Секція: Фізико-математичні науки

ОСІДАЧ АНДРІЙ ОЛЕГОВИЧ
аспірант кафедри загальної екології та
екоінформаційних систем
Національний університет "Львівська Політехніка"
м. Львів, Україна

ЕТАПИ КЛАСТЕРИЗАЦІЇ ВЕБ-ДОКУМЕНТІВ НА ОСНОВІ АЛГОРИТМУ ХЕШУВАННЯ

Структурна схема процесу кластеризації веб-документів на основі алгоритму хешування для отримання частих наборів (КВДХЧН) запропонована на рис.1. Основною характеристикою цього підходу є те, що в ньому реалізовано новий алгоритм інтелектуального аналізу даних з метою подолання недоліків алгоритму Apriori (алгоритм пошуку асоціативних правил).

Рис. 1. Структурна схема процесу 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.

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 бути в змозі перетворити різні синтаксичні форми слова в нормалізованому вигляді, знизити кількість індексних термінів, з метою економії пам'яті, та, деякою мірою, може збільшити продуктивність алгоритмів кластеризації. Портер парадигматичного модуля [1] – метод, який широко застосовується для стовбурових документів. 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;

$tf_{i,j}$ – 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:

$$idf_i = log\left({}^{N}/_{Nt_j}\right), \qquad (2)$$

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

Як тільки схему зважування було обрано, автоматизоване індексування може бути виконано, просто вибравши кращі K-слова, що задовольняють заданій вазі обмеження для кожного документа. Основною перевагою автоматизованої процедури індексування є те, що вона скорочує витрати на індексацію [2].

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 складових і містити цікаві 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 масштабувати документи, 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 [3].

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

а) масив – це масив U з розміру R, де під кожною ячейкою мається на увазі 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, в яких документ кластеризації використовується, наприклад, додаток рекомендацій, який використовує результати кластеризації для рекомендацій користувачам.

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.

Література:

1. По определению Майкла Портера, кластеры [Электронный ресурс]. – Режим доступа: http://davaiknam.ru/text/po-opredeleniyu-majkla-portera-klasteri – 05.02.2016 р. – Загол. с экрана.

2. Fung B., Wang K. and Ester M. Hierarchical document clustering using frequent itemsets // International Conference on Data Mining, 2003. – vol. 30. – P. 59-70.

3. Baghel R. and Dhir Dr. R. A Frequent concept based document clustering algorithm, International Journal of Computer Applications, 2010. – vol. 4. – P. 875-887.