УДК 004.91

Юсин Я.О.,
Заболотня Т.М., к.т.н.

ПІДХОДИ ДО ПОПЕРЕДНЬОГО ОБРОБЛЕННЯ ГРАФУ СУМІСНОЇ ЗУСТРІЧАЛЬНОСТІ ТЕРМІВ В МЕТОДІ ОСТРІВНОЇ КЛАСТЕРИЗАЦІЇ ТЕКСТІВ

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

yusin.yakiv@gmail.com
tetiana.zabolotnia@gmail.com

Запропоновано три нових підходи до попереднього оброблення графу сумісної зустрічальності термів в методі острівної кластеризації текстів. Визначено алгоритми, які реалізують дані підходи. Проведено тестування точності та швидкості виконання острівної кластеризації текстів з використанням запропонованих підходів

Ключові слова: кластеризація, острівна кластеризація, апроксимація графу, попереднє оброблення графу

Вступ

Починаючи з 1950-х років кількість інформації, що генерується людством, невпинно зростає. Це явище отримало назву інформаційного вибуху [1], і в першу чергу стосується текстових документів, що використовуються в науці, бізнесі, навчанні та інших сферах діяльності людини.

Відповідно до досліджень, лише за 2012 рік людство згенерувало 2.8 зетабайти інформації, і за прогнозами ця кількість подвоюватиметься кожні два роки [2]. Проте потенційно корисними є лише 23%, а структурованими – лише 5% цих даних [2]. Наслідком такого зростання обсягу текстових документів стала необхідність в розробленні методів автоматичної попередньої систематизації цих масивів даних перед їх подальшим обробленням та/чи аналізом.

В таких умовах розроблення нових та вдосконалення існуючих методів автоматичної (unsupervised) кластеризації текстових документів, тобто розбиття корпусів документів на наперед не задані підмножини (кластери), стає актуальною задачею [3]. Причому тут слід відмітити, що крім проблем, спільних для всіх задач кластеризації, додаткова складність кластеризації текстів визначається необхідністю індикації змісту знайденого кластеру – оскільки зазвичай результати кластеризації інтерпретуються безпосередньо людиною, вона повинна розуміти зміст знайденого кластеру і чому певні тексти були віднесені саме до нього. Саме тому в якості критеріїв ефективності методів кластеризації текстів зазвичай використовується точність та швидкість виконання кластеризації тестових попередньо розмічених корпусів.

Постановка завдання

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

Це відносно молодий метод, який існує в різних модифікаціях. В основі цього метода 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.

Основною модифікацією методу острівної кластеризації є метод, описаний в роботі [4], який базується на використанні графа сумісної зустрічальності термів. На далі, в даній роботі, під терміном «метод острівної кластеризації», буде матись на

увазі саме ця модифікація. Вона обрана для дослідження та подальшого удосконалення, тому що має такі властивості [5]:

- інтерпретованість знайдених кластерів;
- можливість віднесення документу більше, ніж до одного кластеру;
- обмеженість часу роботи алгоритмів, що реалізують метод, квадратичною залежністю від кількості оброблюваних документів;
- наявність засобів боротьби з омонімією та синонімією.

Метод острівної кластеризації текстових колекцій складається з етапів [4]:

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

В рамках даної роботи розглянемо етап попереднього оброблення графу сумісної зустрічальності термів перед його кластеризацією, оскільки його удосконалення може сприяти підвищенню точності кластеризації.

Таким чином, **метою** даної роботи є підвищення ефективності кластеризації текстових даних методом острівної кластеризації за критерієм точності шляхом розроблення та дослідження різних підходів до процедури отримання наближення графу кореляції термів.

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

- формулювання нових підходів до процедури попереднього оброблення графу кореляції термів;
- розроблення алгоритмів, що реалізують запропоновані підходи;
- аналіз точності кластеризації текстових документів з використанням запропонованих підходів;
- аналіз ефективності розроблених алгоритмів за критерієм швидкості виконання кластеризації.

Оригінальна процедура оброблення графу сумісної зустрічальності термів

Метою попереднього оброблення графу сумісної зустрічальності термів є зменшення кількості вершин та ребер графу, що спрощує оброблення останнього (за критерієм кількості операцій), а також зменшує кількість необхідної пам'яті для його зберігання. На виході процедури оброблення ми отримуємо деяке наближення або апроксимацію вхідного графу, при цьому його досліджувані характеристики повинні не змінюватись або змінюватись на невелику похибку. У випадку використання графу для кластеризації такою характеристикою графу є його кластерна структура – процедура попереднього оброблення повинна не змінювати її або змінювати на величину, якою можливо знехтувати.

В роботі [4], що описує метод острівної кластеризації, попереднє оброблення графу пропонується виконувати таким чином: якщо вага ребра (яка представляє кореляцію відповідних термів між собою: чим вага менша, тим більш корельовані між собою терми) є більшою певного значення p_c, це ребро видаляється з графу. Значення p_c обчислюється за формулою (1), де N_{terms} – загальна кількість термів, а N_{docs} – кількість документів в колекції.

$$p_c = \frac{0.03}{\max(N_{terms}^2, N_{docs})} \qquad (1)$$

Очевидним є той факт, що при використанні одного підходу для кластеризації отриманого наближення графа, саме

від якості процедури попереднього оброблення буде залежати якість остаточного результату кластеризації. Головний недолік вищезгаданої процедури, яку використовує метод острівної кластеризації, полягає в тому, що вона обробляє всі ребра однаково, тобто глобально, що призводить до втрати кластерів термів. Проілюструвати даний недолік і те, як він впливає на якість результату кластеризації, можна за допомогою такого прикладу. Нехай маємо частину графу сумісної зустрічальності термів, що зображена на рис. 1. Ребра, що відповідають відсутній кореляції термів між собою, не показані; суцільним зображено ребра, вага котрих менше значення p_c. Ребра, вага котрих дорівнює $1.1p_c$, зображено штриховою лінією.

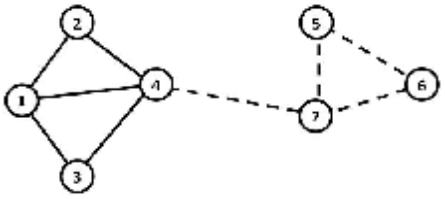

Рис. 1. Частина графу сумісної зустрічальності термів

Після виконання описаної процедури оброблення графу отримане його наближення буде містити лише один кластер, що складається з термів №№1-4, при цьому кластер термів №№5-7 буде втрачено (і відповідний йому кластер документів). Позбутись цього недоліку, підвищивши якість результатів кластеризації, можливо, використовуючи інші підходи до попереднього оброблення графу.

Використання глобального порогу

Даним підходом передбачається певне узагальнення оригінальної процедури попереднього оброблення графу в методі острівної кластеризації [4].

Як зрозуміло з назви підходу, отриманий граф сумісної зустрічальності термів пропонується фільтрувати за допомогою деякого глобального порогу величини кореляції, і всі ребра, що відповідають кореляції меншій, ніж цей поріг, мають бути видалені з графу. Отримана апроксимація графу може бути використана надалі для кластеризації текстів.

Позначимо кореляцію термів i та j між собою як p_{ij}, значення якої відповідає вазі відповідного ребра графу сумісної зустрічальності, а значення глобального порогу – як $threshold$. Тоді даний підхід можна реалізувати у такий спосіб:

1. На вході отримуємо граф $G = (V, E, w)$.

2. Вагу кожного ребра графу $e = (i, j)$ порівнюємо з порогом.

3. Якщо $w_{ij} = p_{ij} \le threshold$, додаємо ребро e до G_{sparse}.

4. Повертаємо отриманий G_{sparse}.

Як вже зазначено при розгляді методу острівної кластеризації, такий підхід впливає на склад найбільш малозначимих кластерів, змінюючи їх.

Описаний підхід є найшвидшим серед запропонованих підходів, оскільки він є однопрохідним і для кожного ребра виконує лише одну операцію – порівняння ваги ребра з порогом. Проте разом з тим він зберігає недоліки оригінальної процедури навіть при використанні різних значень для порогу.

Використання відсоткового порогу

Цей підхід можливо розглядати як модифікацію попереднього підходу, яка замість наперед визначеної величини глобального порогу використовує величину, отриману з самого графу.

В ході застосування цього підходу відбувається сортування всіх ребер графу за їх вагою. В результаті в апроксимацію вхідного графу потрапляє наперед визначений відсоток ребер з найбільшою кореляцією термів між собою.

Якщо ми позначимо значення відсоткового порогу як s, тоді даний підхід матиме таку реалізацію:

1. На вхід подаємо граф $G = (V, E, w)$.

2. Відсортовуємо всі ребра в E за їх вагою w_{ij}.

3. Додаємо верхні s% ребер отриманого відсортованого списку до G_{sparse}.

4. Повертаємо отриманий G_{sparse}.

Проведені авторами експерименти на різноманітних текстових колекціях показали, що значення p_c розглянутого першого підходу відповідає діапазону $(3,7)$ значень s. Таким чином, при використанні більшого значення, ніж права межа цього діапазону, в наближення графу потрапить більше ребер, ніж при використанні першого підходу. Це також призведе до деякого збільшення якості результату кластеризації.

Описаний підхід потребує більше часу та більшої кількості операцій, ніж перший, і є по своїй суті двопрохідним. Перший прохід – це сортування всіх ребер, другий – прохід по частині отриманого списку для додавання ребер до апроксимації.

Використання ефективного опору

В роботі [6] пропонується підхід до апроксимації графу кореляції термів, що заснований на використанні ефективного опору.

Саме поняття ефективного опору для графу будується на основі розгляду цього графу як такого, що представляє собою електричне коло [7]. Ефективний опір R_{ij} між парою вершин графу i та j – це електричний опір в колі, виміряний між вершинами i та j, де ребра графу це резистори з електричною провідністю, що відповідають вазі ребра w_{ij}. Це так званий електричний сенс цієї метрики. Також визначають броунівський сенс – в такому випадку R_{ij} - це величина, що пропорційна середньому часу досяжності вершини j з вершини i при випадковому блуканні по графу.

Цей підхід реалізуємо так [6]:

1. На вхід подаємо граф $G = (V, E, w)$.

2. Вибираємо випадкове ребро e з графу G з ймовірністю p_e, що пропорційна величині $w_{ij}R_{ij}$.

3. Додаємо вибране ребро до G_{sparse} з вагою w_{ij}/qp_{ij}.

4. Повторюємо незалежно q разів з заміною та додаванням ваги ребер, якщо ребро вибирається більше одного разу.

Даний підхід працює за час, наближений до лінійного, а отримана апроксимація графу містить $O\left(n\log n/q^2\right)$ вершин (n – кількість вершин вхідного графу).

Відмова від попереднього оброблення

Відмова від попереднього оброблення графу також може використовуватись в якості окремого підходу до оброблення графу.

В такому випадку замість певного наближення на подальших етапах методу острівної кластеризації відбувається кластеризація цілого графу сумісної зустрічальності термів. Таким підходом гарантується стовідсоткове збереження кластерної структури, на відміну від інших підходів. Недоліками відмови від попереднього оброблення є необхідність зберігати в пам'яті весь граф, а його кластеризація буде потребувати найбільшу кількість часу, в порівнянні з іншими підходами. Проте для невеликих корпусів текстів сумарний час роботи методу острівної кластеризації може не погіршитись, за рахунок переходу відразу до кластеризації.

Формально цей підхід є частковим випадком відсоткового порогу зі значенням 100 і частковим випадком глобального порогу, що дорівнює значенню відсутньої кореляції.

Тестування якості запропонованих підходів

Точність та швидкість кластеризації, що використовує запропоновані підходи до попереднього оброблення графу сумісної зустрічальності термів, відповідно до

методу острівної кластеризації, 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 по кластерам, до загальної кількості текстів в корпусі.

Для оцінювання точності та швидкості запропонованих підходів до попереднього оброблення графу кореляції термів програмно реалізовано модифікацію методу острівної кластеризації, яка виконує ексклюзивну (таку, що кожен текст відноситься лише до одного кластеру) кластеризацію. Програмна реалізація виконана мовою C# на платформі .NET з використанням шаблону проектування «Шаблонний метод» [10]. Його використання забезпечило можливість один раз визначити всі основні кроки методу острівної кластеризації, змінюючи лише реалізацію попереднього оброблення графу сумісної зустрічальності термів.

Для лематизації текстів на першому етапі методу острівної кластеризації використано бібліотеку Stanford CoreNLP [11], портовану на платформу .NET [12].

В якості глобального порогу використано значення p_c, що відповідає оригінальній процедурі попереднього оброблення методу острівної кластеризації. В якості відсоткового порогу обрано значення 15 як таке, що вдвічі перевищує знайдену верхню межу, якій відповідає глобальний поріг.

Точність острівної кластеризації з використанням запропонованих підходів до попереднього оброблення графу кореляції термів наведена на рис. 2. Як бачимо, отримана точність кластеризації корпусів текстів відповідає теоретичним припущенням: найгіршу точність отримано при використанні оригінального оброблення з використанням глобального порогу, найкращу – при відмові від попереднього оброблення графу кореляції.

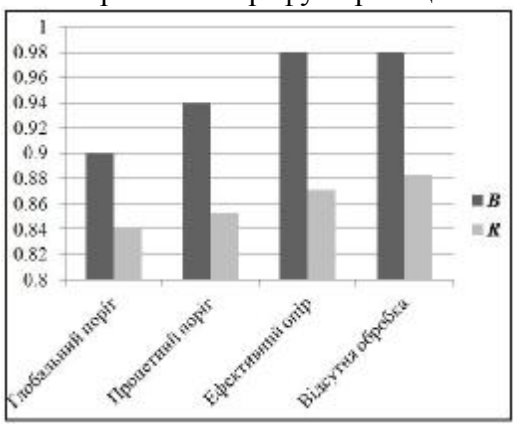

Рис. 2. Точність отриманої кластеризації

Результати тестування швидкості виконання кластеризації тестових корпусів з використанням запропонованих підходів наведені на рис. 3.

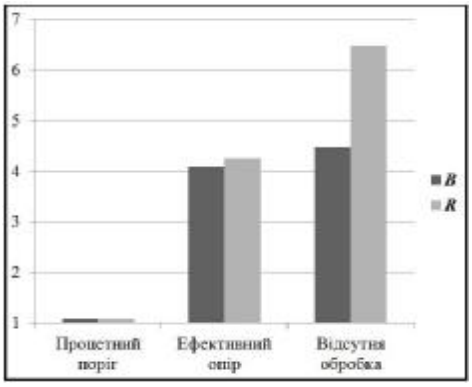

Рис. 3. Швидкість виконання кластеризації

Всі значення подані у вигляді співвідношення часу виконання кластеризації з використанням певного підходу до часу виконання кластеризації з використанням оригінальної процедури. Таким чином, чим ближче до 1 значення, тим показана краща швидкість кластеризації.

Як показало тестування, найвищу швидкість має використання відсоткового підходу – він лише на 7-9% повільніший, ніж оригінальний. Найгіршу швидкість, як очікувалось, отримано при відмові від попереднього оброблення графу, проте в випадку досить малого тестового корпусу вона наближається до швидкості використання підходу ефективного опору.

Висновки

Таким чином, в статті показана доцільність розроблення нових підходів до попереднього оброблення графу сумісної зустрічальності термів в методі острівної кластеризації, запропоновано три нових підходи (використання *відсоткового порогу*, *ефективного опору* та *відмова від оброблення* відповідно), визначено основні кроки алгоритмів, що їх реалізують.

Також проведено тестування точності та швидкості виконання кластеризації відповідно до запропонованих підходів, яке підтвердило теоретичні припущення щодо характеристик цих підходів.

Напрями подальшого вивчення та розвитку запропонованих підходів:

- розроблення методики вибору певного підходу серед запропонованих в залежності від можливих характеристик текстового корпусу;
- створення програмних бібліотек кластеризації текстів, що реалізують дані підходи для різних мов програмування.

Список літератури

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

5. Шмулевич М. М. Метод автоматической кластеризации текстов, основанный на извлечении из текстов имен объектов и последующем построении графов совместной встречаемости ключевых термов : дис. канд. фіз.-мат. наук / Шмулевич Марк Михайлович – Москва, 2009. – 120 с.

6. Spielman D.A., Srivastava N. Graph sparsification by effective resistances // Symposium on Theory of Computing 2004. – 2004. – С.81-90.

7. Ghosh A., Boyd S., Saberi A. Minimizing effective resistance of a graph // 17th International Symposium on Mathematical Theory of Networks and Systems. – 2006. – С.1185-1196.

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. Template Method / E.Gamma, R. Helm, R. Johnson, J. Vlissides // Design Patterns / E.Gamma, R. Helm, R. Johnson, J. Vlissides., 1994. – С. 325–330.

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

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

Статтю подано до редакції 19.12.2017