УДК 004.8

Ряба А.О., Мазурець О.В.

Хмельницький національний університет

РІЗНОВИДИ МЕТОДУ ПОШУКУ КЛЮЧОВИХ СЛІВ У ЦИФРОВИХ ТЕКСТАХ ЗА ДИСПЕРСІЙНИМ ОЦІНЮВАННЯМ

Розглянуто особливості пошуку ключових слів у цифрових текстах, зокрема за методом пошуку ключових слів за дисперсійним оцінюванням. Досліджено, яким чином і в якій кількості визначаються відстані для кожного унікального слова тексту, для різних видів вихідного методу пошуку ключових слів за дисперсійним оцінюванням DE-BM, зокрема DE-BM2, DE-BM3, DE-BM4 та DE-BM5.

The specificities of keyword search in digital texts are discussed, including the variance-based keyword search method. It is investigated how and in what quantity the distances for each unique word of text are determined for different types of the original variance search method DE-BM variance, in particular DE-BM2, DE-BM3, DE-BM4 and DE-BM-5.

Семантика є наукою, яка вивчає сенс слів, речень, проводить аналіз тексту. Формальний аналіз семантики перетинається з багатьма іншими областями дослідження, включаючи лексикологію, синтаксис, прагматику, етимологію тощо [1]. Наприклад, проведення семантичного аналізу дозволяє зробити висновки про необхідність коригування текстового контенту для оптимізації його під пошуковики та/або користувачів в залежності від їх цілей [2].

Першочерговою задачею при семантичному аналізі цифрових текстів є пошук ключових слів, множина яких є найбільш стиснутим виглядом семантики тексту [3].

Відомо ряд методів пошуку ключових слів. Зокрема, метод пошуку ключових слів за дисперсійним оцінюванням DE використовує оцінку відстані між словами у тексті [4], та його різновиди, що мають індекс DE-BM.

Дисперсійна оцінка є оцінкою дискримінантної сили слів й дозволяє відділити із загальної множини широковживаних у тексті слів слова, що розташовані рівномірно [5]. Обчислюється наступним чином.

Якщо деяке слово, наприклад T, позначається як T_k^n, де індекс k – номер появи даного слова у тексті, а n – позиція данного слова у тексті. Інтервалом між послідовними появами слова при таких позначеннях буде

величина $\Delta T_k^m = T_{k+1}^m - T_k^n = m - n$, де на *m-й* та *n-й* позиції в тексті знаходиться слово *A*, яке зустрілось *k*+1-*й* і *k*-*й* рази.

Дана оцінка розраховується як: $DE = \sqrt{(\Delta T^2) - (\Delta T)^2} \,/ (\Delta T)$, де (ΔT) – середнє значення послідовності $\Delta T_1, \Delta T_2, \Delta T_k$, (ΔT^2) – послідовність T_1^2, T_2^2, T_k^2, де *k* – кількість появи слова *A* у тексті [6].

Дисперсійна оцінка дозволяє відокремити слова, що зустрічаються в тексті відносно рівномірно (для рівномірно розподілених слів ця оцінка дорівнює нулю), від слів, розподілених нерівномірно. Тобто це оцінка дискримінантної сили слів, зокрема, для інформаційного пошуку. Ідея даної оцінки близька до TF-IDF, однак більш коректно застосовується до цілих текстів, а не до масивів з великої кількості документів, як TF-IDF.

Дисперсійний аналіз є статистичним методом оцінки зв'язку між факторними й результативними ознаками в різних групах, відібраний випадковим чином, заснований на визначені розходжень (розкиду) значень ознак. В основі дисперсійного аналізу лежить аналіз відхилень всіх одиниць досліджуваної сукупності від середнього арифметичного. Як міра відхилень береться дисперсія – середній квадрат відхилень. Відхилення, викликані впливом факторної ознаки (фактору) порівнюються з величиною відхилень, викликаних випадковими обставинами. Якщо відхилення, викликані факторною ознакою, більш істотні, ніж випадкові відхилення, то вважається, що фактор впливає на результуючу ознаку.

Обрахунок відстаней між словами у тексті є підготовчим етапом до дисперсійного оцінювання слів, за якого визначаються для кожного слова (з кількістю появ у тексті більше одного) всі відстані між сусідніми їх появами.

У залежності від того, яким чином і в якій кількості визначаються відстані для кожного унікального слова тексту, розрізняють різні види DE-BM вихідного методу пошуку ключових слів за дисперсійним оцінюванням DE-BM, а саме: DE-BM2, DE-BM3, DE-BM4, DE-BM5.

Метод пошуку ключових слів DE-BM для *n* появ слова враховує *n*-1 відстані. При цьому за відстань береться різниця між меншим порядковим номером наступного слова й більшим порядковим номером попереднього слова.

Метод пошуку ключових слів DE-BM2 для *n* появ слова враховує *n* відстаней. При цьому за відстань береться різниця між меншим порядковим номером наступного слова й більшим порядковим номером попереднього слова. Також додатково враховується відстань від початку тексту до першої появи слова у тексті.

Метод пошуку ключових слів DE-BM3 для *n* появ слова враховує *n* відстаней. При цьому за відстань береться різниця між меншим порядковим номером наступного слова й більшим порядковим номером попереднього слова. Також додатково враховується відстань від останньої появи слова у тексті до кінця тексту.

Метод пошуку ключових слів DE-BM4 для *n* появ слова враховує *n*+1 відстань. За відстань береться різниця між меншим порядковим номером наступного слова й більшим порядковим номером попереднього слова. Також додатково враховуються відстані: від початку тексту до першої появи слова у тексті, від останньої появи слова у тексті до кінця тексту.

Метод пошуку ключових слів DE-BM5 для *n* появ слова враховує *n* відстаней. За відстань береться різниця між меншим порядковим номером наступного слова й більшим порядковим номером попереднього слова. Також додатково враховується відстань, рівна сумі різниць між початком тексту до першої появи слова та між останньою появою слова до кінця тексту.

Інтелект властивий людям, а також спостерігається у тварин.

Людина застосовує інтелект для обробки наявної інформації, наприклад, з метою побудови або вдосконалення розуміння, позиції, стратегії, методу, правила, комбінації, відношення, пояснення, рішення, плану чи цілі. Інтелект пов'язаний з іншими внутрішніми властивостями людини, такими як сприйняття, пам'ять, мова, уява, самосвідомість, самоконтроль, характер, володіння тілом, творчість, інтуїція і власне формується завдяки функціонуванню означених параметрів особистості. Інтелект найчастіше спрямовується на вирішення питань облаштування побуту і відпочинку, професійну діяльність, міжособистісні стосунки та самовдосконалення.

В повсякденному житті в сучасній розвинутій людині інтелект також проявляє себе у вигляді внутрішніх почуттів і образів мислення, таких як відчуття реальності, часу, простору, себе, ритму, відповідальності, гумору, ситуації, прекрасного, небезпеки, захищеності, такту, комфорту, міри, справедливості, довіри, свободи, поваги, власної гідності та інших, і у вигляді аналітичного, образного, практичного, абстрактного, тактичного або стратегічного образу мислення.

Рисунок 1 – Текст для аналізу [7]

Наприклад, у тексті (рис. 1) [7], що складається з 131 слова, й у якому слово «інтелект» зустрічається на позиціях 1,11,34, 60 та 83, для обрахунку дисперсійної оцінки можна використати наступні відстані:

– відстані, рівні 10, 23, 26, та 23 за класичним підходом до обрахунку;
– додаткова відстань від початку тексту до першої появи слова у тексті 1;
– додаткова відстань від останньої появи слова у тексті до кінця тексту 48;

– додаткова відстань від останньої появи слова до попередньої появи слова 23, 26, 23,10;
– додаткова відстань, рівна сумі різниць між початком тексту до першої появи слова та між останньою появою слова до кінця тексту 49.

Відповідно, для різновидів дисперсійного оцінювання слів буде використано наступні відстані:

– для вихідного методу пошуку ключових слів за дисперсійним оцінюванням DE-BM: 10, 23, 26, 23;
– для методу пошуку ключових слів DE-BM2: 0, 1;
– для методу пошуку ключових слів DE-BM3: 83;
– для методу пошуку ключових слів DE-BM4: 23, 26, 23, 10;
– для методу пошуку ключових слів DE-BM5: 1, 83.

Дослідження ефективності розглянутих різновидів методів пошуку ключових слів може визначити їх особливості та рекомендовані області застосування.

Перелік посилань

1. Серажим К. С. Семантичний і семіотичний аспекти аналізу текстів / К. С. Серажим // Вісник Київського національного університету імені Тараса Шевченка. Журналістика. – Київ, 2013. – № 20. – С.34-36.
2. Chen J. Smart Data Integration by Goal Driven Ontology Learning / J. Chen, D. Dosyn, V. Lytvyn, A. Sachenko // Advances in Big Data. – 2016. – Т. 529. – С. 283-292.
3. Мазурець О. В. Онтологічний підхід до побудови семантичної моделі навчальних матеріалів / О. В. Мазурець // Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2017, №6. – С. 223-229.
4. Крак Ю. В. Практична реалізація інформаційної технології автоматизованого визначення множини семантичних термінів в контенті навчальних матеріалів / Ю. В. Крак, О. В. Бармак, О. В. Мазурець // Науковий журнал «Проблеми програмування». Київ, 2018, №2-3. – С.245-254.
5. Ventura J. New Techniques for Relevant Word Ranking and Extraction / J. Ventura, J. Silva // Proceedings of the artificial intelligence 13th Portuguese conference on Progress in artificial intelligence, EPIA'07. – Berlin: Springer-Verlag, Berlin, Heidelberg, 2007. – С. 691-702.
6. Ландэ Д. В. Компактифицированный горизонтальный граф видимости для сети слов / Д. В. Ландэ, А. А. Снарский // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения» – КПИ, Киев: 2013. – С.158-164.
7. Інтелект. Вікіпедія [Електронний ресурс] – Режим доступу: https://uk.wikipedia.org/wiki/Інтелект