DOI 10.31891/2307-5732-2019-275-4-80-83
УДК 004.94

О.В. ДЖУРАБАЄВ, О.В. БАРМАК, Е.А. МАНЗЮК, Т.К. СКРИПНИК
Хмельницький національний університет

ІНФОРМАЦІЙНА ЗОСЕРЕДЖЕНІСТЬ ЗМІСТОВНОСТІ В ТЕКСТІ

В роботі було запропоновано підхід для дослідження текстової інформації з описом його теоретичної частини. Підхід полягає в дослідженні текстової інформації як сигналу. Було реалізовано інформаційну технологію та проведено дослідження з описом результатів і побудовою графіків тексту.

Ключові слова: інформаційна технологія, аналіз тексту.

O.V. DZHURABAIEV, O.V. BARMAK, E.A. MANZIUK, T.K. SKRYPNYK
Khmelnytskyi National University

SEARCHING FOR CONTEXT IN THE TEXT

Nowadays the search for keywords is not complicated, because there are effective algorithms for their search. The most popular algorithms are TF-IDF, the Bag of Words. There are main disadvantages of these methods: the removal of stopwords, the lack of including the position of each word in the text. The aim of research is development of information technology for searching context and test the effectiveness of to search for keywords in the text without removing stopwords and taking into attention the position of each word. Also, the aim of research is the development of information technology to find places of content concentration in the text with minimal time and low CPU usage returns the correct result for a certain range of tasks in the case of compliance with the limits of input data. The paper proposes an approach based on the analogy of the physical phenomenon of the signal, for constructing a "meaning recognizer", which does not require any training base, nor a deep machine analysis of the text, and returns the approximate result. The approach is to normalize the text, build the amplitude and phase vectors, and then plot the dependencies of the calculated parameters and visualize the text. Also described are the results of experiments on the recognition of content in the test data. The results of research have shown that the greatest effectiveness is obtained with a text belonging to a specific category. Information technology for the search of content in text information allows graphically to present text in the form of a three-dimensional model, which makes it possible to identify grouped concentrations. In the final case, this allows us to visually cluster groups of words that are a vector of signs of content concentration. Thus, the textual information is presented in the form of a clustered three-dimensional model based on the content concentration, presented in the form of key words of content. It is revealed the basic characteristics of text information as the basic representation after transformation in the form of numerical dimensional characteristics. This presentation is the basis for further research in the direction of clustering and text classification. The results of the research have confirmed that this method is effective for the case where the text belongs to one category. In case you research several texts of a similar category, you can create a set of words that best characterize these texts (the classifier's core). You can also conduct visually researches of texts as surfaces.

Keywords: information technology, text processing.

Розвиток інформаційних технологій, включаючи соціальні мережі, форуми та інші ресурси, призводить до збільшення кількості користувачів, що породжує потребу класифікації та систематизації текстової інформації, і для часткового вирішення цієї проблеми існують готові програмні рішення. Мова – непроста річ, і сама ідея реалізації алгоритму розуміння людської мови є цікавим нетривіальним завданням, тому кожна технологія чудово підходить для вирішення конкретного, специфічного, вузького кола завдань. І хоч з 90-х років минулого століття і була реалізована велика кількість програмних продуктів для дослідження тексту, як новачками-аматорами, так і професіоналами, актуальність пошуку сенсу в текстовій інформації стоїть гостро й до сьогодні.

Серед багатьох відомих на даний час методів досліджень текстової інформації найбільшою популярністю користуються: метод частотного аналізу терму з врахуванням інверсії частоти до інших документів (Term Frequency – Inverse Document Frequency) [1], лінійний метод опорних векторів (Linear Support Vector Machine) [2] та метод так званого «мішка» або множини слів (Bag of Words) [3]. До переваг TF-IDF та Bag of Words можна віднести швидкість та масовість; головною перевагою методу опорних векторів є точність. До недоліків – повільність виконання, «відкидання» стоп-слів, що неминуче веде до втрати сенсу; відсутність врахування позиції кожного слова у тексті, що може провокувати ускладнення пошуку змісту в конкретному тексті.

Розпізнавання змісту в тексті, в ідеальному випадку, потребує розробки синтаксичного аналізатора мови, а також ряду програмних лінгвістичних компонентів. Окрім цього, для коректних результатів часто потрібно мати навчальну вибірку або базу, а також проводити ресурсозатратні обрахунки. Це пов’язано з тим, що результати проведення аналізу тексту можна умовно поділити на дві групи: до першої групи віднесемо випадок для якого достатньо наближених або приблизних результатів; в іншому випадку необхідно отримати на виході точні, конкретні результати [4].

Метою роботи є розробка інформаційної технології для пошуку місць зосередженості змісту в тексті, яка з мінімальними затратами часу та потужності ЕОМ повертає коректний результат для певного кола завдань у разі дотримання обмежень вхідних даних.

В роботі запропоновано підхід для побудови «розпізнавача» сенсу, який не потребує ні навчальної бази, ні глибокого машинного аналізу тексту і повертає наближений результат. Підхід полягає в нормалізації тексту, побудові векторів амплітуди та фази, після чого будуються графіки залежностей обрахованих параметрів та з’являється можливість аналізувати текст візуально. Також описані результати експериментів по розпізнаванню змісту в підготовлених текстах.

Вхідними даними для застосування технології є текстовий контент, який має певний сенс. Це може бути, наприклад, наукова стаття, кулінарний рецепт або якась новина чи повідомлення. Тематика інформації обмежується уявою користувача і може бути довільною. Важливим є лише те, що ця інформація повинна бути у вигляді текстового файлу.

В основі функціонування інформаційної технології пошуку сенсу в тексті є дослідження тексту як скінченної множини слів. Блок-схема роботи програми зображена на рис. 1. Розглянемо кожен пункт детальніше.

Зчитування даних передбачає подання на вхід користувачем файлу, що містить необхідну для аналізу текстову інформацію.

Фільтрація включає в себе видалення з тексту розділових знаків, html-тегів та усіх інших символів, які відсутні в алфавіті. Нормалізація – приводить кожне слово до нормальної форми [5], наприклад, для слова «технологією» нормальна форма – «технологія». Після нормалізації текстовий файл є множиною слів, приведених до нормальної форми.

В свою чергу, експрес-аналіз складається з кількох етапів, зображених на рис. 2.

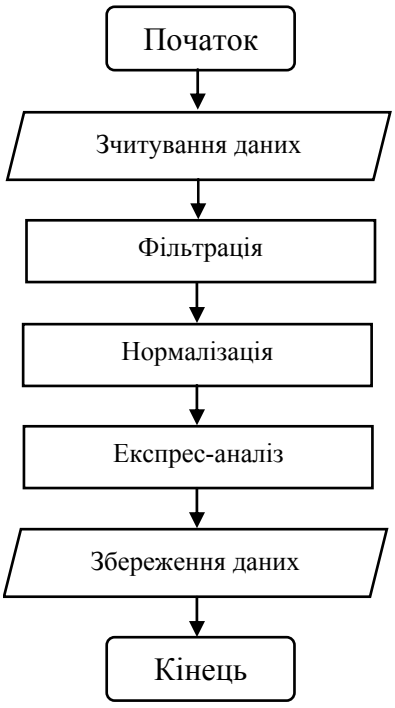

Рис. 1. Блок-схема роботи інформаційної технології для пошуку сенсу в текстовій інформації

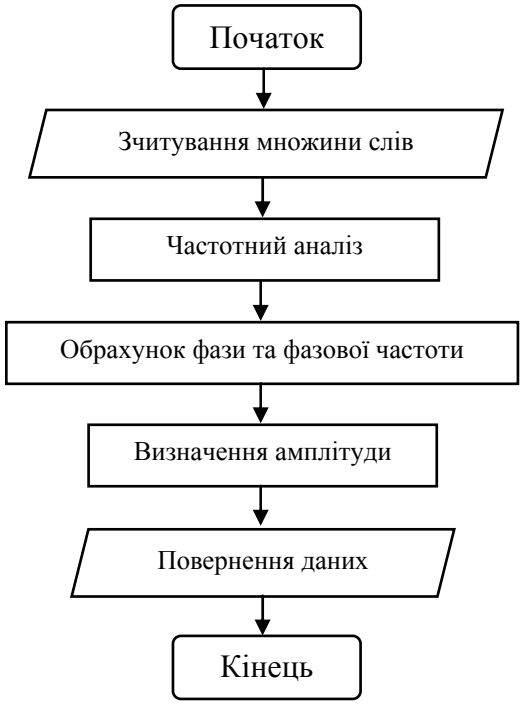

Рис. 2. Блок-схема роботи експрес-аналізу текстової інформації

Для кращого розуміння суті здійснених у розрахунках величин, проведемо аналогію з таким фізичним явищем, як сигнал. Так само, як сигнал є сукупністю змін фізичної величини [6], множину слів можна представити як сукупність термів та їх властивостей, які утворюють скінченний потік даних. Звідси слідує, що існує можливість дослідити кожен терм у тексті на предмет частоти, фази і амплітуди.

Позначимо множину слів як *W*, слово як *w*, множину унікальних термів як *S*, а терм як *s*.

Частота терму – це кількість його входжень у множину слів:

$$F(s_i) = \underset{w_i \in W}{count}(w_i), s \in S, i = 1, \ldots, |S| \tag{1}$$

Фазова частота терму – кількість входжень терму в підмножину множини слів. Нехай *W′* – підмножина множини слів, тоді:

$$F_\omega(s_i) = \underset{w_i \in W', W' \in W}{count}(w_i), s \in S, i = 1, \ldots, |S| \tag{2}$$

Фаза терму – індекс, при якому його фазова частота має максимальне значення:

$$\omega(s_i) = \underset{s_i \in S}{index}\left(argmax(F_\omega(s_i))\right), i = 1, \ldots, |S| \tag{3}$$

Амплітуда терму – максимальне значення зміщення або зміни кількості входжень терму від його середнього значення, іншими словами розсіювання конкретного терму в множині слів:

$$A(s_i) = \sum \left| \underset{s_i \in S}{index}(s_i) - \omega(s_i) \right| \tag{4}$$

Підставивши $\omega(s_i)$, отримаємо:

$$A(s_i) = \sum \left| \underset{s_i \in S}{index}(s_i) - \underset{s_i \in S}{index}\left(argmax(F_\omega(s_i))\right) \right| \quad (5)$$

Результати розрахунків за вище вказаними формулами записуються в файл, після чого можуть бути використані для візуалізації аналізу тексту.

Для прикладу візьмемо статтю «Приготування їжі як вид мистецтва». Вхідний файл містить 126 слів, з них унікальних, тобто термів – 80.

Нижче зображений точковий графік залежності амплітуди, частоти та її фази на прикладі вхідної статті. Кожна точка цього графіку є унікальним термом без повторень.

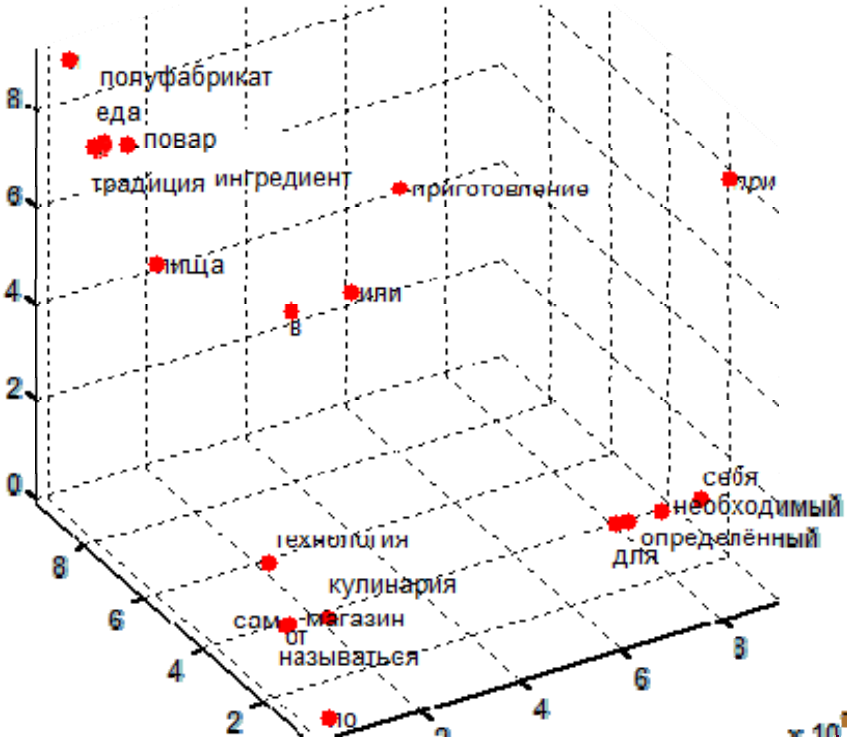

Рис. 3. Точковий графік обробленої множини слів

З тримірного представлення тексту (рис. 3) можна виділити кілька груп або кластерів. Якщо згрупувати множини точок частоти і амплітуди, утвориться графік так званої «густини» термів, приклад якого зображений на рисунку 4. Тут чітко можна розрізнити 4 кластери. Розглянемо їх.

Вектор термів, які мають велику амплітуду і малу частоту – рівномірно розподілені по всьому тексту та рідко зустрічаються. В даному випадку це слова: «для», «определённый», «необходимый», «при», «себя».

Вектор термів, які мають малу амплітуду та малу частоту – зустрічаються рідко, але концентровано. До нього входять наступні слова: «технология», «в», «магазин», «от», «сам», «кулинария», «называться», «по», «от», «или».

В ході ряду експериментів було виявлено, що терм або група термів, яка є рівновіддаленою від інших груп та має помірну амплітуду і частоту, утворює вектор контексту та її можна вважати головною тематикою текстової інформації. В даному випадку це один терм – «приготовление». Хоча за певних умов ця закономірність не справджується, особливо, коли вхідна текстова інформація не має чіткої тематики.

Слова «традиция», «полуфабрикат», «ингредиент», «кухня» та всі інші, які залишилися, мають високу частоту та малу амплітуду, що говорить про те, що вони часто зустрічаються на відносно малих ділянках тексту.

Дослідження тексту не обмежується побудовою «плоских» графіків, в деяких випадках може виявитись корисною поверхня. Наприклад, при дослідженні подібних за тематикою статей і побудові їх поверхонь можна дослідити точки перетину, – в широкому розумінні це буде ядро класифікатора або набір слів, які найкраще характеризують тему вхідного тексту.

Варто зазначити, що кращі результати можна отримати тільки в тому випадку, коли текстова інформація несе зміст з певної конкретної категорії. При «суміші» різноманітних тем в одній статті отримуються результати, за яких не вдається зробити коректний висновок щодо змісту тексту.

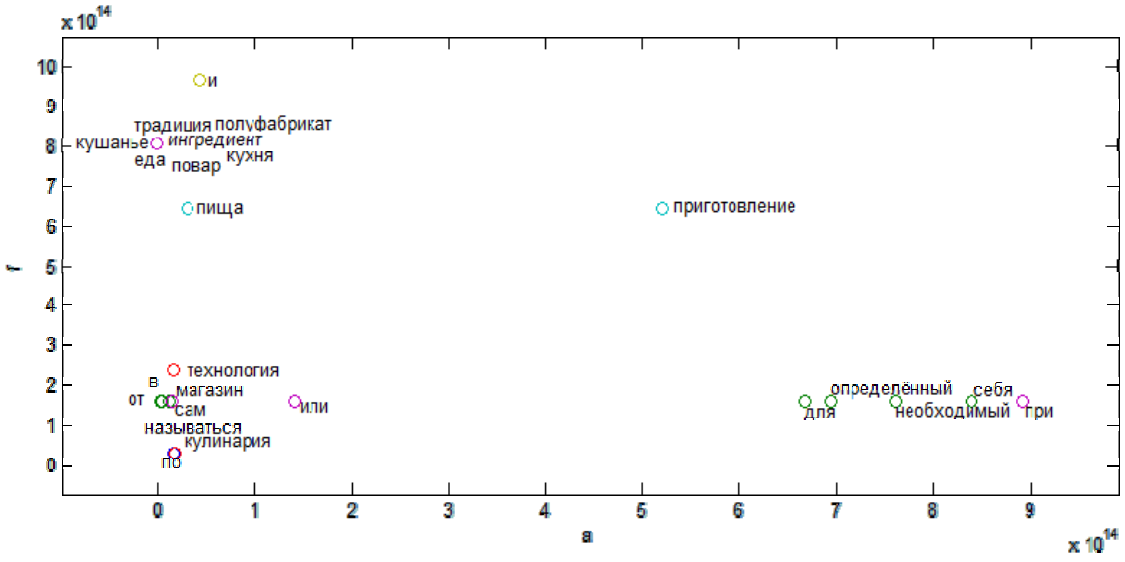

Рис. 4. Графік густини термів

Висновки

Результати досліджень показали, що найбільша ефективність отримується при тексті, що належить до конкретної категорії. Інформаційна технологія для пошуку змісту в текстовій інформації дозволяє графічно представити текст у вигляді тримірної моделі, що дає можливість виявити груповані зосередженості. В кінцевому випадку це дозволяє візуально кластеризувати групи слів, які є вектором ознак концентрації змісту. Таким чином текстова інформація подається у вигляді кластеризованої тримірної моделі за ознаками зосередженості змісту, що представлена у вигляді ключових слів змістовності.

Виявлено основні ознакові характеристики текстової інформації як базове представлення при трансформації у вигляді числових розмірних характеристик. Це представлення є основою для подальших досліджень в напрямку кластеризації та класифікації тексту.

Література

1. B. Das, S. Chakraborty. An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation, India, 2013, pp. 1–3.

2. M. Labbé, L.I. Martínez-Merino, A.M. Rodríguez-Chía. Mixed Integer Linear Programming for Feature Selection in Support Vector Machine, Departamento de Estad´ıstica e Investigaci´on Operativa, Universidad de C´adiz, Spain, August 8, 2018, pp. 2–5.

3. B. Heap, M. Bain, W. Wobcke, A. Krzywicki, S. Schmeidl. Word Vector Enrichment of Low Frequency Words in the Bag-of-Words Model for Short Text Multi-class Classification Problems, Sydney NSW, Australia, 2017. pp. 1-2.

4. Єрмаков А.Є. Статистична модель для розпізнавання сенсів у текстах іноземною мовою з навчанням на прикладах з паралельних текстів : підручник / Єрмаков А.Є., Поляков П.Ю. – Москва, 2017. – 397 с.

5. Erica K. Shimomoto, Lincon S. Souza. Text Classification based on Word Subspace with Term-Frequency, University of Tsukuba, Japan, 2018, pp. 2–4.

6. Прэтт У. Цифровая обработка изображений / Прэтт У. ; пер. с англ. – Москва : Мир, 1982. – С. 32–35.

References

1. B. Das, S. Chakraborty. An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation, India, 2013, pp. 1–3.

2. M. Labbé, L.I. Martínez-Merino, A.M. Rodríguez-Chía. Mixed Integer Linear Programming for Feature Selection in Support Vector Machine, Departamento de Estad´ıstica e Investigaci´on Operativa, Universidad de C´adiz, Spain, August 8, 2018, pp. 2–5.

3. B. Heap, M. Bain, W. Wobcke, A. Krzywicki, S. Schmeidl. Word Vector Enrichment of Low Frequency Words in the Bag-of-Words Model for Short Text Multi-class Classification Problems, Sydney NSW, Australia, 2017. pp. 1-2.

4. Yermakov A.Ye. Statistichna model dlya rozpiznavannya sensiv u tekstah inozemnoyu movoyu z navchannyam na prikladah z paralelnih tekstiv : pidruchnik / Yermakov A.Ye., Polyakov P.Yu. – Moskva, 2017. – 397 s.

5. Erica K. Shimomoto, Lincon S. Souza. Text Classification based on Word Subspace with Term-Frequency, University of Tsukuba, Japan, 2018, pp. 2–4.

6. William K. Pratt / Digital Image Processing: Translate from English. – Moscow: Mir, 1982. – pp. 32–35.

Рецензія/Peer review : 27.6.2019 р. Надрукована/Printed :17.7.2019 р.
Рецензент: д.т.н., проф. Сорокатий Р.В.