УДК 004.91

Я.О. ЮСИН, Т.М. ЗАБОЛОТНЯ
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

МОДИФІКОВАНИЙ МЕТОД ОСТРІВНОЇ КЛАСТЕРИЗАЦІЇ ТЕКСТОВИХ ДАНИХ

Запропоновано модифікований метод острівної кластеризації текстів, що базується на розбитті графу кореляції термів на групи. Проведено тестування точності та швидкості виконання кластеризації текстових документів за допомогою оригінального методу острівної кластеризації та модифікованого методу.

Ключові слова: кластеризація, острівна кластеризація, розбиття графу, k-medoids

Y.O. YUSYN, T.M. ZABOLOTNYA
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»

MODIFIED METHOD OF ISLAND TEXT CLUSTERING

The implementation of clustering of natural-language text data is an important scientific and practical task. Thus, the improvement of existing and the development of new methods of clustering is an urgent task. In this paper, the method of island clustering is considered, in which the step of clustering of the received approximation of the terms co-occurrence graph is highlighted for the further improvement. On the basis of consideration of the original procedure of clustering of the terms co-occurrence graph it disadvantages were identified. As a solution to the described shortcomings of the original procedure of clustering the term correlation graph, using of the graph partitioning procedure in the clustering of the terms co-occurrence graph instead of the described original procedure is proposed. The use of the k-medoids method as a procedure of the terms correlation graph partitioning into groups is proposed. The three most common implementations of the k-medoids method are described: PAM, s-heuristics and clara. The modified method of island text clustering which is based on correlation graph partitioning is proposed. Testing of accuracy and speed of text clustering using the original island text clustering method and the proposed modified method has been carried out. This testing was conducted on two corpuses of news texts. Tests have shown that the results obtained by the proposed modified method of island text clustering is better by 3-8% by the accuracy criterion (depends on the chosen implementation of k-medoids method). At the same time, clustering is slowed down by 4-15%. Thus, the proposed modified method of island text clustering is appropriate to be used in practice in tasks where the accuracy of the results is more important than the speed of clustering execution.

Keywords: clustering, island clustering, graph partitioning, k-medoids.

Вступ

Починаючи з 1950-х років, людство стикнулося з явищем, яке згодом отримало назву «інформаційного вибуху» – невпинним зростанням кількості інформації, що генерується [1]. В першу чергу, це явище стосується текстових документів, що мають місце в науці, бізнесі, навчанні та інших сферах людської діяльності. Наприклад, тільки в 2012 році було згенеровано 2.8 зетабайти текстових даних, і відповідно до результатів проведених досліджень ця цифра на далі тільки зростатиме, подвоюючись кожні два роки [2]. При цьому потенційно корисними є лише 23% згенерованої інформації, а структурованими – лише 5% [2]. Наслідком такого розростання множини текстових документів стала необхідність в розробленні методів автоматичної попередньої систематизації масивів даних перед їх подальшим обробленням та/чи аналізом.

В таких умовах формулювання нових та вдосконалення існуючих методів автоматичної (unsupervised) кластеризації текстових документів, тобто розбиття корпусів документів на наперед не задані підмножини (кластери), стає актуальною задачею [3].

Тут слід відмітити, що оскільки зазвичай результати кластеризації текстів інтерпретуються безпосередньо людиною, остання повинна розуміти зміст знайденого кластеру і чому певні тексти були віднесені саме до нього. Це є важливим фактором, що відрізняє задачу кластеризації текстів від задач кластеризації інших видів даних. З цієї причини в якості критеріїв ефективності методів кластеризації текстів зазвичай використовується точність та швидкість виконання кластеризації тестових попередньо розмічених корпусів.

Постановка завдання

Одним з існуючих ефективних методів кластеризації текстів, що забезпечує зрозумілість процесу отримання кластерів для людини, є метод острівної кластеризації [4].

В основі цього методу лежить виконання двоетапної процедури: на першому етапі передбачено проведення кластеризації термів, з яких складаються документи; на другому етапі – побудова кластерів документів, виходячи з отриманих на першому етапі кластерів термів. Наведемо нижче більш детальний перелік кроків даного методу [4]:

Етап І. Кластеризація термів

1. Попереднє оброблення текстів з вхідної колекції документів: видалення стоп-слів, лематизація тощо.
2. Виділення з текстів множини термів, з яких вони складаються.
3. За необхідності – фільтрація отриманої множини термів (наприклад, в ситуаціях, коли відомі початкові центроїди кластерів або отримана множина є занадто великою).
4. Побудова графу кореляції термів між собою.
5. Попереднє оброблення графу і отримання його наближення.

6. Кластеризація отриманого наближення графу.

Етап 2. Побудова кластерів документів

1. Розбиття документів на кластери на основі отриманих кластерів термів.

В рамках даної роботи автори пропонують зосередити увагу на кластеризації отриманого наближення графу кореляції термів між собою (етап I, крок 6) (далі будемо його називати просто «граф кореляції термів»), оскільки удосконалення цього процесу може сприяти підвищенню точності кластеризації.

Таким чином, **метою** даної роботи є підвищення ефективності кластеризації текстових даних методом острівної кластеризації за критерієм точності шляхом розроблення модифікованого методу острівної кластеризації текстів.

Відповідно до вказаної мети в роботі поставлені і розв'язані такі **задачі**:

- визначення переваг та недоліків оригінальної процедури кластеризації графу кореляції термів;
- розроблення модифікованого методу острівної кластеризації;
- аналіз точності кластеризації текстових документів згідно з модифікованим методом;
- аналіз ефективності розробленого модифікованого методу за критерієм швидкості виконання кластеризації.

Оригінальна процедура кластеризації графу кореляції термів

В роботі [4], що описує метод острівної кластеризації, для кластеризації графу кореляції термів пропонується процедура, що отримує на вхід ребра графу, відсортовані за зменшенням ступеня кореляції термів. В якості позначення для ребра використовується позначення $< i, j, \tilde{p}_{ij} >$, де відповідно i та j – це терми (вершини графу), а $\tilde{p}_{ij}$ – ступінь їх кореляції (вага ребра). В рамках процедури кластеризації також використовуються такі терміни та позначення:

- множина незафіксованих кластерів термів G;
- множина зафіксованих кластерів термів F;
- множина зафіксованих термів PRO, що включає терми, які відносяться до повністю сформованих кластерів термів;
- параметр T_S, що визначає мінімальні розміри кластерів, що отримуються;
- $Pop(C)$ – множина документів, що буде відповідати цьому кластеру термів.

Кожний кластер термів в рамках даної процедури визначається множиною його термів (вершин) та множиною зв'язків між ними (ребер).

Крім цього, в ході виконання даної процедури часто використовується операція фіксації кластеру термів, яка включає в себе перенесення кластеру термів з множини G до F та додавання всіх термів кластеру до множини PRO.

Кожна ітерація процедури кластеризації графу кореляції термів полягає в отриманні чергового ребра $< i, j, \tilde{p}_{ij} >$ та його аналізі, поки не будуть проаналізовані всі ребра графу. Аналіз ребра відбувається наступним чином [4]:

1. Якщо $i \in PRO$ та $j \in PRO$ – перейти до наступного ребра.

2. Якщо $i \in PRO$ – якщо існує кластер, що містить терм j, то він фіксується; інакше – $PRO \leftarrow PRO \cup \{j\}$.

3. Якщо $j \in PRO$ – якщо існує кластер, що містить терм i, то він фіксується; інакше – $PRO \leftarrow PRO \cup \{i\}$.

4. Якщо і терм i, і терм j не належать до жодного кластеру, то до множини G додається кластер, що містить ці два терми і зв'язок між ними.

5. Якщо терми i та j належать до одного кластеру, то до цього кластеру додається зв'язок між ними.

6. Якщо терми i та j належать до різних кластерів C та D – якщо $|Pop(C)| > T_S$ або $|Pop(D)| > T_S$, то обидва кластери термів фіксуються; інакше кластери C та D видаляються з множини G, а замість них до неї додається кластер, що є об'єднанням цих кластерів і зв'язку між i та j.

7. Якщо терм i належить до деякого кластеру, а терм j не належить жодному кластеру – до кластеру, що містить терм i додається терм j та зв'язок між ними.

8. Якщо терм j належить до деякого кластеру, а терм i не належить жодному кластеру – до кластеру, що містить терм j додається терм i та зв'язок між ними.

Коли аналіз всіх ребер закінчено, останнім кроком процедури кластеризації графу кореляції термів є фіксація всіх кластерів термів, що залишилися в множині G. В якості результату кластеризації приймається множина F.

Описана процедура кластеризації графу кореляції термів має такі переваги:

- лінійна залежність обчислювальної складності від кількості ребер графу;
- всі зв'язки між термами одного кластеру є сильнішими будь-яких інших зв'язків термів цього кластеру з іншими термами.

Проте в багатьох практичних застосуваннях кластеризації текстових документів дана процедура має свої недоліки. Такими недоліками є:

- неможливість ручного встановлення очікуваної кількості кластерів – часто перед виконанням кластеризації вже є припущення про їх кількість;
- дана процедура виконує неексклюзивну кластеризацію [4] (один терм може потрапити більше, ніж до одного кластеру), коли часто потрібне виконання ексклюзивної кластеризації (один терм – один кластер).

Вирішити описані недоліки може використання іншої процедури для кластеризації графу кореляції термів, яка передбачає ручне встановлення кількості кластерів та ексклюзивну кластеризацію, тобто виконує розбиття графу на групи.

Модифікований метод, що базується на розбитті графу кореляції термів на групи

Таким чином, основною ідеєю запропонованого модифікованого методу острівної кластеризації текстів є використання процедури розбиття графу на групи (graph partitioning) під час кластеризації графу сумісної зустрічальності термів замість описаної оригінальної процедури кластеризації. Інші кроки методу острівної кластеризації текстів в модифікованому методі залишаються без змін.

В рамках даної роботи в якості процедури розбиття графу кореляції термів на групи пропонується використання методу *k-medoids* [5]. Даний метод є модифікацією класичного методу кластеризації даних k-means, яка спеціально розроблена для кластеризації графу, оскільки оригінальний метод *k* найближчих не може бути застосований з цією метою. Це пов'язано з тим, що в оригінальному методі в якості центроїдів кластерів може виступати будь-яка випадкова точка простору вимірювань. У випадку графу встановлення відстані до такої точки є неможливим. Саме тому в методі *k-medoids* накладено обмеження на центроїди, в якості яких можуть виступати лише точки графу.

Найбільш поширеним алгоритмом, що реалізує метод *k-medoids* є PAM (Partitioning Around Medoids) алгоритм [4]. PAM використовує жадібний пошук, який може не знайти оптимальне рішення, проте він є швидшим, ніж повноцінний вичерпний пошук.

Алгоритм PAM складається з таких кроків:

1. На вхід алгоритм отримує граф $G = (V, E, w)$, а також число k – задану кількість кластерів.
2. Ініціалізація: обираємо випадково k вершин в якості початкових медоїдів.
3. Для кожної точки знаходимо найближчий медоїд, формуючи початкове розбиття на кластери.
4. Знаходимо *minCost* , як значення функції втрат від початкової конфігурації.
5. Поки медоїди не стабілізуються, повторюємо наступні кроки алгоритму.
6. Для кожного медоїду m повторюємо кроки 7-12.
7. Для кожної вершини $v \neq m$, що знаходиться всередині кластеру з центром в m повторюємо кроки 8-12.
8. Переміщуємо центр кластеру з m в v.
9. Перерозподіляємо всі вершини між новими медоїдами.
10. Знаходимо *cost* $cost$ – значення функції втрат від поточної конфігурації.
11. Якщо *cost* < *minCost* $cost < minCost$, запам'ятовуємо медоїди і прирівнюємо $minCost = cost$.
12. Повертаємо медоїд на місце (в m).
13. Робимо найкращу знайдену заміну зі всіх, що розглядались.

Складність однієї ітерації алгоритму в найгіршому випадку складає $O((n-k)^2 k)$ $O(k*(n-k)^2)$.

Даний алгоритм може бути вдосконаленим за допомогою ще більш жадібної евристики, яка буде проводити пошук найкращої заміни лише по невеликій частині одного кластеру (кількість точок в якій буде задаватись параметром s). Складність однієї ітерації такого алгоритму складає $O(n*k*s)$, при цьому $s \ll n/k$, що радикально зменшує обчислювальну складність алгоритму. В роботі, що розглядає подібний алгоритм, показується, що зниження параметру s s до 2 і, навіть, до 1, практично не погіршує різноманітні метрики якості кластерів [6].

Також досить відомою є модифікація PAM під назвою clara [7]. Дана модифікація передбачає в ході ітерації вибір випадковим чином підмножини вершин і кластеризацію підграфу, який вони утворюють. Інші вершини просто розподіляються по найближчим отриманим медоїдам. Втрату інформації (оскільки

кластеризується лише підграф) пропонується компенсувати послідовним виконанням ітерацій на різних підмножинах вершин і вибором найкращого результату за наперед визначеними метриками якості.

Тестування запропонованого модифікованого методу

Точність та швидкість виконання кластеризації запропонованим модифікованим методом острівної кластеризації текстів перевірено на двох корпусах документів.

Перший корпус ***B*** складається з 50 текстів, присвячених Євробаченню та діяльності компанії SpaceX, відібраних з сайту BBC [8]. В даному корпусі текстів обох тематик порівну – по 25 новин.

Другий корпус ***R*** складається з 574 текстів, розподілених порівну між сімома різними тематиками. Тексти цього корпусу є попередньо обробленою підмножиною популярного тестового набору Reuters-21578 [9]. Попереднє оброблення даної підмножини полягало в приведені текстів до формату, з яким працювала програмна реалізація (виділення міток кластерів, перейменування файлів з текстами).

У зв'язку з простотою обох тестових корпусів (кількість текстів кожного кластеру є однаковою, кожен текст належить лише до одного кластеру) в якості міри точності кластеризації використано просте відношення кількості текстів, розподілених правильно за кластерами, до загальної кількості текстів в корпусі.

Програмна реалізація виконана мовою C# на платформі .NET. Для лематизації текстів на першому етапі методу острівної кластеризації використано бібліотеку Stanford CoreNLP [10], портовану на платформу .NET [11].

Точність виконання кластеризації оригінальним методом острівної кластеризації та модифікованим (з різними реалізаціями методу *k-medoids*) наведена на рис. 1. Як очікувалося, запропонований модифікований метод показав кращий результат для всіх реалізацій методу *k-medoids*, ніж оригінальний метод, завдяки ручному встановленню очікуваної кількості кластерів.

Рис. 1. Точність отриманої кластеризації

Результати тестування швидкості виконання кластеризації тестових корпусів з використанням оригінального та модифікованого методу наведені на рис. 2. Всі значення подані у вигляді співвідношення часу виконання кластеризації модифікованим методом до часу виконання кластеризації оригінальним острівним методом (в %). Таким чином, чим ближче до 0 значення, тим показана краща швидкість кластеризації (менше уповільнення, в порівнянні з оригінальним острівним методом).

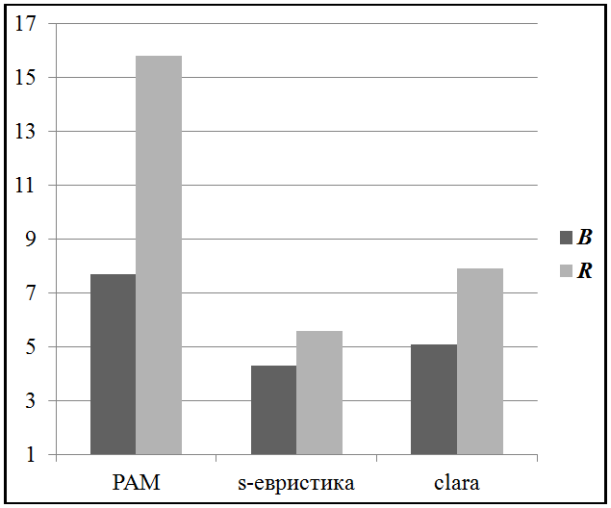

Рис. 2. Швидкість виконання кластеризації

Висновки

В даній роботі запропоновано модифікований метод острівної кластеризації текстових документів, який базується на ідеї розбиття графу кореляції термів на групи, що забезпечує можливість виконання ексклюзивної кластеризації та ручного встановлення очікуваної кількості кластерів. Як показало тестування точності та швидкості виконання кластеризації запропонованим методом, його використання є доцільним при наявності припущень про очікувану кількість кластерів документів та необхідності виконати ексклюзивну кластеризацію.

Можна виділити такі напрями подальшого вивчення та розвитку запропонованого модифікованого методу:

- тестування методу на корпусах відмінної від новинної тематики;
- розроблення модифікації, яка допускає автоматичне визначення кількості кластерів текстів;
- створення програмних бібліотек на різних мовах, що реалізовують запропонований метод.

Література

1. Information explosion [Електронний ресурс]. – Режим доступу : https://en.oxforddictionaries.com/definition/information_explosion. – Назва з екрану. – (Дата звернення: 15.12.2017).

2. Gantz J., Reinsel D. The digital universe in 2020: Big data bigger digital shadows and biggest growth in the far east // IDC iView: IDC Anal. Future. – 2012. – № 2007. – С. 1–16.

3. Berry M.W. Survey of Text Mining // Springer. – 2003.

4. Шмулевич М.М. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к аналізу тематической структуры новостного потока, а также ее динамики / М.М. Шмулевич, М.В. Киселев, В.С. Пивоваров // Интернет-математика 2005. – 2005. – С. 412–435.

5. Kaufman L. and Rousseeuw P.J. Clustering by means of Medoids. In: Y. Dodge and North-Holland, editor. Statistical Data Analysis Based on the L1-Norm and Related Methods. Springer US. – 1987. – P. 405–416.

6. Michael Ovelgonne Scalable Algorithms for Community Detection in Very Large Graphs. – 2011. – 146 p.

7. Kaufman L. and Rousseeuw P.J. Finding Groups in Data: An Introduction to Cluster Analysis. – 1990. – P. 126–163.

8. BBC News [Електронний ресурс]. – Режим доступу : http://www.bbc.com/news. – Назва з екрану. – (Дата звернення: 15.11.2017).

9. Reuters-21578 [Електронний ресурс]. – Режим доступу : http://www.daviddlewis.com/resources/testcollections/reuters21578/. – Назва з екрану. – (Дата звернення: 13.11.2017).

10. The Stanford CoreNLP Natural Language Processing Toolkit / [C. D. Manning, M. Surdeanu, J. Bauer] // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations / [C. D. Manning, M. Surdeanu, J. Bauer]., 2014. – P. 55–60.

11. SimpleNetNlp [Електронний ресурс]. – Режим доступу : https://github.com/yakivyusin/SimpleNetNlp. – Назва з екрану. – (Дата звернення: 15.11.2017).

References

1. Information explosion [accessed 2017 Dec 15]. https://en.oxforddictionaries.com/definition/information_explosion.

2. Gantz J., Reinsel D. The digital universe in 2020: Big data bigger digital shadows and biggest growth in the far east // IDC iView: IDC Anal. Future. – 2012. – №2007. – pp. 1-16.

3. Berry M.W. Survey of Text Mining // Springer. – 2003.

4. Shmulevich M.M., Kiselev M.V., Pivovarov V.S. Metod klasterizatsii tekstov, uchityivayuschiy sov-mestnuyu vstrechaemost klyuchevyih terminov, i ego primenenie k analIzu tematicheskoy strukturyi novostnogo potoka, a takzhe ee dinamiki // Internet-matematika 2005. – 2005. – pp. 412 435.

5. Kaufman, L. and Rousseeuw, P.J., Clustering by means of Medoids. In: Y. Dodge and North-Holland, editor. Statistical Data Analysis Based on the L1-Norm and Related Methods. Springer US. – 1987. – pp. 405–416.

6. Michael Ovelgonne Scalable Algorithms for Community Detection in Very Large Graphs. – 2011. – 146 p.

7. Kaufman, L. and Rousseeuw, P.J. Finding Groups in Data: An Introduction to Cluster Analysis. – 1990. – pp.126–163.

8. BBC News [accessed 2017 Nov 15]. http://www.bbc.com/news.

9. Reuters-21578 [accessed 2017 Nov 11]. http://www.daviddlewis.com/resources/testcollections/reuters21578/.

10. The Stanford CoreNLP Natural Language Processing Toolkit / [C. D. Manning, M. Surdeanu, J. Bauer] // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations / [C. D. Manning, M. Surdeanu, J. Bauer]., 2014. – pp. 55–60.

11. SimpleNetNlp [accessed 2017 Nov 15]. https://github.com/yakivyusin/SimpleNetNlp.

Рецензія/Peer review : 25.05.2018 р. Надрукована/Printed :14.07.2018 р.
Рецензент: д.т.н., проф. І.А. Дичка