БОРОЗЕННИЙ С.О. МЕЛЬНИК Г.В.

Київський національний університет «Києво-Могилянська академія»
sborozenny@gmail.com, gansuk@gmail.com

ПОШУК ДОКУМЕНТІВ НА ОСНОВІ АЛГОРИТМУ LSA

Одним з перспективних методів, що дозволяє отримувати дані про значення наведеного тексту, є метод латентного семантичного аналізу (ЛСА).

ЛСА дозволяє виявити значення слів з урахуванням контексту їх використання шляхом обробки великого набору текстів. Принцип дії методу полягає в тому, що порівняння всіх контекстів, у яких слова або групи слів вживаються, і контекстів, у яких вони не вживаються, дозволяє зробити висновок про ступені близькості змісту цих слів чи груп слів.

Вперше метод ЛСА був описаний в роботі "An Introduction to Latent Semantic Analysis" Landauer, T. K., Foltz, P., and Laham, D. у 1998 році [1J і потім розвинений в працях Scott Deerwester, Susan Duniais, George Furnas.

На даний момент лідером в області застосування ЛСА є компанія Pearson Knowledge Technologies. [2] їхні комерційні продукти дозволяють впевнитись в ефективності методу ЛСА. Проте конкретні алгоритми реалізації цього методу не опубліковані, оскільки є комерційною таємницею, тому дуже важливою частиною цієї роботи є створення програмного забезпечення, яке дозволить відтворити і перевірити результати роботи алгоритму ЛСА.

І Іредставлення слова і абзацу за допомогою алгоритму ЛСА здебільшого моделює сприйняття тексту людиною [3]. Наприклад з його допомогою можна оцінити есе на відповідність темі або порівняти зміст уривків тексту.

ЛСА можна розглядати вдвох аспектах:

- Як практичний прийом для отримання приблизних оцінок контекстного зв'язку слів у великих фрагменту по змісту, або оцінок змістових кореляцій між словом і набором слів
- Як комп'ютерну моделі, отримання і використовування знань людиною яка читає текст

В якості практичного методу, що характеризує значення слова ЛСА дозволяє визначити кореляції типу «слово-слово», «слово-уривок» і «уривок-уривок». Ці кореляції моделюють механізм мислення людини, співставляючого частини тексту за змістом. Досвід показує наявність зв'язку між результатами роботи методу ЛСА і людським сприйняттям. Важливо зазначити, що результати, що даються методом ЛСА, залежать не тільки від частотності використання слів *в* уривках. Метод грунтується на виявлення більш глибоких ("латентних") зв'язків і, таким чином, краще моделює людське сприйняття тексту ніж прості методи, засновані на частотності вживання слів [4].

Слід зазначити, що у методу ЛСА існують деякі обмеження. У ньому не використовується інформація про порядок слів, отже, метод не враховує синтаксичні відношення, логіку або морфологію. Незважаючи на це, результати методу досить вірогідно відображають смислові кореляції між словами та уривками [1].

С дві основні відмінності методу ЛСА від інших пошукових методів

- В якості початкових даних ЛСА використовує частоту зустрічання кожного слова окремо в тексті, а не частоту зустрічання фрази
- метод збирає данні не про попарно сумісне використовування слів, а про використання множини слів у великому масиві текстів

Таким чином, метод розглядає вплив вибору, а не порядку слів на сенс уривка. Можна сказати, що ЛСА представляє значення слова як середнє значень уривків, в яких воно зустрічається, а значення уривка - як середнє значень всіх слів, що складають уривок.

Кластеризація тексту

Латентно-семантичний аналіз застосовується при вирішенні завдання що часто зустрічається — структурування та аналізу уривків тексту. Для вирішення цього завдання погрібно зробити розбиття текстових масивів на систему (можливо ієрархічних) підмножин, помічених якимись то їх змістовими описувачами (автоматичну кластеризацію).

У методах кластеризації, що ґрунтуються на метриці близькості, документ представляється у вигляді багагомірного вектора. Підходи

до формування вектора, що представляє документ, можуть істотно різнитися. У найпростішому випадку кожен елемент вектора відповідає наявності в тексті однієї з словоформ, що зустрічається в] розглянутому наборі текстів. Основною проблемою методів, I заснованих на такий матриці, виявляється занадто велика розмірність | простору слів, велика частина яких є надлишковими і навіть шкідливими. Наприклад, при кластеризації текстів в даній предметній області терміни, не відносяться до цієї області, можуть маскувати схожість між документами. Сингулярний розклад, що використовується в методі ЛСА, дозволяє зменшити розмірність матриці, зменшити обсяг обчислень і позбутися від надлишкових даних.

Латентне семантичне індексування

Результат, що видається звичайною пошуковою машиною, заснований на безпосередньому порівнянні запиту та документа. Користувачеві видаються тільки ті документи, в яких є одне або кілька слів із запиту. Результат, отриманий таким чином, часто буває неточний - він може містити «зайві» документи (не мають відношення до теми запиту користувача) і не містити тематичних документів, в яких відсутні слова з запиту. Для отримання кращих результатів необхідно враховувати змістову близькість запиту користувача і видавати документів. Латентне семантичне індексування (LSI) дозволяє поліпшити результат пошуку. При відборі документів враховується не тільки наявність у них слів із запиту, а й змістову близькість до запиту (виміряна за використанням методу ЛСА). Семантичний пошук знаходить вирази, які відповідають пошуковим запитам, користувача і видає результати, які близькі їм за змістом. Сторінка з результатом пошуку може не містити тих слів, які ви шукали, але слова на сторінці будуть близькі до вашого пошукового запиту.

Цей підхід використовується в пошуковій машині Google. Наприклад, запит «appie computer» буде інтерпретовано як належить до комп'ютерної компанії Apple, тому серед результатів будуть документи, що описують технології цієї компанії (навіть, якщо ці документи не містять слів «apple» або «computer»).

Матриця вживаності

ЛСА використовує матрицю, яка описує вживаність слів у текстах. Нехай стовпці матриці будуть означати тексти, а рядки — слова. Годі елементи матриці являють собою кількість зустрічань конкретного слова в даному тексті. Такий підхід стандартний для семантичних моделей

Отже матриця відображає зв'язки між словами та текстами. При аналізі текстів виникає дві проблеми

- синонімія — коли різні слова позначають одне й теж саме поняття
- полісемія — коли одне і теж саме слово або декілька слів можуть означати різні поняття

Пониження рангу

Після отримання матриці вживаності необхідно понизити її ранг. На це с декілька причин:

- За рахунок великої розмірності матриці обчислення в ній часто можуть бути неможливими
- Початкова матриця містить "шуми", тобто випадкове потрапляння слова в текст, при цьому це слово не впливає на зміст. Пониження рангу дозволяє частково позбавитися таких шумів
- Початкова матриця досить розріджена, вона враховує тільки слова які зустрічаються в тексті, але не слова пов"язані з текстом (через синонімію)

Також необхідно вводити певні вагові коефіцієнти для слів. Наприклад при пошуку за словами (машина, квіти, вантажівка) визначити вагові коефіцієнти цим ловам

{(машина), (вантажівка), (квіти)} —»{(1.563*машина+0.661 ""вантажівка), (квіти)} оскільки всі вантажівки є машинами.

Ця операція дозволяє зменшити вплив синонімії, оскільки вона "зливає" слова близькі за змістом. Також зменшується вплив полісемії, оскільки якщо слово має "правильне значення", то воно буде "злите" зі словами близькими по змісту, а якщо ні, то буде відкинуте.

На основі алгоритму ЛСА можна реалізувати два варіанти пошуку. Перший, більш класичний, гобто користувач вводить пошуковий запит, зазвичай це певні ключеві слова або фрази, і як результат

пошуку отримує документи відсортовані в порядку відповідності *до* запиту. Другий варіант реалізує пошук схожих текстів. На вхід користувач подає документ, а в результаті отримує документи які є схожими на вхідний.

Пошук за ключовими словами

В даному випадку пошук відбувається серед файлів які знаходяться у визначеному каталозі. З цих файлів будується матриця вживаності, але перед цим слова оброблюються, щоб зменшити розмірність матриці. Усі слова перевіряються на належність до так званого списку "стоп-слів" , тобто найбільш вживаних слів які зустрічаються в усіх можливих текстах (не тільки серед тих, в яких буде здійснювати пошук), оскільки вони жодним чином не будуть надавати документу відмінності від інших. Також кожне слово обробляється стеммером Портера. В результаті після обробки до матриці вживаності потрапляють словоформи без суфіксів та закінчень, що дає змогу не враховувати рід і відмінок слів.

Пошук схожих документів.

Такий пошук відбувається за схожою схемою. Тільки на вхід користувач подає документ. Після побудови матриці вживаності та її сингулярного розкладу, критерієм схожості документів є рангова кореляція Спірмена.

В подальшому роботу можна розвинути в напрямках багатомовності, підключення словників для перекладу пошукових запитів та знаходження необхідних документів іншими мовами. Створення криерію мінімальної відповідності запиту. Створення «альтернативного» пошукового запиту при відсутності результатів які задовольняють критерій відповідності запиту.

ЛІТЕРАТУРА

1. Landauer, T. K., Foltz, P., and Laham, D. 1998. An Introduction to Latent Semantic Analysis. DiscourseProcesses, 25: 259-284.
2. Веб-сайт Pearson Knowledge Technologies – http://www.k-a-t.com
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.A. 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41: 391-407.
4. Foltz, P. W. 1996. Latent Semantic Analysis for text-based research. Behavior Research Methods, Instruments and Computers. 28, 197-202.