УДК 004.91

Галкіна Р.І., Мазурець О.В.

Хмельницький національний університет, Україна

ВИКОРИСТАННЯ МЕТОДУ BM25 ДЛЯ АВТОМАТИЗОВАНОГО ПОШУКУ КЛЮЧОВИХ ТЕРМІНІВ

Halkina R.I., Mazurets A.V.

USING THE BM25-METHOD FOR AUTOMATED SEARCHING OF KEY TERMS

Семантичний аналіз тексту є етапом у послідовності дій алгоритму автоматичного розуміння тексту, що полягає у виділенні семантичних відношень і формуванні семантичного представлення. Результати семантичного аналізу можуть бути застосовані для рішення задач у таких областях, як системи автоматичного перекладу, пошукові системи, психіатрія, політологія, торгівля, філологія тощо.

Для автоматизованого пошуку ключових слів використовуються різноманітні методи аналізу текстів, серед яких найбільш відомими є частотна оцінка TF, оцінка TF-IDF, дисперсійна оцінка DE, оцінка ранжування BM тощо. Частотна оцінка визначається відношенням числа входження деякого слова до загальної кількості слів документу. Оцінка TF-IDF є статистичною мірою, що використовується для оцінки важливості слова в контексті документу, що є частиною колекції документів або корпусу, причому складова IDF зменшує вагу широковживаних слів.

Оцінка ранжування базується на імовірнісній моделі й використовується відомими пошуковими системами для впорядкування документів по їх релевантності даному пошуковому

запиту. Зважаючи на те, що оцінка ранжування BM (Best Match) на сучасному етапі широко використовується як метод пошуку ключових слів при вирішення різноманітних задач, визначено актуальним її використання в прикладних інформаційних технологіях, які включають етап визначення множин ключових термінів.

Метод Best Match має багато модифікацій. Серед них BM11, BM15, BM25 та BM25F. Відмінність BM11 та BM15 від більш популярного BM25 полягає у екстремальних значеннях коефіцієнта *b* (при BM11 *b*=1, при BM15 *b*=0).

Okapi BM25 носить назву пошукової системи Okapi, у якій ця функція була вперше застосована. Базується на вірогіднісній моделі, що була розроблена в 1970-х та 1980-х роках Стівеном Робертсоном, Карен Спарк Джоунс та іншими. Оцінка BM25 визначає залежність релевантності від входження чи не входження слів в запитах з більш ніж одного слова. Нехай є декілька запитів, що складаються з декількох слів. Порівнюються два документи, причому перший документ не містить слова *x*. Згідно з обрахунками, оцінка слова - це сума релевантності кожного із слів:

$$score(D,Q) = IDF(q_i) \cdot \sum_{i=1}^{n} \frac{f(q_i, D) \cdot (k_i + 1)}{f(q_i, D) + k_i (1 - b + b \cdot \frac{|D|}{avgdl})}. \quad (1)$$

Релевантність кожного із слів рівна його *IDF*, що помножений на другий його множник у виразі (1). Релевантність всього пошукового запиту рівна сумі релевантності всіх його слів. Таким чином, відсутність слова (його частота $f(q_i,D)$ рівна нулю) дає релевантність 0. Тому якщо по деяким двом словам score буде рівним, то більш релевантним буде той документ, що має слово *x*. Тому перевага у

пошуку в запитах з більш ніж двох слів (одно з яких менш вживане – більш вузькоспеціалізоване) буде надаватись документам, що містять це вузькоспеціалізоване слово. Відсотковий ріст BM25 від числа входжень показано на рисунку 1. IDF приймає від'ємні значення для слів, що входили більш ніж до половини документів. Слова, що часто зустрічались, суттєво зменшують кінцеву оцінку документа.

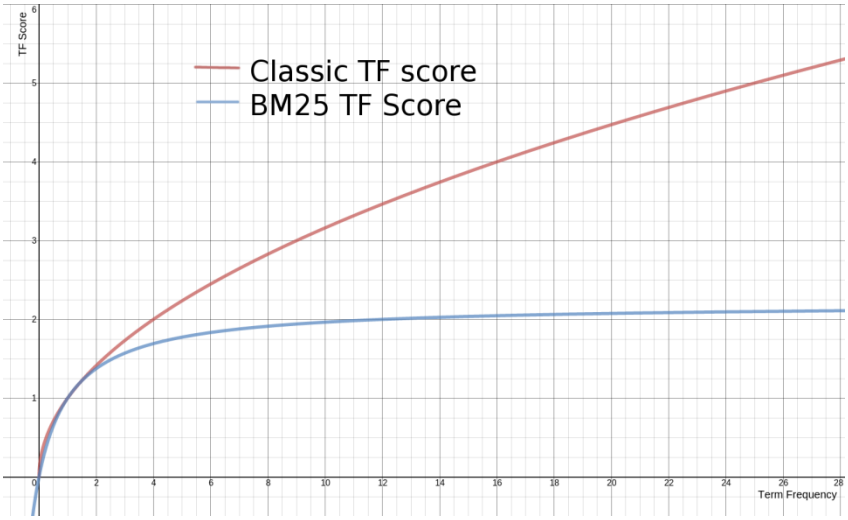

Рис. 1. Відсотковий ріст BM25 від числа входжень

Використання методу BM25 дає основу для нейромереж багатьох ведучих пошукових систем, таких як Google. Зокрема, обчислення BM25, вірогідно, входять до формули ранжування Google у якості окремих факторів. Це визначає перспективність використання оцінки ранжування методу BM25 в прикладних інформаційних технологіях, які включають етап визначення множин ключових термінів, зокрема при формуванні семантичного ядра електронних навчальних матеріалів для оцінки їх відповідності вимогам, оцінки якості тестів, генерації прототипів тестових завдань, автоматизації формування рефератів та анотацій до елементів навчальних матеріалів тощо.