Правова інформатика

УДК 004.67

ЛАНДЕ Д.В., *доктор технічних наук*

МЕТОДИ ОЦІНКИ РІВНЯ ДИСКРИМІНАНТНОЇ СИЛИ СЛІВ У ТЕКСТАХ З ПРАВОВОЇ ТЕМАТИКИ

Анотація. *Розглянуто підходи до оцінки дискримінантної сили слів у текстах з правової тематики. Підходи перевірені на колекції законодавчих актів України та масиві новинних повідомлень. Запропоновано метод візуалізації рівня нерівномірності входження слів у тексти.*

Ключові слова: *фрагменти текстів, автоматичний пошук та аналіз.*

Аннотация. *Рассмотрены подходы к оценке дискриминантной силы слов в текстах правовой тематики. Подходы проверены на коллекции законодательных актов Украины и массиве новостных сообщений. Предложен метод визуализации уровня неравномерности вхождения слов в тексты.*

Ключевые слова: *фрагменты текстов, автоматический поиск и анализ.*

Summary. *Approaches are considered to the estimation of discrimination force of words in the texts of legal subject. Approaches are tested on collection of legislative acts of Ukraine and array of news-related reports. The method of visualization of level of unevenness of including of words in texts is offered.*

Keywords: *fragments of texts, automatic search and analysis.*

Постановка проблеми. Ключові слова для пошуку в тексті, опорні слова для автоматичного екстрагування значущих фрагментів текстів або формування автоматичних рефератів, вибираються з урахуванням такої властивості слів, як “розпізнавальна” або дискримінантна сила. Адже якщо слово відносно рівномірно розподілено по тексту документа, то воно навряд чи може використовуватися для ефективного змістовного пошуку або служити основою вибору якогось значущого фрагмента, який може розглядатися як деяка надфразова єдність [1]. При аналізі текстів з правової тематики, зокрема, при вирішенні завдання формування електронної енциклопедії на основі аналізу всього масиву законодавчих актів України, оцінка дискримінантної сили окремих слів має найважливіше значення.

Одна з перших технологій оцінки якості ключових слів була “матеріалізована” Солтоном в векторно-просторовій моделі пошуку [2], в якій саме для обліку дискримінантної сили слів було введено поняття інверсної частоти появи слова в окремих документах масиву. Запропонований метод зважування слів має сьогодні стандартне позначення – TF IDF, де TF вказує на частоту появи слів у документі, а IDF – на величину, зворотну до кількості документів у масиві, що містять дане слово (точніше, логарифм, монотонну функцію від цієї величини):

$$w_i = tf_i \cdot \log \frac{N}{n_i},$$

де: w_i – вага слова t_i, tf_i – частота слова t_i у документі, n_i – кількість документів в інформаційному масиві, у яких застосовується слово t_i, слова N – загальна кількість документів в інформаційному масиві.

Оцінка нерівномірності входження слів можлива і на основі чисто статистичних, дисперсійних оцінок. В роботі [3] запропонована така оцінка дискримінантної сили слова:

$$\sigma_i = \frac{\sqrt{\langle d^2 \rangle - \langle d \rangle^2}}{\langle d \rangle},$$

де: $\langle d \rangle$ – середнє значення послідовності $d_1,\ d_2,...,\ d_n$, n – кількість появ слова t_i в інформаційному масиві.

Якщо позначити координати (номери) входження слова t_i в інформаційний масив як $e_1,\ e_2,...,\ e_n$, то $d_k = e_{k+1} - e_k \quad (e_0 = 0)$.

Для візуалізації нерівномірності входження слів в тексти в [3] була запропонована технологія спектограмм, які зовні нагадують штрих-коди товарів [4], разом з тим не дозволяють розглядати входження слів у різних масштабах вимірювань, як це робиться, наприклад, у вейвлет-аналізі [5].

Метою статті є дослідження методів оцінки рівня дискримінантної сили слів у нормативних текстах.

Виклад основних положень. Автором запропоновані та реалізовані інструментальні засоби, що дозволяють візуалізувати щільність появи слова в тексті в залежності від ширини вікна спостереження. Через веб-інтерфейс відповідної програми (http://ling.infostream.ua/jag/jag.html) вводиться текст і слово для аналізу. У результуючій спектограмі по горизонталі відкладаються номери входження слів у тексті, а по вертикалі – ширина вікна спостереження. Одному входженню слова відповідає світло-сірий колір. Якщо у відповідне вікно спостереження потрапляє кілька цільових слів, то воно зафарбовується більш темним відтінком. Експерт – прикладний лінгвіст за зовнішнім виглядом відразу може визначити ступінь рівномірності входження в текст слова, що аналізується [6].

Розраховані автором коефіцієнти нерівномірності входження окремих слів у добірці законодавчих актів України, що відносяться до формування та розвитку інформаційного простору держави (закони України “Про доступ до публічної інформації”, “Про Основні засади розвитку інформаційного суспільства в Україні на 2007-2015 роки”, “Про телекомунікації”, “Про захист персональних даних”, “Про основи національної безпеки України”) наведені у Табл. 1, а відповідні спектограми – на Рис. 1 – 5.

При розрахунку коефіцієнта w_i використовувався штучний прийом, вихідний текст розбивався на фрагменти фіксованої довжини по 500 слів, які при розрахунках TF IDF розглядаються як окремі фрагменти документів. Як видно, нерівномірність входження окремих слів, що точно виражається в коефіцієнтах w_i і σ_i, може бути визначена візуально у спектограмах. Однак монотонність зростання значень w_i порушується в одному випадку (слова “Безпека” і “Електронний”), що пояснюється різними підходами, що застосовуються для розрахунку w_i та σ_i і частою появою першого слова.

Табл. 1. Значення коефіцієнтів нерівномірності для окремих слів у добірці законодавчих актів України

Слово	**Входжень**	w_i **(TF IDF)**	σ_i
Технології	46	53,62	1,99
Оприлюднення	33	53,98	2,21
Безпека	102	96,15	2,41
Електронний	50	85.24	3.22
Регулювання	220	129,92	3,62

Рис. 1 – Спектограма входження слова “Технології”
Рис. 2 – Спектограма входження слова “Оприлюднення”
Рис. 3 – Спектограма входження слова “Безпека”
Рис. 4 – Спектограма вхождения слова “Електронний”
Рис. 5 – Спектограма входження слова “Регулювання”

Аналогічні розрахунки були проведені для масиву з 50 новинних веб-публікацій 2012 р. з тематикою, яка визначається запитом до системи контент-моніторингу InfoStream [7] (Табл. 2, Рис. 6 – 10):

(захист~персональн~даних) | кібербезпек | **(**інформац~безпек**).**

У цьому випадку монотонність зростання значень по відношенню до слова “Безпека” порушується.

Слід звернути увагу, що дискримінантна сила окремих слів на двох розглянутих добірках істотно розрізняється, що пов’язано з стилем і змістом відповідних текстів.

Табл. 2. Значення коефіцієнтів нерівномірності для окремих публікацій у масиві новинних повідомлень

Слово	**Входжень**	w_i **(TF IDF)**	σ_i
Регулювання	16	33,07	1,26
Оприлюднення	18	35,49	1,50
Безпека	56	71,59	1,75
Електронний	28	50.53	2.02
Технології	39	61,45	2,06

Рис. 6 – Спектограма входження слова “Регулювання”
Рис. 7 – Спектограма входження слова “Оприлюднення”
Рис. 8 – Спектограма входження слова “Безпека”
Рис. 9 – Спектограма входження слова “Електронний”
Рис. 10 – Спектограма входження слова “Технології”

Висновки.

Можна зробити висновок, що крім традиційного підходу до оцінки дискримінантної сили слів у текстах, запропонованого Солтоном, дисперсійний аналіз дає близькі за якістю результати. Незважаючи на те, що підхід TF IDF за останній час

пройшов ряд трансформацій, доповнюється допоміжними параметрами, зокрема, отримав популярність метод BM25, що враховує довжину документів, дисперсійний аналіз виявляється досить перспективним.

Розглянуті приклади показали, що штучний прийом, що полягає в тому, що вихідний текст великого розміру розбивався на фрагменти фіксованої довжини, цілком виправдався, результати багато в чому збіглися з результатами, отриманими іншим методом.

Наведені приклади показують, нерівномірність слів в масивах новинних повідомлень і в офіційних документах має близьку, багато в чому аналогічну природу, проте дискримінантна сила окремих слів на двох розглянутих добірках істотно розрізняється, що пов’язано з стилем і змістом відповідних текстів.

І, нарешті, запропонований метод візуалізації нерівномірності входження слів, у порівнянні з існуючими, додав ще один вимір – величину вікна спостереження, що виявилося зручним при розгляді текстових (у тому числі документальних) масивів великих обсягів. Техніка спектограм дозволяє експертам без додаткових зусиль якісно оцінювати значення окремих слів при формуванні так званих надфразових єдностей, екстрагуванні фрагментів текстів для формування довідкових документів.

Використана література

1. Ягунова Е.В., Ландэ Д.В. Динамические частотные характеристики как основа для структурного описания разнородных лингвистических объектов : труды 14-й Всероссийской научной конференции [“Электронные библиотеки: перспективные методы и технологии, электронные коллекции”], (Россия, Переславль-Залесский, 15-18 октября 2012 г). – С. 196-205. – RCDL-2012. .

2. Salton G., McGill M. J. Introduction to Modern Information Retrieval. – New York : McGraw-Hill, 1983. – 448 p.

3. Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA / Europhys. Lett., 2002, 57. – P. 759 – 764.

4. Carpena P., Bernaola-Galván P., Hackenberg M., Coronado A.V., Oliver J.L. Level statistics of words: Finding keywords in literary texts and symbolic sequences / Phys Rev E Stat Nonlin Soft Matter Phys. 2009, E 79. – P. 035102-1 – 035102-4.

5. Чуи К. Введение в вэйвлеты / К. Чуи. – М. : Мир, 2001. – 416 с.

6. Ландэ Д.В. Визуализация статистики вхождения слов / MegaLing'2009. Горизонты прикладной лингвистики и лингвистических технологий : материалы международной конференции (Украина, Киев, 21-26 сентября 2009 г.). – К. : Довіра, 2009. – С. 63-64.

7. Ланде Д.В. Програмно-апаратний комплекс інформаційної підтримки прийняття рішень : науково-методичний посібник / Д.В. Ланде, В.М. Фурашев, О.М. Григор’єв. – К. : Інжиніринг, 2006. – 48 с.

~~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~~
~~~~~~~~~~~~~~