УДК 004.62

Ст. викладач Мальчиков В. В., студент Пирогов Є. В.

**Національний технічний університет України
«Київський політехнічний інститут»**

МЕТОД ПОШУКУ ДОКУМЕНТІВ НА ОСНОВІ КОМБІНАЦІЇ КЛАСТЕРНИХ АЛГОРИТМІВ

Abstract

V. V. Malchykov, senior lecturer, E. V. Pirogov, student
Document search methods based on combination of cluster algorithms

This article is dedicated to a review and analysis of current approaches to data retrival, based on clustering algorithms. The article includes a comparison of current known algorithms, as well as conclusions regarding them. Authors proposed a new algorithm as a combination of already known ones, with the use of several modern techniques to storing and accessing the data.

Вступ

Кластеризація документів вважається перспективною технологією [1] для поліпшення пошуку документів та отримання даних. Вважається, що подібні (за деякою ознакою) документи можуть бути отримані одним і тим самим запитом. Тому автоматичне об'єднання документів у групи може ефективно розширити та якісно поліпшити результат пошуку [2]. У випадку розбиття документів на окремі кластери пошуковий запит знаходить найбільш підходящий кластер і повертає множину документів, що той містить. У момент видачі документи всередині кластера можуть бути відсортовані за релевантністю. У разі ієрархічного розбиття запит на кластері виконується послідовно "зверху вниз" доти, поки не буде виконуватися умова завершення (наприклад, до тих пір, поки значення коефіцієнта релевантності не стане нижчим за деяке порогове значення). У цьому випадку користувачеві буде повернено залишкове дерево з кластерів і, відповідно, документи в ньому.

В основі цих стратегій лежить пошук “найближчих сусідів”, де "близькість" сусідів визначається деякою мірою кореляції двох документів, завдяки чому і формується кластер. Як і будь-який інший метод, даний підхід має певні недоліки і обмеження. Зокрема, склад і кількість кластерів залежить від обраних критеріїв розбиття, визначення яких може потребувати додаткового аналізу.

Постановка задачі

З метою поліпшення якості результатів та підвищення швидкості пошуку пропонується метод кластерного аналізу документів на основі комбінації і модифікації низки існуючих алгоритмів.

Огляд існуючих рішень

На даний момент існує багато методів [3], які виконують кластеризацію документів. Деякі методи можуть використовувати кілька альтернативних алгоритмів. На основі аналізу методів були виділені їх основні особливості. Найчастіше використовуються наступні методи:

1. Custom Search Folders — метод дозволяє звузити результати пошуку шляхом розподілу їх папок (folders);
2. LSA/LSI — Latent Semantic Analysis/Indexing. Шляхом факторного аналізу множини документів виявляються латентні (приховані) фактори, які надалі є основою для створення кластерів документів;
3. STC — Suffix Tree Clustering. Кластери утворюються у вузлах дерева спеціального виду — суфікс-дерева, яке будується з слів і фраз вхідних документів;
4. Single Link, Complete Link, Group Average — ці методи розбивають множину документів на кластери, розташовані в деревоподібній структурі (дендрограмі), що одержується за допомогою ієрархічної агломеративної кластеризації;
5. K-means. Відноситься до не-ієрархічних алгоритмів. Кластери представлені у вигляді центроїдів, що є центром "маси" всіх документів, які входять в кластер;
6. CI — Concept Indexing. Розбиває множину документів методом рекурсивної бісекції, тобто розділяючи множину документів на дві частини на кожному кроці рекурсії. Метод може використовувати інформацію, отриману на етапі навчання;
7. SOM — Self-Organizing Maps. Виробляє класифікацію документів самостійно;
8. Buckshot — швидкий, проте недостатньо точний алгоритм розбиття на кластери.
9. Fractional — точний, проте повільний алгоритм.

Практично всі згадані підходи мають одні й ті самі недоліки. Вони потребують достатньо складних структур даних для зберігання документів; показують невисокі результати швидкості пошуку на величезних колекціях документів; не використовують сучасні можливості щодо паралельного обчислення та доступу до даних. Тому пропонується

метод, який представляє собою комбінацію методів кластеризації Buckshot і Fractionation [4].

Опис запропонованого методу

На початку система розбиває документи на невелику кількість груп (перша фаза). Ґрунтуючись на коротких описах вмісту груп, користувач вибирає одну або кілька груп для подальшого розгляду. Документи об'єднуються (друга фаза) і розглядаються як одна група, і процес повторюється вже над нею. Це схоже на послідовність штучних запитів всередині основного.

На фазі розбиття метод може використовувати два алгоритми: Buckshot і Fractionation [5]. Алгоритм Buckshot менш точний, проте здатний до швидкої рекластеризації при виконанні ітерацій. Fractionation ж є більш точним, але й більш повільним алгоритмом, і використовується для попереднього розбиття на групи множини документів та виконується в режимі off-line. Обидва алгоритми (Buckshot і Fractionation) належать до алгоритмів висхідної (bottom-up) кластеризації.

Ідея алгоритму Buckshot полягає в тому, що з усієї кількості документів обирається випадкова вибірка розміром $\sqrt{x * k}$, де *k* — необхідна кількість утворених груп, n — кількість документів; далі над цією вибіркою проводиться процедура знаходження центрів кластерів. Процедура полягає у послідовному "об’єднанні" документів, що знаходяться в найбільшій близькості один від одного (використовується значення близькості між документами). Процедура виконується доти, поки потрібну – заздалегідь задану кількість центрів – не буде знайдено. Швидкість роботи алгоритму — *O(kn)*.

Fractionation – більш складний алгоритм. Спочатку всі документи розбиваються на групи, кількість яких більша за кількість передбачуваних кластерів. Потім групи об'єднуються, щоб їх кількість дорівнювала заданій кількості кластерів. На *j*-му кроці об’єднана кількість груп дорівнює $r^j n$, де *r* – якийсь чинник редукції, менший 1. Швидкість роботи алгоритму *O(mn),* де *m* — число початкових груп, на які поділяється множина документів

Після закінчення роботи алгоритму Fractionation у множині проводиться пошук центрів та присвоєння документів до отриманих центрів. Присвоєння відбувається за принципом Assign-To-Nearest — документ присвоюється до найближчого центру. На цьому етапі йде більше класифікація, ніж кластеризація. Швидкість класифікації дуже велика порівняно з кластеризацією, і пропорційна кількості кластерів.

Для присвоєння назви отриманим кластера обираються слова, що найчастіше зустрічаються в наборі документів кластера, і з них складається назва кластера (наприклад, простим перелічуванням слів).

Висновки

В умовах наявності величезної кількості об'єктів природним бажанням користувача є бачити досить короткий список рубрик-категорій, під які потрапляють всі отримані документи. Користуючись цими категоріями, користувач істотно звужує рамки пошуку. Підхід, запропонований в даній роботі, грунтується саме на динамічному формуванні категорій (кластерів).

Серед переваг описаної комбінації методів слід відзначити: розбиття на кластери з допомогою алгоритму Buckshot забезпечує високу швидкість, а алгоритм Fractionation гарантує високу точність визначення центроїдів кластерів; гарна наочність представлення даних; алгоритм не потребує "навчання"; використовується матриця близькості документів.

На даний момент проводиться тестування запропонованого методу.

Література

1. *Z. Dong*, Towards Web Information Clustering, doctoral dissertation, Southeast Univ., Nanjing, China, 2002. – p. 1.

2. *Grabmeier J., Rudolph A.*, "Techniques of cluster algorithms in data mining", Data Mining and Knowledge Discovery. – October 2002. – Vol. 6, № 4. – pp. 303-360.

3. *Runkler T.A., Bezdek J.C.*, "Web mining with relational clustering", International Journal of Approximate Reasoning. – February 2003. – Vol. 32, №. 2-3. – pp. 217-236.

4. *D. R. Cutting, J. Pedersen, D. R. Karger and J. W. Tukey*. "Scatter/gather: A cluster-based approach to browsing large document collections", in SIGIR Forum, Copenhagen, Denmark, 1992. – pp. 318–329

5. *Tantrum, J., Murua, A. and Stuetzle*, W. 2002 Hierarchical model-based clustering of large datasets through fractionation and refractionation. In Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. New York: ACM Press. – p. 2.