УДК 004.85:025.042

ВЕКТОРНІ МОДЕЛІ ІНФОРМАЦІЙНОГО ПОШУКУ

Никоненко В.І.
Науковий керівник – канд. пед. наук, доц. Ситникова Ю.В.
Харківський національний університет радіоелектроніки, каф. ПМ,
м. Харків, Україна
e-mail: valeriia.nykonenko@nure.ua

The post-industrial development of society has made mathematics as the foundation for creating the most advanced technologies that would make our lives easier. In this regard, special information retrieval models have been invented. Most of them are based on mathematical knowledge, so they are considered the most convenient to use. However, there are errors in everything that only complicate the process of any activity. Some factors that cause the inefficiency of the information retrieval vector model have been studied. Problems affect the result have been analyzed. The possibilities of eliminating certain shortcomings are clarified. Prospects for further research of the model under consideration and positive aspects of the use of mathematical tools in information retrieval systems are determined.

Проблема знаходження зв’язку між математикою та повсякденним життям залишається в центрі уваги, оскільки світ розвивається завдяки новітнім технологіям. Поєднання точних наук та людської фантазії дозволяє пізнавати та створювати нове. Таким чином, зараз ми можемо знаходити відповіді на власні питання, використовуючи мережу Інтернет. Для знаходження інформації ми здійснюємо її пошук та обираємо найбільш відповідну до нашого запиту. Як відомо існує декілька моделей для виконання ефективного пошуку.

Метою даної роботи є дослідження алгебраїчного підходу до інформаційного пошуку. Об’єктом вивчення в роботі стала векторна модель (Vector Space Model) та методи, що використовуються в цьому просторі.

Документ або запит подаються у вигляді векторів, і головне завдання – знайти найбільшу релевантність [1]. Більшість авторів (зокрема, Gerard Salton, Christopher D. Manning, Peter Norving, Stuart J. Russell) вказують на косинусну подібність та частотно-інверсну частоту, тож їх можна вважати найпоширенішими методами у векторній моделі. Як показало дослідження, є ще декілька технік, які можуть покращити аналіз на відповідність.

Зазначимо, що векторна модель може використати значну кількість часу на обробку та представлення документа, де багато термінів [2], мовою математики, – це велика розмірність. Дану проблему можна вирішити завдяки латентно (прихованого) семантичного індексування (LSI) – метод, про який найменше згадують дослідники та науковці. Проте він залишається важливим методом, особливо для задач тематичного моделювання та у випадках, коли важлива ефективність та інтерпретованість результатів.

LSI шукає приховані закономірності між словами в самому документі, що призводить до зменшення набору термінів [3]. Завдяки фільтруванню результат стає точним, а обробка ефективнішою, оскільки розмірність стає меншою. Зміна розміру обумовлена не тільки латентно семантичним індексуванням, а також використання розкладання сингулярного значення (SVD), яке, на нашу думку, є суто математичною технікою [3]. VSM важко працювати з синонімами та багатозначними словами, оскільки вона вважає кожне слово незалежною частинкою, не враховуючи можливих зв'язків [2]. Вищезазначений нами метод шукає об'єкти зі схожими характеристиками, намагається зрозуміти концепцію та залишає суттєве [3].

Як було нами досліджено, результати попередньої дії можуть бути використані для нового бачення контексту, що допоможе з подальшим аналізом та швидшим знаходженням відповіді на поставлений запит. Така техніка як аналіз формальної концепції (Formal concept analysis) базується на побудові концептуальних решіток, в яких найважливіші атрибути обираються з ранжованого списку, який в свою чергу був створений за допомогою LSI [1–3]. Латентно семантичне індексування часто використовують разом з іншими методами для інформаційного пошуку через його ефективність. Розмір вектору зменшується через обробку слів, які мають приблизно однакове значення. Тож, виділення головного сенсу є основою для побудови математичної моделі документа або запита.

Підсумовуючи результати дослідження можна зазначити, що LSI – необхідна техніка для семантичного розуміння між словами в документі. Даний метод містить математичну основу, а саме SVD та косинус подібності, можна стверджувати, що у VSM справді алгебраїчна структура, яка допомагає знайти релевантні результати [3].

Перспективою нашого дослідження є детальне вивчення векторної моделі інформаційного пошуку, знаходження проблемних зон застосування й обмежень, які спричиняють неефективність роботи та відбір правильного підходу до вирішення завдань.

Список використаних джерел:

1. Kant Singh V., Kumar Singh V. Vector space model: An Information retrieval system. Int. J. Adv. Engg. Res. Studies. 2015. Vol. 4, No 2. P. 141–143. URL: https://www.researchgate.net/publication/362060638.

2. Van Otten N. Vector space model made simple with examples & tutorial in python. Spot Intelligence. URL: https://spotintelligence.com/2023/09/07/vector-space-model/ (date of access: 07.02.2024).

3. D. Lan, A. Tian, Y. Wang, Y. Li. An overview of the principle, algorithm improvement and application based on the theory of latent semantic indexing. Academic Journal of Computing & Information Science. 2021. Vol. 4, No. 5. P. 71–75. DOI: 10.25236/AJCIS.2021.040510