УДК 681.301

К.т.н, доцент Петрашенко А.В., магістрант Максименко І.Г.

Національний технічний університету України
«Київський політехнічний інститут»

АНАЛІЗ МЕТОДІВ ТА ОПТИМІЗАЦІЯ ЗАДАЧІ КЛАСТЕРИЗАЦІЇ ДОКУМЕНТІВ У ІНТЕРНЕТ

Abstract

Andriy V. Petrashenko, assoc. prof., PhD; Igor G. Maksimenko, student
Document clustering methods analysis and optimization at the Internet

This paper presents a comparison of document clustering methods at the Internet, choosing one of the most optimized and its modification. The modification will provide an opportunity to reduce consumption of resources in carrying out the algorithm of clustering method.

Вступ

Кластеризація грає важливу роль в отриманні релевантної інформації, навігації, реферування та організації текстових документів, доступних в Інтернеті, електронних бібліотеках і корпоративних мережах. Тому, на сьогоднішній день, актуальною є проблема пошуку оптимальних шляхів вирішення задачі кластеризації документів.

Нерідко трапляється, коли на введений користувачем запит повертається величезна кількість посилань на текстові та інші види матеріалів. Як правило, результати запиту є лінійними списками з простим перерахуванням об'єктів, впорядкованих за релевантністю до запиту.

В умовах, коли кількість об'єктів велика, природним бажанням користувача є отримання короткого списку рубрик знайдених документів. Розбиття всього обсягу документів на рубрики підпадає під задачу кластеризації даних.

Постановка задачі

Метою роботи є дослідження та аналіз методів кластеризації документів в глобальній мережі Інтернет з метою вибору оптимізованого за критеріями співвідношення швидкості і точності, можливості роботи в «інкрементному» режимі, простоти алгоритму та його реалізації, наявності «перетину» кластерів.

Термінологія

Кластеризація в Інтернет - задача, в якій потрібно розбити задану вибірку об'єктів (web-документів) на підмножини, які називають кластерами, так, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися.

Аналіз існуючих методів кластеризації документів

Оцінка методів проводитиметься за критеріями: відношення швидкості і точності; точність системи; можливість роботи алгоритму в «інкрементному» режимі - здійснювати рубрикацію документів, одночасно з надходженням їх з джерела, не виконуючи кожного разу повний цикл обчислень; наявність «перетинів» - можливість попадання одного документа в різні рубрики за суміжними темами; якомога менша кількість попередньої інформації, необхідної для процесу рубрикації.

У методі кластеризації Suffix Tree Clustering [1] побудова дерева здійснюється поетапно. Спочатку отримані з мережі документи піддаються попередній обробці – очищенню від знаків пунктуації, приведення слів до початкової форми тощо. Потім для набору документів будується дерево, але одиницею, що знаходиться на ребрах дерева є слово чи словосполучення. Кожній вершині дерева відповідає фраза. Її можна отримати, об'єднавши всі слова або словосполучення, що знаходяться на ребрах на шляху від кореня дерева до даної вершини дерева. У тих вершинах дерева, які мають нащадків, є посилання на документи, в яких зустрічається фраза, відповідна вершині. Множини документів, на які вказують ці посилання, утворюють базові кластери. Після цього проводиться комбінування базових кластерів та отримання набору кластерів.

Переваги методу: висока швидкість роботи; наочність представлення результатів (загальні фрагменти текстів і фраз виступають в якості назви кластерів); алгоритм не потребує навчання і встановлення порогу спрацьовування; алгоритм інкрементний і допускає перетин областей видимості кластерів.

Недоліки методу: необхідність повторної обробки текстів документів; не виявляється прихована семантика серед документів, яка може бути присутня не тільки на текстовому рівні; синонімія та омонімія.

Особливістю методів Single Link, Complete Link, Group Average, є те, що вони розділяють документи на кластери шляхом розбиття їх на ієрархічні групи, множина кластерів має ієрархічну структуру. Принцип роботи ієрархічних агломеративних процедур полягає в послідовному

об'єднанні груп елементів, спочатку найближчих, а потім все більш віддалених один від одного.

Переваги методів: алгоритми не потребують навчання; використання матриці близькості між документами; алгоритми інкрементні [2].

Недоліки методів: необхідно встановлення порогу – максимальної кількості документів в кластері; для отримання хороших результатів кластеризації значення близькості між парами документів повинні приходити в певному порядку, тобто робота алгоритму не детермінована; кластери не перетинаються.

В основі методу k-середніх (k-means) лежить ітеративний процес стабілізації центроїдів кластерів. Основною характеристикою кластера є його центроїд [5] і вся робота алгоритму спрямована на стабілізацію або повне припинення зміни центроїда кластера. Спершу вибираються початкові центроїди для множини документів, потім усі документи розподіляються за значенням близькості між цими центроїдами, таким чином формуючи кластери. Після цього відбувається перерахунок центроїдів кластерів. Це триває до того часу, поки кількість документів у кластері перестане змінюватися.

Переваги методу: доволі висока швидкість роботи; використовує значення матриці близькості; метод не потребує навчання і за необхідності може накопичувати відомості для подальшого збільшення точності роботи;

Недоліки методу: потрібно задавати кількість кластерів, як мінімум на початкових етапах – до використання апріорної інформації; у тому випадку, коли центроїд кластерів вибирається випадковим чином, результати, отримані над однією і тією ж вибіркою документів, будуть відрізнятися; алгоритм не інкрементний; кластери не перетинаються.

У методі Concept Indexing [3, 4] пошук кластерів може відбуватися або шляхом безпосереднього виділення k документів, що виконуватимуть роль центроїдів, або шляхом використання методу рекурсивної бісекції. На початку процесу безпосереднього розбиття довільно вибираються k документів, які є центрами для кластерів, потім до цих центрів відносять документи, які мають найвище значення коефіцієнта близькості, виконується перерахунок центроїдів кластерів, що виникли. Метод рекурсивної бісекції має за мету розбиття всього обсягу документів на 2 частини, потім ці частини теж піддають розбиттю, за необхідністю. Процес триває до отримання k кластерів. Алгоритм рекурсивної бісекції використовує значення «сумарної неподібності» (aggregate dissimilarity) між документами в кластері, для того щоб вирішити який кластер розбивати далі.

Переваги методу: алгоритм рекурсивної бісекції має високу швидкість роботи; у разі, коли не використовується рекурсивна бісекція

для знаходження кластерів, алгоритм інкрементний; простота використання; зрозумілість і прозорість алгоритму.

Недоліки методу: потрібно задання кількості кластерів, на які буде розбиватися множина документів.

Результати порівняння

В результаті порівняння методів кластеризації був обраний метод Concept Indexing, а саме його варіант, що використовує метод рекурсивної бісекції. При звичайному розбитті на фіксовану кількість кластерів важче регулювати точність та витрату ресурсів при виконанні алгоритму. Метод поєднує в собі високу швидкість роботи разом з ясністю та прозорістю алгоритму.

Модифікація полягає у введенні додаткового коефіцієнту t, який буде виражати мінімальну допустиму подібність, між крайніми протилежними елементами кластеру. Якщо вона буде менша за t, то процес подальшого рекурсивного розбиття такого кластеру припиняється.

У загальному випадку кінцевий алгоритм кластеризації буде мати вигляд (рис.1):

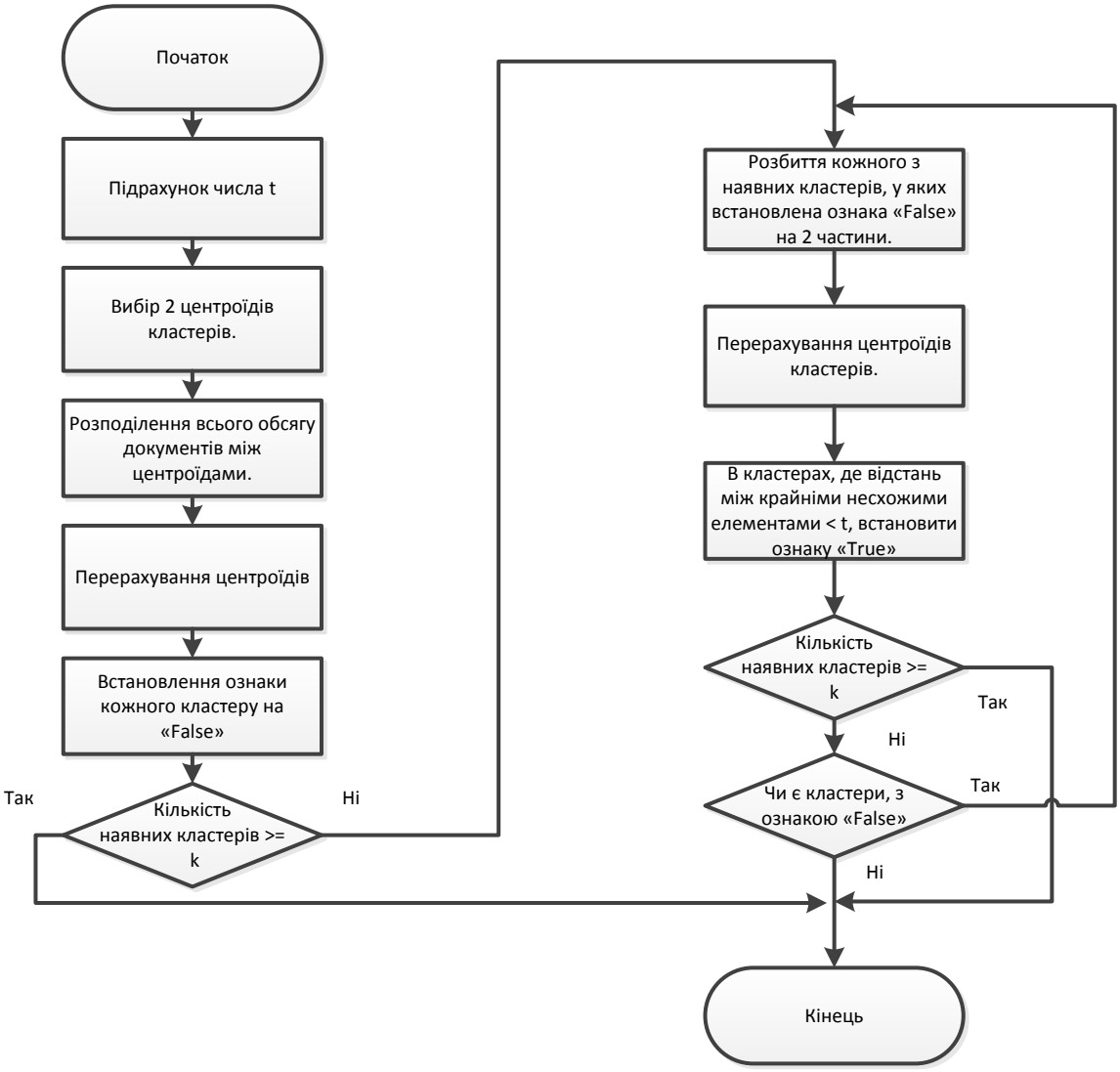

Рис.1. Модифікований алгоритм кластеризації.

Висновки

В ході проведеної роботи проведено дослідження та аналіз методів кластеризації документів в глобальній мережі Інтернет і вибір оптимізованого за параметрами швидкості роботи, прозорості алгоритму та простоти його реалізації на автоматизованих системах.

Оцінка методів була проведена за критеріями співвідношення швидкості і точності, можливості роботи в «інкрементному» режимі, простоти алгоритму та його реалізації, наявності «перетинання» кластерів.

В кінці дослідження обраний метод був модифікований. Він може бути успішно реалізований для подальшого вирішення задачі кластеризації документів в Інтернет. Запропонована модифікація дозволить збільшити співвідношення точності до часу виконання методу.

Література

1. *Oren Eli Zamir*. A Phrase-Based Method for Grouping Search Engine Results. University of Washington, Department of Science & Engineering. University of Washington, 1999. – p.p. 65-117.
2. *R. Sibson*. SLINK: An optimally efficient algorithm for the single-link cluster method. King's College Research Center, King's College, Cambridge, and Cambridge University Statistical Laboratory. Cambridge, 1972 – p.p. 1-34.
3. *Eui-Hong (Sam) Han and George Kapyris.* Centroid-Based Document Classification: Analysis & Experimental Results. University of Minnesota, Department of Computer Science / Army HPC Research Center. 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD*)*, 2000 – p.p. 1-11.
4. *Eui-Hong (Sam) Han and George Kapyris*. Concept Indexing. A Fast Dimensionality Reduction Algorithm with Application to Document Retrieval & Categorization. University of Minnesota, Department of Computer Science / Army HPC Research Center. University of Minnesota, 2000 – p.p. 1-20.
5. *A.K. Jain, M.N. Murty and P.J. Flynn*. Data Clustering: A Review. Michigan State University, Indian Institute of Scienc, The Ohio State University. Michigan State University, 1999. – p.p. 4-47.
6. http://www.dialog-21.ru/Archive/2001/volume2/2_26.htm