ВИСНОВКИ

Наведено головні характеристики кластерних систем на основі розподілених компонентів та розглянуто основні методи їх оптимізації. Варто відзначити, що цей підхід не унікальний, тому що кожна задача має свої параметри та алгоритми виконання, тому необхідно індивідуально налаштовувати систему під конкретні потреби.

ЛІТЕРАТУРА

1. Оптимізація продуктивності обчислювального кластера на базі розподілених слабкозв'язаних компонентів / О. О. Судаков, Ю. В. Бойко, Т. В. Ничипорук, Т. П. Короткова // Математичні машини і системи. – 2004. – (ISSN 1028-9763; кн. 4). – С. 57–65.

Андрющенко Д.Ю.,
Криворізький національний університет
Сенько А.О.
асистент, Криворізький національний університет

ФУНКЦІЇ РАНЖУВАННЯ КОМЕРЦІЙНИХ ПОШУКОВИХ СИСТЕМ ДЛЯ ОБРОБКИ ТЕКСТОВИХ ЗАПИТІВ

Розглянуто класичні способи реалізації пошукових системи в Інтернеті. Проаналізовані недоліки цих систем за кількома кількісними та якісними критеріями, порівнюючи компроміси прийняття нейронної архітектури з успішними та зрілими традиційними методами пошуку інформації.

Ключові слова: пошукова система, машинне навчання, information retrieval

Завданням фундаментального IR (information retrieval) є завдання спеціального пошуку, в якому система повинна відповідати ранжованим списком документів, що є актуальним для інформаційної потреби користувача, зазвичай повідомленої системі як одноразовий запит, наприклад: "готелі в Нью-Йорку".

Один із способів: ранжувати набір документів для обчислення оцінки релевантності для кожного документа d даного запиту q. Отже, завдання зводиться до представлення та задачі відповідності, тобто, як ви представляєте запит і документ таким

чином, що функція posterior matching повертає найвищі значення для найбільш релевантних документів для даного запиту q. Цей процес оцінювання релевантності можна в цілому виражати:

$$rel(q, d) = f(\Phi q(q), \Phi d(d)), \quad (1)$$

де Φq і Φd є функціями відображення представлення, тобто вони обчислюють представлення запиту і документа, відповідно, f є функцією узгодження, заснованої на взаємодії між запитами і поданнями документів.

З досвіду відомо, що функція ранжування для будь-якої комерційної пошукової системи використовує комбінацію незалежних від запиту сигналів і залежних від запиту сигналів, тобто може бути щось подібне до $rel(d, q) = \psi(d) \times f(\Phi q(q),\Phi d(d))$. Незалежні від запиту сигнали ψ(d) побудовані на основі алгоритмів авторитету / популярності, таких як PageRank[1] або оцінка виявлення спаму[3]. Ця робота залишає ці сигнали в стороні, оскільки ми розуміємо, що вони можуть бути досліджені ортогонально для обчислення залежних від запиту сигналів.

Далі розглянуті деякі класичні прийоми для текстового IR і найсучасніші підходи.

Одне з перших IR-рішень було запропоновано в 1950-х роках, булева модель: документ і запит представлені просто булевим вектором, який вказує, які слова лексики вони містять. Функція узгодження f є просто булевою операцією, що вказує, чи є слова у запиті присутніми в документі.

Це може здатися занадто простим, але це є основою широко і успішно прийнятого рейтингу на основі TF-IDF, який можна розглядати як розширену булеву модель, що використовує модель векторного простору, щоб надати вагам слова відповідно до частоти її рівня колекції (IDF) і частоту документообігу (TF), запропонована в 1970-х рр.[2]

Іншим недоліком класичних методів є те, що вони припускають: терміновий витяг отримує достатньо повний список документів-кандидатів. Проте, через добре відомий розрив словників, це припущення не є цілком точним[4]. Документ може бути доречним для запиту навіть без точних відповідних термінів, через використання множини, сполучення, синонімів або семантичних відносин, щоб навести декілька. Розширення документів і запитів допомага-

ють вирішити цю проблему, і вони самі по собі є повноцінними полями досліджень. Іншими словами, використання точного пошуку відповідності термінів призводить до втрати зворотнього зв'язку, оскільки документ, що містить тільки «штучний інтелект», не буде вилучений, наприклад, для запиту «інтелектуальні машини».

ВИСНОВКИ

Отже, у традиційних методах IR функції Фq і Фd обчислюють векторне представлення у вигляді TF-IDF, а функція узгодження f пари обчислюється за допомогою рівняння закритої форми, наприклад, BM25 враховуючи перекриваються терміни. Більш точні представлення можуть розглядати більш довгі послідовності слів, замість окремих слів, відстань між термінами, синоніми та некомпозиційні сполуки. Крім того, часто документ може складатися з декількох компонентів, таких як заголовок, основний текст, метатеги та якірні тексти, виконувати побудову функцій з цих джерел, а також рішення про те, як зважити їх у кінцевій функції узгодження яка не є тривіальною.

ЛІТЕРАТУРА

1. L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: bringing order to the web, Technical Report 1999-66, Stanford InfoLab, 1999.
2. K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1972) 11–21.
3. G.V. Cormack, M.D. Smucker, C.L. Clarke, Efficient and effective spam filtering and re-ranking for large web datasets, Inf. Retr. 14 (5) (2011) 441–465, doi:10.1007/s10791-011-9162-z.
4. L. Boytsov, D. Novak, Y. Malkov, E. Nyberg, Off the beaten path: Let's replace term-based retrieval with k-nn search, in: Proceedings of the Twenty-Fifth ACM International on Conference on Information and Knowledge Management, 2016, pp. 1099–1108, doi:10.1145/2983323.2983815.