Аналіз алгоритмів кластеризації для наукових статей на основі підсумовування за допомогою нейронних мереж

М. В. Бевза

Київський національний університет імені Тараса Шевченка
*Corresponding author. E-mail:maksymbevza@gmail.com

Paper received 02.07.21; Accepted for publication 19.07.21.

https://doi.org/10.31174/SEND-NT2021-255IX32-07

Анотація. Вступ. Ріст кількості інформації доступної в мережі інтернет є експоненційним. Фільтрування кісної інформації від неякісної стає дедалі важчим. Пошук інформації з надійних джерел є важливим завданням. Одне з високоякісних джерел інформації - це наукові дослідження. **Метою роботи** є побудова системи, яка може кластеризувати наукові статті, враховуючи її семантичні властивості, та порівняти алгоритми кластеризації для цієї задачі. **Результати.** Проаналізовано структуру наукових статей та визначено особливості, які слід враховувати під час кластеризації статей за їх текстовими семантичними властивостями. Описано процес побудови кластерів. Характеристики текстової частини контенту попередньо сформовано за допомогою нейронної мережі BERT. Розроблено систему, яка опрацьовує текстову частину наукової статті і будує кластеризацію. Аналіз роботи розробленої системи показав переваги і недоліки методів кластеризації. **Висновки.** Використання розробленої системи, яка на вхід отримує характеристики текстової частини наукової статті, дає змогу комплексно оцінити зв'язок наукових статей один між іншим, а також інтерпретувати ці зв'язки.

Ключові слова: *штучний інтелект, оброблення природної мови, кластеризація.*

Вступ. Кількість інформації, доступна пересічній людині росте експоненційно, це зв'язано з легкістю створення і поширення інформації за допомогою мережі "Інтернет", та сучасних соціальних мереж. Інформацію може поширювати кожен, хто має доступ до цих мереж, а це - мільярди людей. Все нагально постає питання перевірки правильності інформації.

Одним з джерел якісної і достовірної інформації є наукові статті в рецензованих журналах. У кожної такої статті є декілька людей-рецензентів, які критично прочитали статтю, перевірили логічність пар гіпотеза-висновок, кількість статистичних даних та способів їх обробки.

Пошук відповіді на поставлене запитання серед великої кількості статей автоматизованим способом - це нетривіальне завдання. У цій роботі ми будуємо алгоритм, який спрощує групує наукові статті у категорії для спрощення пошуку відповіді на поставлене запитання

Постановка проблеми. Кластеризація - одна з найбільших тем науки про дані. Задача кластеризації - це поділ об'єктів на групи (кластери), які у випадку текстових даних несуть схожі семантичні характеристики. Кластеризація буває строга і м'яка. Строга кластеризація передбачає побудову груп - множин об'єктів, які не перетинаються і які в об'єднанні дають всю множину об'єктів - універсум. Така кластеризація є покриттям множини об'єктів. М'яка кластеризація - це такий поділ об'єктів на групи, коли один об'єкт може належати від одного до декількох груп одночасно.

У роботі Hu et al. (2008) Enhancing Text Clustering by Leveraging Wikipedia Semantics [1] розглядається кластеризація текстів базуючись на онтології WordNet. В їх підході будуються так звані "мішки слів" (bag of words) - це векторизація тексту по кількості слів які в них входять. Кластеризацію проведено на статтях з "Вікіпедії".

У роботі Beil et al. (2002) Frequent term-based text clustering [2] виділяють спеціальні слова-представники текстів для здійснення кластеризації. Ці представники отримано за рахунок обрахунки частоти вживання слів і таким чином було виділено ключові слова тексту, по яким була зроблена кластеризація.

- У цій роботі розглядається кластеризація текстів наукових робіт. Для цих робіт є характерним наявність структури. Кожна зі статей має таку структуру:
- анотація;
- вступ;
- аналіз інших робіт;
- секції, що описують спосіб перевірки гіпотез;
- висновки;
- додаткові секції.

Наша робота ставить перед собою завдання побудувати алгоритм, який враховує саме цю структуру, гіпотези і висновки для кластеризації текстів, а також порівняти ці алгоритми кластеризації.

Підготовка вхідних даних для кластеризації. Наукові статті мають специфічну структуру розділів. Для задачі кластеризації по тематиці розглянутих питань (гіпотезою і її підтвердженням або спрощенням) ми виділяємо такі признаки, які будуть перетворені в текстові характеристики.

1. Ключові слова
2. Анотація
3. Вступ
4. Текст роботи
5. Висновок

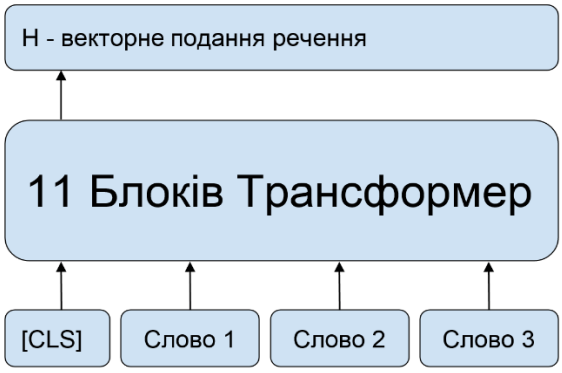
Гіпотезу і відповідь подано в анотації, вступі та висновку, тому вони враховуються окремо від загального тексту роботи.

Отримання характеристик з текстової частини контенту. Для генерації ознак з текстових даних було використано нейронну мережу, побудовану на основі архітектури Transformer [3]. Оскільки навчання нейронної мережі для оброблення потребує величезних обсягів даних, то цю систему ми тренували на основі попередньо навченої мережі на мовних моделях системи BERT [4]. На відміну від системи word2vec [5], система BERT враховує контекстну залежність слів в межах речення. На відміну від системи RoBERTa [6], система BERT має кращу швидкодію, що суттєво впливає на час тренування і виконання алгоритму.

На Рис 1. зображено схему нейронної мережі BERT. Ця система дає змогу приймати на вході ціле речення, розраховувати 12-рівневу нейронно-мережеву архітектуру і отримати контекстно-залежні репрезентації слів. Цей підхід має перевагу над підходом незалежного оброблення слів, оскільки враховує контекст, в якому кожне слово використовується.

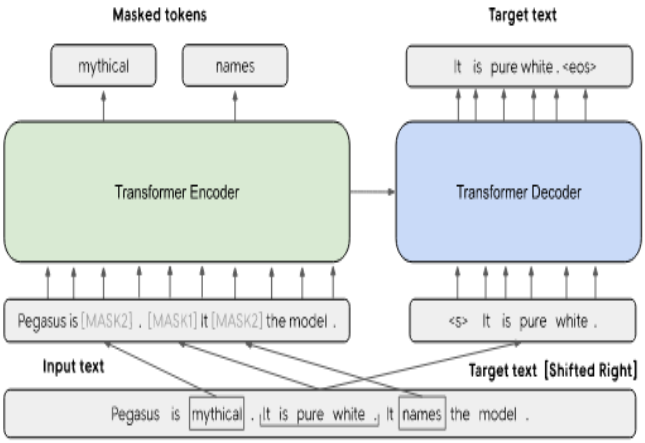
Система BERT маскує слова і намагається по контексту передбачити ці слова, таким чином досягається додатковий рівень узагальнення і зменшується перенавчання.

Кожне слово має відповідний 768-розмірний вектор, який відповідає його числовій репрезентації. Для побудови єдиного значення для всього тексту здійснено усереднення за всіма словами в тексті як спосіб узагальнення.



1. Схема роботи нейронної мережі BERT.

Підсумувовування. Для розв'язання задачі підсумовування всього тексту статті ми використовуємо мережу Pegasus [7], яка була натренована для розв'язання задачі підсумовування текстів. Це нейронна мережа побудована на основі мережі BERT, яка в свою чергу базується на моделі побудови нейронних мереж Transformer. Особливістю нейронної мережі Pegasus є те, що вона аналогічно до BERT маскує частини тексту. У системі Pegasus аналогічно до BERT маскується частина тексту, у цьому разі - ціле речення, а не окремі слова, і система має завдання відновити це речення. Це покращує узагальненість мережі і покращує надійність її роботи.



2. Схема роботи нейронної мережі Pegasus.

Кластеризація. Алгоритм кластеризації зважений відповідно до надійності і розмірності ознак. Найбільш надійними ознаками ми вважаємо ключові слова, далі - висновок, далі - анотація, далі - підсумування статті. Кластеризація проводилась з допомогою алгоритму K-Means, Ієрархічної кластеризації та алгоритму "поширення спорідненості".

Вимірювання відстаней у векторному просторі речень та слів було зроблено через косинусну відстань. Відстань між векторами A і B визначається за такою формули 3:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

3.

Кластеризація K-Means

Згідно з алгоритмом описаним у роботі [8] для побудови кластерів алгоритм ітеративно робить такі кроки:

1. підрахунок центроїдів кластерів як середнє значення вектора у відповідному кластері;
2. для кожного об'єкту кластеризації підраховується до якого центроїда вона має найменшу відстань;
3. призначити для кожного об'єкту номер кластеру відповідно до того який центроїд мав найменшу відстань до кластеру.

Значення K для K-means було обрано 63 для отримання 63 кластерів.

Ієрархічна кластеризація

Алгоритм побудови кластерів такий.

Спочатку всі об'єкти належать до своїх власних кластерів. Далі ітеративно виконуються такі кроки:

1. вибирається пара кластерів A та B, які мають найменшу відстань між середніми значеннями їх векторів;
2. кластери A та B об'єднуються в один кластер

Алгоритм зупиняється коли досягнуто певного порогу дальності між кластерами, або коли отримано необхідну кількість кластерів K.

Алгоритм "поширення спорідненості"

Цей алгоритм аналогічно до ієрархічної кластеризації працює ітеративно.

Нехай x_1 - x_n - набір точок даних, не роблячи припущень про їх внутрішню структуру, і нехай s - функція, яка кількісно визначає подібність між будь-якими двома точками.

Діагональ s (тобто s(i,i)} особливо важлива, оскільки вона представляє перевагу екземпляра, що означає, наскільки ймовірно, що конкретний екземпляр стане зразком (представником) для кластеру. Коли діагоналі встановлено одне і те ж значення для всіх входів, це визначає кількість кластерів алгоритм дасть на виході. Значення діагоналі, близьке до мінімально можливої подібності - це створює менше кластерів, тоді як значення, близьке або більше за максимально можливу подібність, дає велику кількість кластерів. Це значення зазвичай встановлюють як середня подібності всіх пар входів.

Алгоритм працює шляхом чергування двох кроків передачі повідомлень, які оновлюють дві матриці:

Матриця "відповідальності" R має значення r(i,k), які кількісно визначають, наскільки добре підходить x_k, щоб служити зразком для x_i, порівняно з іншими зразками кандидатів для x_i.

Матриця "доступності" A містить значення a(i,k), які представляють, наскільки "доцільним" було б, щоб x_i вибрав x_k як його приклад, враховуючи перевагу інших точок щодо x_k як зразка.

Обидві матриці ініціалізовані до всіх нулів і можуть розглядатися як таблиці ймовірностей журналів. Потім алгоритм виконує наступні оновлення ітеративно:

Спочатку надсилаються оновлення про відповідальність 4:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k} \{a(i,k') + s(i,k')\} \quad 4.$$

Потім доступність оновлюється за наступними формулами 5

$$a(i,k) \leftarrow \min\left(0, r(k,k) + \sum_{i' \notin \{i,k\}} \max(0, r(i',k))\right)$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k)) \quad 5.$$

Ітерації виконуються до тих пір, поки межі кластера не залишаться незмінними протягом ряду ітерацій, або не буде досягнуто якесь заздалегідь визначене число (ітерацій). Зразки витягуються з кінцевих матриць як ті, чия "відповідальність + доступність" для себе є позитивною (тобто r(i,i) + a(i,i) > 0).

Апробація алгоритму. Для формування датасету використано наукові статті опубліковані в роки з 2015 до 2020. Для роботи ми взяли 4000 статей з ресурсу https://arxiv.org/. Обробка цих даних і отримання текстових характеристик на нейронній мережі Pegasus зайняла 144 години робочого часу на GPU.

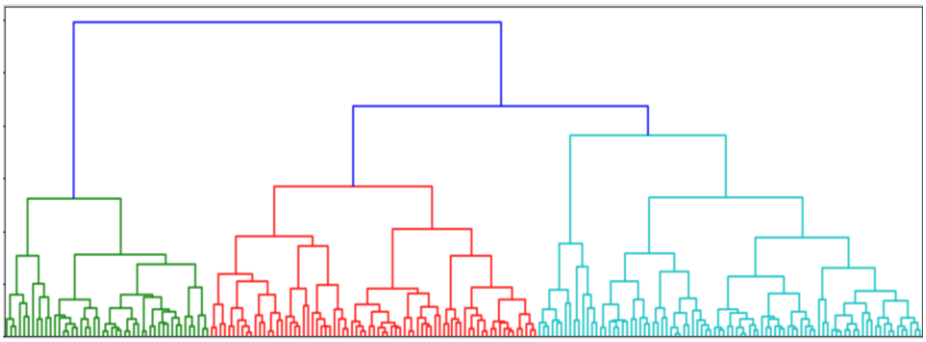
Кластеризація K-Means. Для алгоритму K-Means ми використали значення 63 для кількості кластерів. Це число вибрано з розрахунку K = sqrt(N), де N - кількість кластерів.

Для візуалізації результатів побудуємо кластеризації з'єднань різних статей.

Переваги алгоритму: легко інтерпретувати кластери; легкість в контролі кількості кластерів

Недоліки алгоритму: об'єкти на межі кластерів важко віднести до якоїсь категорії

Ієрархічна кластеризація. Верхівка алгоритму ієрархічної кластеризації зображена на Рис 6.



6. Візуалізація роботи ієрархічної кластеризації

Ми чітко прослідковуємо, три категорії текстів за їх тематикою:

1. Computer Science, Mathematics
2. Physics
3. Quantitative Biology

Переваги алгоритму: легко інтерпретувати кластери і їх частини; швидкодія алгоритму. Недоліки алгоритму: низькорівнева ієрархія не має сенсу

Алгоритм поширення спорідненості. При аналізі утворених кластерів ми побачили, що об'єкти в кластерах мають чітко вираженого представника, який описує весь кластер вцілому. Це дозволяє інтерпретувати кластер і робити наступні висновки або пошук по кластеру.

Переваги алгоритму: легко інтерпретувати кластери; простота контролю дальності кластерів через параметризацію діагоналі; легкість інтерпретації

Недоліки алгоритму: значний час роботи і повільна сходимість алгоритму

Висновки. Розроблений алгоритм кластеризації, який на вхід отримує характеристики текстової частини наукової статті та ключових слів, отриманої за допомогою нейронної мережі BERT та Pegasus і зважений відповідно до важливості і розмірності вхідних ознак.

Проаналізовано такі алгоритми до кластеризації наукових статей:

1. K-means;
2. ієрархічна кластеризація;
3. поширення спорідненості.

Проаналізовано переваги і недоліки кожного з методів, а також особливості використання.

REFERENCES

1 Hu et al. (2008) Enhancing Text Clustering by Leveraging Wikipedia Semantics. http://www.cse.ust.hk/faculty/qyang/Docs/2008/fp422Hujian.pdf
2 Beil et al. (2002) Frequent term-based text clustering https://dl.acm.org/doi/abs/10.1145/775047.775110
3 Vaswani et al. Attention Is All You Need. NIPS 2017 Proceedings https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
4 Devlin et al. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/pdf/1810.04805.pdf
5 Mikolov, T., Chen, K., Carrado, G. and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. 1st ed.
6 Lie et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/pdf/1907.11692.pdf
7 Zhang et al. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization https://arxiv.org/pdf/1912.08777.pdf
8 J. A. Hartigan, M. A. Wong, A K-Means Clustering Algorithm https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2346830

Analysis of clustering algorithms of scientific papers using summarisation via neural networks.

M. Bevza

Abstract. Introduction**.** The growth in the amount of information available on the Internet is exponential. Filtering quality information from low-quality information is becoming increasingly difficult. Finding information from reliable sources is an important task. One of the high-quality sources of information is scientific research papers. The aim of the work is to build a system that can cluster scientific articles, taking into account its semantic properties, and compare clustering algorithms for this task. **Results.** The structure of scientific articles is analyzed and the features that should be taken into account when clustering articles by their textual semantic properties are determined. The process of cluster construction is described. The characteristics of the text part of the content are pre-formed using the neural network BERT. A system has been developed that processes the text part of a scientific article and builds clustering. Analysis of the developed system showed the advantages and disadvantages of clustering methods. **Conclusions.** The use of the developed system, which receives the characteristics of the text part of a scientific article, allows to comprehensively assess the relationship of scientific articles to each other, as well as to interpret these links.

Keywords: *artificial intelligence, natural language processing, clustering,neural networks, summarization, bert, pegasus, clusterization, neural networks.*