Bandery, 12, Lviv, 79000, Ukraine. E-mail: osidach.publications@gmail.com

Purpose. To develop a block scheme of the Web documents clustering process tied on the frequent itemset-based hierarchical algorithm (FIHC). **Methodology.** An evaluation of the algorithm effectiveness has been presented in comparing with the result of global hashing algorithm for frequent sets of algorithm Apriori. Mentioned approach has been based on the clustering of web-documents. **Results.** As a result of the research a mechanism of the web documents clustering has been proposed using the frequent itemset-based hierarchical algorithm (FIHC) regardless of documents size. The main feature of the FIHC approach is the implementation of an effective global hashing algorithm for the frequent itemset. **Originality.** An originality of the approach is the implementation of an effective global hashing algorithm for frequent sets. **Practical value.** According to the achieved results, FIHC algorithm involves abilities to create the hash table in the process of scanning documents only once and to receive frequent sets with the different minimum threshold of support without having to re-scan the documents. This is the main factor in speeding up the process of clustering. References 12, figures 5.

Key words:clustering, web document, hashing, algorithm, frequent sets, indexing, association rules, mining.

REFERENCES

1. Sviridenko, Y. S. (2010), "Bayes classification of the text documents. Hierarchical clustering on the basis of frequent itemsets", *Eastern-European Journal of Enterprise Technologies*, vol. 3, no. 4 (45), pp. 45–49.

2. Al Qady, M., Kandil, A. (2014), "Automatic clustering of construction project documents based on textual similarity", *Automation in Construction Journal*, vol. 42, pp. 36–49.

3. Wang, L., Tian, L., Jia, Y., Han, W.A (2010), "Hybrid algorithm for web document clustering based on frequent term sets and k-means", *Lecture Notes in Computer Science*, vol. 4537, pp. 198–203.

4. Rozlomii, I. O. (2015), "Deploying and measuring the effectivness of a secure document management system", *Transactions of Kremenchuk Mykhailo Ostrohradskyi National University*, vol. 6, no. 95, pp. 119–124.

5. Kyselov, M., Pyvovarov, V., Shmulevych, M. (2005), "Method of text clustering which includes the co-occurrence of key terms, usage of this method for the analysis of the news flow thematic structure and also its dynamics", *Internet-matematyka 2005. Avtomaticheskaja obrabotka veb-danych*, pp. 412–435.

6. Fung, B., Wang, K., Ester, M. (2003), "Hierarchical document clustering using frequent itemsets", *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM)*,San Francisco, CA, May 1-3, 2003, pp. 59–70.

7. Anaya, H., Pons, A., Berlanga, R.A. (2010), "Document clustering algorithm for discovering and describing topics", *Pattern Recognition Letters*, vol. 31, pp. 502–510.

8. Porter, M. (2000), *Konkurencija* [Competition], Williams, St. Petersburg, Russia.

9. Information agency of Ukraine (2016), available at http://www.unian.net/ (accessed at July 25, 2016).

10. Bezhenov, D. (2012), "Estimation of the classifier (accuracy, completeness , F- measure)", Bazhenov blog, available at http://bazhenov.me/blog/ 2012/07/ 21/classification-performance-evaluation.html (accessed at July 25, 2016).

11. Barsegyan, A. (2007), *Technologii analiza danych* [Data Mining, Visual Mining, Text Mining], Peterburg, St. Peterburg, Russia.

12. Baghel, R., Dhir, Dr. R. A. (2010), "Frequent concept based document clustering algorithm, International", *Journal of Computer Applications*, vol. 4, pp. 875–887.

Стаття надійшла 23.08.2016.