УДК 004.89:004.912:378.147

Савосін В. С., здобувач вищої освіти,
Штовба С. Д., д-р техн. наук, професор,
професор кафедри інформаційних технологій,
Донецький національний університет імені Василя Стуса

АЛГОРИТМИ СЕМАНТИЧНОГО АНАЛІЗУ ДЛЯ ПОКРАЩЕННЯ ТОЧНОСТІ РЕКОМЕНДАЦІЙ НАУКОВИХ ТВОРІВ НА ОСНОВІ КЛЮЧОВИХ СЛІВ

Анотація. У роботі досліджуються сучасні алгоритми семантичного аналізу для покращення точності рекомендацій наукових творів. Представлено порівняльний аналіз алгоритмів TF-IDF, Word2Vec, та BERT, а також їх ефективність у підборі релевантної наукової літератури.

Ключові слова: семантичний аналіз, рекомендаційна система, наукові твори, ключові слова, TF-IDF, Word2Vec, BERT.

Вступ. З кожним роком обсяг наукових публікацій стрімко зростає, що ускладнює пошук релевантних матеріалів. Традиційні пошукові системи використовують точний збіг ключових слів, що часто не відображає повну картину наукової спорідненості робіт. Тому виникає потреба у розробці більш точних інструментів, які враховують семантичну схожість між документами.

Мета цієї роботи полягає в аналізі та розробці рекомендаційної системи для наукових творів на основі сучасних алгоритмів семантичного аналізу, зокрема TF-IDF, Word2Vec та BERT.

Основний текст. Один із найпоширеніших методів аналізу тексту – це TF-IDF (Term Frequency-Inverse Document Frequency). Цей метод розглядає частоту вживання слова в документі відносно його загальної частоти в корпусі [1]. Хоча TF-IDF дає змогу виявити важливі ключові слова, він не враховує контекст, що є ключовим для точності рекомендацій.

Більш досконалий підхід представляє алгоритм Word2Vec, який моделює семантичні зв’язки між словами, перетворюючи їх у вектори в багатовимірному просторі [2]. Такий підхід дає змогу обчислювати подібність між словами на основі їх контексту. Наприклад, слова «алгоритм» та «модель» будуть розташовані ближче у векторному просторі, якщо часто зустрічаються в схожих контекстах. В експериментах було показано, що Word2Vec на 12–15 % покращує точність рекомендацій, порівняно з TF-IDF [3].

Найбільш інноваційним підходом є використання BERT (Bidirectional Encoder Representations from Transformers), моделі, що базується на технології трансформерів. BERT аналізує не лише окремі слова, але й їх оточення з обох сторін, що дає можливість краще враховувати контекст [4]. Порівняно з TF-IDF та Word2Vec, BERT демонструє найбільш високі показники точності, оскільки може враховувати багатозначність слів і складні мовні структури. У дослідженні використання BERT дало приріст точності рекомендацій на 20 % [5].

Висновки. Кожен алгоритм був оцінений за метриками точності, повноти та F-міри. Найкращі результати були досягнуті алгоритмом BERT, що дає змогу робити більш релевантні рекомендації завдяки врахуванню контексту ключових слів.

Алгоритми семантичного аналізу, як-от Word2Vec і BERT, показують значні переваги у підвищенні точності рекомендацій наукових творів, порівняно з традиційними методами, як-от TF-IDF. У перспективі ці методи можуть бути інтегровані з іншими технологіями, зокрема з мережевими підходами для покращення рекомендацій, що враховують не лише семантику тексту, але й інші метадані, як-от цитування та авторські зв'язки. Дослідження продовжуватиметься в напрямі гібридних моделей рекомендацій для наукових творів.

Список використаних джерел

1. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press. Cambridge, England. 2009. 544 p.

2. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781. 2013. DOI: 10.48550/arXiv.1301.3781.

3. Goldberg Y., Levy O. Word2Vec Explained: Deriving Mikolov's Negative-Sampling Word-Embedding Method. arXiv:1402.3722. 2014. DOI: 10.48550/arXiv.1402.3722.

4. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M. W. Chang, K. Lee, K. Toutanova. arXiv:1810.04805. 2018. DOI: 10.48550/arXiv. 1810.04805.

5. Vaswani A., Shazeer N., Parmar N. Attention is All You Need. *Advances in Neural Information Processing Systems.* 2017. P. 5998–6008.