УДК 004.89
DOI: 10.18523/2617-3808.2020.3.107-113

Жежерун О. П., Борозенний С. О., Ніверовський М. М.

ВИКОРИСТАННЯ АЛГОРИТМУ LSA ДЛЯ КЛАСТЕРИЗАЦІЇ ЗАДАЧ ІЗ ГЕОМЕТРІЇ

У роботі розглянуто метод LSA (латентно-семантичного аналізу), зокрема його найпоширеніший варіант, що базується на сингулярному розкладі матриці (SVD). На його основі реалізовано алгоритм кластеризації задач і застосовано на прикладі кластеризації задач із геометрії.

Ключові слова: LSA, LSI, SVD, кластеризація.

Вступ

На цей момент існує величезна кількість алгоритмів кластеризації. Основна ідея більшості з них – об'єднати однакові послідовності в один клас або кластер на основі подібності. Як правило, вибір алгоритму залежить від поставленого завдання. Що ж до текстових даних, то тут порівнюваними складовими слугують послідовності слів і їхніх атрибутів (наприклад, вага слова у тексті, тип іменованої сутності, тональність тощо). Тож тексти спочатку перетворюються на вектори, з якими проводять різного типу маніпуляції. При цьому, як правило, постає низка проблем, пов'язаних із:

- вибором первинних кластерів;
- залежністю якості кластеризації від довжини тексту;
- визначенням загальної кількості кластерів тощо.

Однак найскладнішою проблемою є відсутність зв'язку між близькими за змістом текстами, в яких використовується різна лексика. У таких випадках об'єднання має відбуватися не тільки на основі подібності, а ще й на основі семантичної суміжності або асоціативності.

Одним із методів, що дає змогу розв'язувати подібні задачі, є латентно-семантичний аналіз (ЛСА, *англ*. Latent semantic analysis, LSA).

ЛСА – це метод оброблення інформації, який аналізує набір документів і знаходить терміни, які там ужито, й на основі цього виявляє характерні фактори, тематики, які характеризують зміст документа.

Визначають такі типи кореляції:

«слово – слово»;

«слово – параграф»;

«параграф – параграф».

Саме цими трьома типами мислить людина, зіставляючи частини тексту зі змістом. Технологія ЛСА враховує не лише частотність вживання тексту, а й латенті (глибинні) зв'язки.

Першу статтю з автоматичної класифікації документів за кназвою «Automatic Document Classification» [4] було опубліковано в журналі «Journal of the ACM» на початку 1963 р., у ній було уперше описано метод факторного аналізу як засіб для пошуку інформації. Факторний аналіз – це метод, який визначає зв'язок між значеннями змінних.

Алгоритм комп'ютерного пошуку інформації, який використовував латентно-семантичну структуру, запатентувала у 1988 р. команда дослідників: Scott C. Deerwester (Chicago, IL), Susan T. Dumais (Berkeley Heights, NJ), George W. Furnas (Madison, NJ), Richard A. Harshman (London, CA), Thomas K. Landauer (Summit, NJ), Karen E. Lochbaum (Chatham, NJ), Lynn A. Streeter (Summit, NJ).

Опис та застосування алгоритму ЛСА для вирішення задач пошуку було розглянуто в роботах [1] і [2].

У цій роботі було досліджено можливість застосування латентно-семантичного аналізу для кластеризації текстів (задач із геометрії), для цього було розроблено алгоритм і необхідне програмне забезпечення.

Результати

Для проведення дослідження було вибрано такий набір задач із геометрії [3]:

1. Знайдіть кут трикутника, якщо два інші його кути дорівнюють 35° і 96°.
2. Один із кутів трикутника у 3 рази менший від другого кута та на 35° менший від третього. Знайдіть кути трикутника.
3. Знайдіть кути трикутника, якщо їхні градусні міри відносяться як 2 : 3 : 7.
4. Знайдіть кути рівностороннього трикутника.
5. Знайдіть кути рівнобедреного прямокутного трикутника.
6. Кут при основі рівнобедреного трикутника дорівнює 63°. Знайдіть кут при вершині цього трикутника.
7. Знайдіть кути при основі рівнобедреного трикутника, якщо кут при вершині дорівнює 104°.

8. Знайдіть кути рівнобедреного трикутника, якщо кут при вершині в 4 рази більший за кут при основі.
9. Знайдіть кути рівнобедреного трикутника, якщо кут при основі на 48° менший від кута при вершині.
10. Знайдіть кути рівнобедреного трикутника, якщо один із них дорівнює: 1) 110°; 2) 50°. Скільки розв'язків має задача?
11. Периметр трикутника дорівнює 30 см. Чи може одна з його сторін дорівнювати: 1) 20 см; 2) 15 см?
12. Довжини двох сторін трикутника дорівнюють 7 см і 9 см. Чи може периметр цього трикутника дорівнювати: 1) 20 см; 2) 32 см; 3) 18 см?
13. Чи існує трикутник, одна зі сторін якого на 2 см менша від другої та на 6 см менша від третьої, а периметр дорівнює 20 см?
14. У трикутнику ABC відомо, що AC > 90°. На стороні BC позначили довільну точку B. Доведіть, що AB > AC.
15. Доведіть, що відрізок, який сполучає вершину рівнобедреного трикутника з точкою, яка лежить на його основі, не більший за бічну сторону трикутника.
16. Доведіть, що відрізок, який сполучає вершину трикутника з точкою, яка лежить на протилежній стороні, не більший хоча б за одну з двох інших сторін.
17. Дві сторони рівнобедреного трикутника дорівнюють 9 см і 20 см. Знайдіть третю сторону трикутника.
18. Дві сторони рівнобедреного трикутника дорівнюють 8 см і 16 см. Знайдіть довжину основи цього трикутника.
19. Основа D висоти AD трикутника ABC є внутрішньою точкою відрізка BC. Відомо, що кут BAD > кута CAD. Порівняйте довжини сторін AB і AC.
20. Основа D висоти AD трикутника ABC є внутрішньою точкою відрізка BC. Відомо, що AB > AC. Порівняйте величини кутів BAD і CAD.
21. Доведіть, що в трикутнику будь-яка сторона менша від половини периметра.
22. У трикутнику ABC провели бісектрису BD. Доведіть, що AB > AD і BC > CD.
23. На стороні AC трикутника ABC позначили точку D. Відомо, що AD > BD. Доведіть, що AC > BC.
24. Один із гострих кутів прямокутного трикутника дорівнює 43°. Знайдіть другий гострий кут.
25. У рівнобедреному трикутнику ABC (AB = BC) проведено висоту AH. Знайдіть кут CAH, якщо кут B = 76°.
26. Кут між основою рівнобедреного трикутника та висотою, проведеною до бічної сторони, дорівнює 19°. Знайдіть кути цього трикутника.
27. На сторонах кута з вершиною в точці B позначили точки A і C так, що AB = BC. Через точки A і C провели прямі, які перпендикулярні до сторін BA і BC відповідно та перетинаються в точці O. Доведіть, що промінь BO – бісектриса кута ABC.
28. Доведіть, що висоти рівнобедреного трикутника, проведені до його бічних сторін, є рівними.
29. Доведіть, що коли дві висоти трикутника рівні, то цей трикутник є рівнобедреним.
30. Доведіть рівність прямокутних трикутників за катетом і бісектрисою, проведеною з вершини прямого кута.
31. Доведіть рівність прямокутних трикутників за катетом і висотою, проведеною з вершини прямого кута.
32. Доведіть рівність прямокутних трикутників за катетом і бісектрисою, проведеною з вершини прилеглого до цього катета гострого кута.
33. Доведіть рівність прямокутних трикутників за катетом і медіаною, проведеною до другого катета.
34. Пряма перетинає сторони AB і BC трикутника ABC відповідно в точках M і K, які є серединами цих сторін. Доведіть, що вершини цього трикутника рівновіддалені від прямої MK.
35. Пряма перетинає сторони AB і BC трикутника ABC у точках M і K відповідно. Вершини цього трикутника рівновіддалені від прямої MK. Доведіть, що точки M і K є серединами сторін AB і BC відповідно.
36. Висоти AM і CK трикутника ABC перетинаються у точці H, HK = HM. Доведіть, що трикутник ABC рівнобедрений.
37. Висоти ME і NF трикутника MKN перетинаються у точці O, OM = ON, MF = KE. Доведіть, що трикутник MKN рівносторонній.

Для покращення результатів було проведено попереднє оброблення, зокрема вилучено так звані стоп-символи. Це слова, які не мають смис-

лового навантаження, вони трапляються в кожному тексті: сполучники, частки, прийменники і безліч інших слів. Далі було проведено операцію стемінгу. Вона не є обов'язковою, деякі джерела стверджують, що хороші результати виходять і без неї. Зокрема, якщо набір текстів досить великий, то цей крок можна пропустити. Для стемінгу використовували модифікований алгоритм Портера для російської мови, оскільки варіанта для української немає.

['знайдіт', 'кут', 'трикутник', 'інш', 'кут', 'дорівнюют']
['із', 'кут', 'трикутник', 'раз', 'менш', 'друг', 'кут', 'менш', 'треть', 'знайдіт', 'кут', 'трикутник']
['знайдіт', 'кут', 'трикутник', 'їхн', 'градусн', 'мір', 'віднос']
['знайдіт', 'кут', 'рівностороннь', 'трикутник']
['знайдіт', 'кут', 'рівнобедрен', 'прямокутн', 'трикутник']
['кут', 'осн', 'рівнобедрен', 'трикутник', 'дорівню', 'знайдіт', 'кут', 'вершин', 'трикутник']
['знайдіт', 'кут', 'осн', 'рівнобедрен', 'трикутник', 'кут', 'вершин', 'дорівню']
['знайдіт', 'кут', 'рівнобедрен', 'трикутник', 'кут', 'вершин', 'раз', 'більш', 'кут', 'осн']
['знайдіт', 'кут', 'рівнобедрен', 'трикутник', 'кут', 'осн', 'менш', 'кут', 'вершин']
['знайдіт', 'кут', 'рівнобедрен', 'трикутник', 'із', 'них', 'дорівню', 'скільк', 'розв'язк', 'задач']
['периметр', 'трикутник', 'дорівню', 'см', 'одн', 'сторін', 'дорівнюв', 'см', 'см']
['довжин', 'двох', 'сторін', 'трикутник', 'дорівнюют', 'см', 'см', 'периметр', 'трикутник', 'дорівнюв', 'см', 'см', 'см']
['існ', 'трикутник', 'одн', 'зі', 'сторін', 'як', 'см', 'менш', 'другої', 'см', 'менш', 'третьої', 'периметр', 'дорівню', 'см']
['трикутник', 'відом', 'сторон', 'позначил', 'довільн', 'точк', 'доведіт']
['доведіт', 'відрізок', 'сполуч', 'вершин', 'рівнобедрен', 'трикутник', 'точк', 'лежит', 'осн', 'більш', 'бічн', 'сторон', 'трикутник']
['доведіт', 'відрізок', 'сполуч', 'вершин', 'трикутник', 'точк', 'лежит', 'протилежн', 'сторон', 'більш', 'одн', 'двох', 'інш', 'сторін']
['дві', 'сторон', 'рівнобедрен', 'трикутник', 'дорівнюют', 'см', 'см', 'знайдіт', 'трет', 'сторон', 'трикутник']
['дві', 'сторон', 'рівнобедрен', 'трикутник', 'дорівнюют', 'см', 'см', 'знайдіт', 'довжин', 'основ', 'трикутник']
['осн', 'висот', 'трикутник', 'внутрішнь', 'точк', 'відрізк', 'відом', 'кут', 'кут', 'порівняйт', 'довжин', 'сторін']
['осн', 'висот', 'трикутник', 'внутрішнь', 'точк', 'відрізк', 'відом', 'порівняйт', 'величин', 'кут']
['доведіт', 'трикутник', 'будьяк', 'сторон', 'менш', 'половин', 'периметр']
['трикутник', 'провел', 'бісектрис', 'доведіт']
['сторон', 'трикутник', 'позначил', 'точк', 'відом', 'доведіт']
['із', 'гостр', 'кут', 'прямокутн', 'трикутник', 'дорівню', 'знайдіт', 'друг', 'гостр', 'кут']
['рівнобедрен', 'трикутник', 'проведен', 'висот', 'знайдіт', 'кут', 'кут']
['кут', 'основ', 'рівнобедрен', 'трикутник', 'висот', 'проведен', 'бічної', 'сторон', 'дорівню', 'знайдіт', 'кут', 'дан', 'трикутник']
['сторон', 'кут', 'вершин', 'точц', 'позначил', 'точк', 'точк', 'провел', 'прям', 'перпендикулярн', 'сторін', 'відповідн', 'перетинают', 'точц', 'доведіт', 'промін', 'бісектрис', 'кут']
['доведіт', 'висот', 'рівнобедрен', 'трикутник', 'проведен', 'бічн', 'сторін', 'рівн']
['доведіт', 'дві', 'висот', 'трикутник', 'рівн', 'трикутник', 'рівнобедрен']
['доведіт', 'рівніст', 'прямокутн', 'трикутник', 'катет', 'бісектрис', 'проведен', 'вершин', 'прям', 'кут']
['доведіт', 'рівніст', 'прямокутн', 'трикутник', 'катет', 'висот', 'проведен', 'вершин', 'прям', 'кут']
['доведіт', 'рівніст', 'прямокутн', 'трикутник', 'катет', 'бісектрис', 'проведен', 'вершин', 'прилегл', 'катет', 'гостр', 'кут']
['доведіт', 'рівніст', 'прямокутн', 'трикутник', 'катет', 'медіан', 'проведен', 'друг', 'катет']
['прям', 'перетин', 'сторон', 'трикутник', 'відповідн', 'точк', 'середин', 'сторін', 'доведіт', 'вершин', 'дан', 'трикутник', 'рівновіддален', 'прямої']
['прям', 'перетин', 'сторон', 'трикутник', 'точк', 'відповідн', 'вершин', 'дан', 'трикутник', 'рівновіддален', 'прямої', 'доведіт', 'точк', 'середин', 'сторін', 'відповідн']
['висот', 'трикутник', 'перетинают', 'точц', 'доведіт', 'трикутник', 'рівнобедрен']
['висот', 'трикутник', 'перетинают', 'точц', 'доведіт', 'трикутник', 'рівностороннь']

У результаті було отримано таку матрицю вживаності.

більш [0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
бісектрис [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0.
0. 0. 1. 0. 0. 1. 0. 1. 0. 0. 0. 0. 0.]
бічн [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
вершин [0. 0. 0. 0. 0. 1. 1. 1. 1. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 1. 0. 0. 1. 1. 1. 0. 1. 1. 0. 0.]
висот [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0.
1. 1. 0. 1. 1. 0. 1. 0. 0. 0. 0. 1. 1.]
внутрішнь [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
відом [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 1. 1. 0. 0. 1. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
відповідн [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 1. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0.]
відрізк [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
відрізок [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
гостр [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.
0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]
дан [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0.]
двох [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
дві [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
доведіт [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 1. 0. 0. 0. 0. 1. 1. 1. 0.
0. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
довжин [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
дорівню [0. 0. 0. 0. 0. 1. 1. 0. 0. 1. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.
0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
дорівнюв [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
дорівнюют [1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
друг [0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.
0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]
знайдіт [1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 1.
1. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
катет [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 1. 1. 1. 1. 0. 0. 0. 0.]
кут [1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 1.
1. 1. 1. 0. 0. 1. 1. 1. 0. 0. 0. 0. 0.]
лежит [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
менш [0. 1. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
одн [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
осн [0. 0. 0. 0. 0. 1. 1. 1. 1. 0. 0. 0. 0. 0. 1. 0. 0. 0. 1. 1. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
основ [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.
0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
перетин [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0.]
перетинают [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1.]
периметр [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 1. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
позначил [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0.
0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
порівняйт [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
проведен [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
1. 1. 0. 1. 0. 1. 1. 1. 1. 0. 0. 0. 0.]
провел [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0.
0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
прям [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 1. 0. 0. 1. 1. 0. 0. 1. 1. 0. 0.]
прямокутн [0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.
0. 0. 0. 0. 0. 1. 1. 1. 1. 0. 0. 0. 0.]
прямої [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0.]
раз [0. 1. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
рівн [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
рівнобедрен [0. 0. 0. 0. 1. 1. 1. 1. 1. 1. 0. 0. 0. 0. 1. 0. 1. 1. 0. 0. 0. 0. 0. 0.
1. 1. 0. 1. 1. 0. 0. 0. 0. 0. 0. 1. 0.]
рівновіддален [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0.]
рівніст [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 1. 1. 1. 1. 0. 0. 0. 0.]
середин [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0.]
см [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 1. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
сполуч [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
сторон [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 1. 1. 1. 0. 0. 1. 0. 1. 0.
0. 1. 1. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0.]
сторін [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 1. 0. 0. 1. 0. 0. 1. 0. 0. 0. 0. 0.
0. 0. 1. 1. 0. 0. 0. 0. 0. 1. 1. 0. 0.]
точк [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 1. 0. 0. 1. 1. 0. 0. 1. 0.
0. 0. 1. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0.]
точц [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1.]
трикутник [1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
із [0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
інш [1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Найпоширеніший варіант LSA оснований на використанні розкладання матриці за сингулярними значеннями (SVD, Singular Value Decomposition) [6]. За допомогою SVD-розкладання будь-яка матриця розкладається у безліч ортогональних матриць, лінійна комбінація яких є досить точним наближенням до початкової матриці. Згідно з теоремою про сингулярне розкладання матриць [6], матрицю можна представити як додаток трьох матриць:

$$A = U \Sigma V^t,$$

де Σ – діагональна матриця $m \cdot n$ із додатних елементів, які є сингулярними числами; U – матриця $n \cdot m$, яка складається із лівих сингулярних векторів; V – матриця $n \cdot n$, яка складається із правих сингулярних векторів; V^t – транспонована матриця до V. Таке розкладання має чудову особливість: якщо в матриці Σ залишити тільки k найбільших сингулярних значень, а в матрицях U і V – тільки відповідні цим значенням стовпці,

то добуток матриць Σ, U і V буде найкращим наближенням початкової матриці A до матриці $\tilde{A}$ рангу k:

$$\tilde{A} \approx A = U\Sigma V^t.$$

У роботі було досліджено поведінку алгоритму для різних значень k.

1. Кількість сингулярних векторів і чисел – 3. Порахувавши евклідову відстань між точками, отримали таку дендограму:

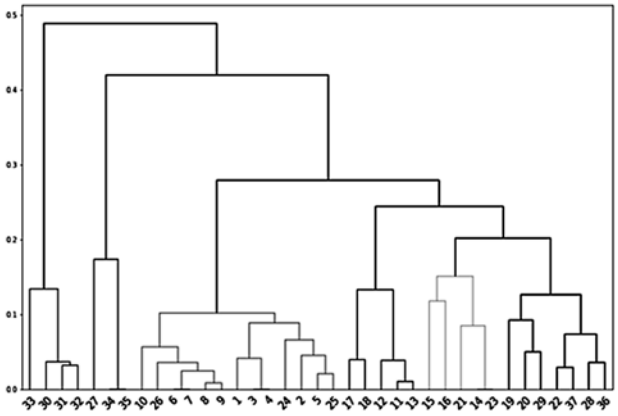

Рис. 1. Дендограма для трьох сингулярних чисел

(1, (30, ‘Доведіть рівність прямокутних трикутників за катетом і бісектрисою, проведеною з вершини прямого кута.\n’))
(1, (31, ‘Доведіть рівність прямокутних трикутників за катетом і висотою, проведеною з вершини прямого кута.\n’))
(1, (32, ‘Доведіть рівність прямокутних трикутників за катетом і бісектрисою, проведеною з вершини прилеглого до цього катета гострого кута.\n’))
(1, (33, ‘Доведіть рівність прямокутних трикутників за катетом і медіаною, проведеною до другого катета.\n’))
(2, (27, ‘На сторонах кута з вершиною в точці B позначили точки A і C так, що AB = BC. Через точки A і C провели прямі, які перпендикулярні до сторін BA і BC відповідно та перетинаються в точці O. Доведіть, що промінь BO — бісектриса кута ABC.\n’))
(2, (34, ‘Пряма перетинає сторони AB і BC трикутника ABC відповідно в точках M і K, які є серединами цих сторін. Доведіть, що вершини цього трикутника рівновіддалені від прямої MK.\n’))
(2, (35, ‘Пряма перетинає сторони AB і BC трикутника ABC у точках M і K відповідно. Вершини цього трикутника рівновіддалені від прямої MK. Доведіть, що точки M і K є серединами сторін AB і BC відповідно.\n’))
(3, (1, ‘Знайдіть кут трикутника, якщо два інші його кути дорівнюють 35° і 96°.\n’))
(3, (2, ‘Один із кутів трикутника в 3 рази менший від другого кута та на 35° менший від третього. Знайдіть кути трикутника. \n’))
(3, (3, ‘Знайдіть кути трикутника, якщо їхні градусні міри відносяться як 2 : 3 : 7.\n’))
(3, (4, ‘Знайдіть кути рівностороннього трикутника.\n’))
(3, (5, ‘Знайдіть кути рівнобедреного прямокутного трикутника.\n’))
(3, (6, ‘Кут при основі рівнобедреного трикутника дорівнює 63°. Знайдіть кут при вершині цього трикутника.\n’))
(3, (7, ‘Знайдіть кути при основі рівнобедреного трикутника, якщо кут при вершині дорівнює 104°.\n’))
(3, (8, ‘Знайдіть кути рівнобедреного трикутника, якщо кут при вершині в 4 рази більший за кут при основі.\n’))
(3, (9, ‘Знайдіть кути рівнобедреного трикутника, якщо кут при основі на 48° менший від кута при вершині.\n’))
(3, (10, ‘Знайдіть кути рівнобедреного трикутника, якщо один із них дорівнює: 1) 110°; 2) 50°. Скільки розв’язків має задача?\n’))
(3, (24, ‘Один із гострих кутів прямокутного трикутника дорівнює 43°. Знайдіть другий гострий кут.\n’))
(3, (25, ‘У рівнобедреному трикутнику ABC (AB = BC) проведено висоту AH. Знайдіть кут CAH, якщо кут B = 76°.\n’))
(3, (26, ‘Кут між основою рівнобедреного трикутника та висотою, проведеною до бічної сторони, дорівнює 19°. Знайдіть кути цього трикутника.\n’))
(4, (11, ‘Периметр трикутника дорівнює 30 см. Чи може одна з його сторін дорівнювати: 1) 20 см; 2) 15 см?\n’))
(4, (12, ‘Довжини двох сторін трикутника дорівнюють 7 см і 9 см. Чи може периметр цього трикутника дорівнювати: 1) 20 см; 2) 32 см; 3) 18 см?\n’))
(4, (13, ‘Чи існує трикутник, одна зі сторін якого на 2 см менша від другої та на 6 см менша від третьої, а периметр дорівнює 20 см?\n’))
(4, (17, ‘Дві сторони рівнобедреного трикутника дорівнюють 9 см і 20 см. Знайдіть третю сторону трикутника.\n’))
(4, (18, ‘Дві сторони рівнобедреного трикутника дорівнюють 8 см і 16 см. Знайдіть довжину основи цього трикутника.\n’))
(5, (19, ‘Основа D висоти AD трикутника ABC є внутрішньою точкою відрізка BC. Відомо, що кут BAD > кута CAD. Порівняйте довжини сторін AB і AC.\n’))

(5, (20, 'Основа D висоти AD трикутника ABC є внутрішньою точкою відрізка BC. Відомо, що AB > AC. Порівняйте величини кутів BAD і CAD.\n'))
(5, (22, 'У трикутнику ABC провели бісектрису BD. Доведіть, що AB > AD і BC > CD.\n'))
(5, (28, 'Доведіть, що висоти рівнобедреного трикутника, проведені до його бічних сторін, є рівними.\n'))
(5, (29, 'Доведіть, що коли дві висоти трикутника рівні, то цей трикутник є рівнобедреним.\n'))
(5, (36, 'Висоти AM і CK трикутника ABC перетинаються в точці H, HK = HM. Доведіть, що трикутник ABC рівнобедрений.\n'))
(5, (37, 'Висоти ME і NF трикутника MKN перетинаються в точці O, OM = ON, MF = KE. Доведіть, що трикутник MKN рівносторонній.'))
(6, (14, 'У трикутнику ABC відомо, що AC > 90°. На стороні BC позначили довільну точку B. Доведіть, що AB > AC.\n'))
(6, (15, 'Доведіть, що відрізок, який сполучає вершину рівнобедреного трикутника з точкою, яка лежить на його основі, не більший за бічну сторону трикутника.\n'))
(6, (16, 'Доведіть, що відрізок, який сполучає вершину трикутника з точкою, яка лежить на протилежній стороні, не більший хоча б за одну з двох інших сторін.\n'))
(6, (21, 'Доведіть, що в трикутнику будь-яка сторона менша від половини периметра.\n'))
(6, (23, 'На стороні AC трикутника ABC позначили точку D. Відомо, що AD > BD. Доведіть, що AC > BC.\n'))

2. Кількість сингулярних векторів і чисел – 5.

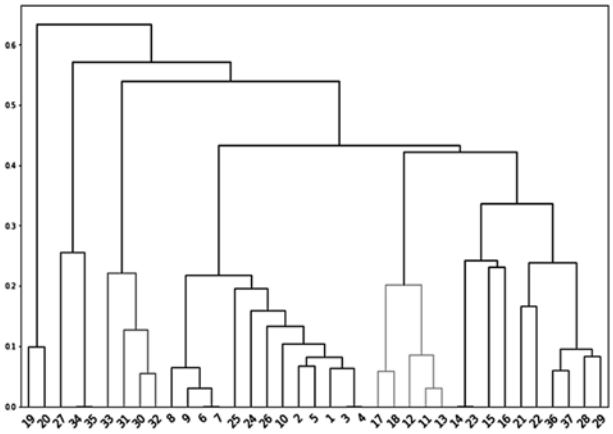

Рис. 2. Дендограма для п'яти сингулярних чисел

3. Кількість сингулярних векторів і чисел – 10.

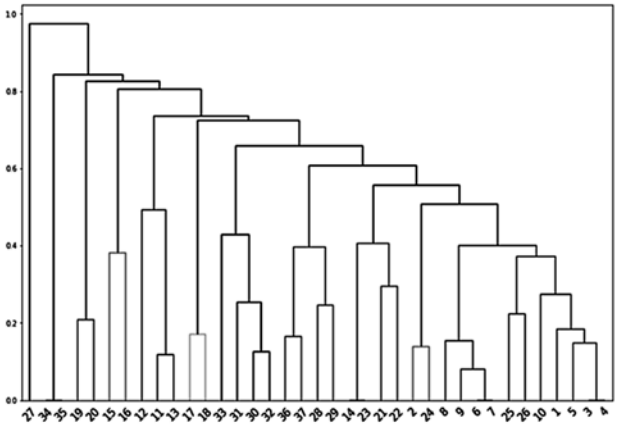

Рис. 3. Дендограма для десяти сингулярних чисел

Висновки

У роботі було досліджено можливість застосування латентно-семантичного аналізу для кластеризації документів. У результаті дослідження зроблено висновок, що, незважаючи на трудомісткість і непрозорість, ЛСА можна успішно застосовуватися для розв'язку подібного типу завдань. Важливою перевагою цього методу є можливість знаходити неочевидні зв'язки між документами.

Один із найбільших недоліків – дуже велика обчислювальна складність методу.

Практичним результатом роботи стала реалізація засобу для класифікації текстів, досліджено його роботу на прикладі кластеризації задач із геометрії.

Список літератури

1. Борозенний С. О. Про особливості використання алгоритму LSA / С. О. Борозенний, Г. В. Мельник // Матеріали Дванадцятої міжнародної науково-практичної конференції «Теоретичні та прикладні аспекти побудови програмних систем» (TAAPSD'2011), 23–26 листопада 2011 р., Ялта, Україна. – С. 16–18.
2. Борозенний С. О. Пошук документів на основі алгоритму LSA [Електронний ресурс] / С. О. Борозенний // Perspective innovations in science, education, production and transport '2014. – Режим доступу: https://www.sworld.com.ua/konfer37/706.pdf.
3. Мерзляк А. Г. Геометрія. Пропедевтика поглибленого вивчення : навч. посіб. для 7 кл. з поглибленим вивченням математики / А. Г. Мерзляк, В. Б. Полонський, М. С. Якір. – Харків : Гімназія, 2015. – 192 с.
4. Borko H. Automatic Document Classification / Harold Borko, Myrna Bernick // Journal of the ACM. – 1963. – Vol. 10, Issue 2. – Pp. 151–162.
5. Foltz P. W. Latent semantic analysis for text-based research / Peter W. Foltz // Behavior Research Methods, Instruments&Computers. – 1996. – Vol. 28 (2). – Pp. 197–202.
6. Golub G. H. Chapter Singular Value Decomposition and Least Squares Solutions / G. H. Golub, C. Reinsch // Linear Algebra. Handbook for Automatic Computation / ed. F. L. Bauer. – Springer, Berlin, Heidelberg, 1971. – Vol. 2.

References

Borko, H., & Bernick, M. (1963). Automatic Document Classification. *Journal of the ACM*, *10* (2), 151–162.

Borozennyi, S. (2014). Poshuk dokumentiv na osnovi alhorytmu LSA. In *perspective innovations in science, education, production and transport*. Retrieved from https://www.sworld.com.ua/konfer37/706.pdf [in Ukrainian].

Borozennyi, S., & Melnyk, H. (2011). Pro osoblyvosti vykorystannya alhorytmu LSA. In *Materialy Dvanadtsyatoyi mizhnarodnoyi naukovo-praktychnoyi konferentsiyi "Teoretychni ta prykladni aspekty pobudovy prohramnykh system". (TAAPSD'2011)*, 23–26 lystopada 2011 r., Yalta, Ukraina (pp. 16–18) [in Ukrainian].

Foltz, P. W. (1996). Latent semantic analysis for text-based research. In *Behavior Research Methods, Instruments&Computers*, *28* (2), 197–202.

Golub, G. H., & Reinsch, C. (1971). Singular Value Decomposition and Least Squares Solutions. In F.L. Bauer (ed.), *Linear Algebra. Handbook for Automatic Computation* (Vol. 2). Springer, Berlin, Heidelberg.

Merzlyak, V., Polons'kyy, M., & Yakir, M. S. (2015). *Heometriya. Propedevtyka pohlyblenoho vyvchennya*. Kharkiv: Himnaziia [in Ukrainian].

O. Zhezherun, S. Borozennyi, M. Niverovskyi

USING THE LSA ALGORITHM FOR CLUSTERING GEOMETRY PROBLEMS

Currently, there are a huge number of clustering algorithms. The basic idea of most of them is to combine identical sequences into one class or cluster based on similarity. As a rule, the choice of algorithm is determined by the task. As for textual data, the compared components are sequences of words and their attributes (for example, the weight of a word in the text, the type of the named entity, tonality, etc.). Thus, the texts are first transformed into vectors, which are used for various types of manipulation. At the same time, as a rule, there are a number of problems connected with: selection of primary clusters, the dependence of the quality of clustering on the length of the text, determining the total number of clusters, etc.

But the most difficult problem is the lack of connection between similar texts, which use different vocabulary. In such cases, the association should take place not only on the basis of similarity, but also on the basis of semantic contiguity or associativity.

One of the methods that allows to solve such problems is Latent semantic analysis (LSA). LSA is a method of information processing that analyzes a set of documents and finds the terms that occur there, and on this basis identifies the characteristic factors, topics that characterize the content of the document.

Define the following types of correlation:

"Word-word";

"Word-paragraph";

"Paragraph-paragraph".

These are the three types that a person thinks, comparing parts of the text with the content. LSA technology takes into account not only the frequency of the text use, but also latent (deep) connections.

The first article on the Automatic Document Classification [4] was published in the Journal of the ACM in early 1963, and was the first to describe the method of factor analysis as a means of finding information. Factor analysis is a method that determines the relationship between the values of variables.

In this paper, the possibility of using latent-semantic analysis for clustering of texts (geometry problems) has been investigated, for which an algorithm and the necessary software have been developed.

Keywords: LSA, LSI, SVD, clustering.

Матеріал надійшов 11.06.2020