УДК 004.048

В.Ю. Царик, О.І. Михальов

# ІНФОРМАЦІЙНО-СИНЕРГЕТИЧНІ МЕТОДИ В ЗАДАЧАХ TEXT MINING

*Анотація. В статті досліджено застосування інформаційно-синергетичних методів для інтелектуального аналізу текстових даних. Запропоновані методи аналізу текстів, довжина яких менша, чим характеристична довжина мови, на якій вони написані. Проаналізовано ряд літературних творів, що написані російською та українською мовами з використанням функцій відображення (R) та розвитку (D). Запропоновано використання комплексного критерію K, що об'єднує в собі значення функцій R і D. Показано можливість застосування синергетичного підходу до аналізу різних дискретних систем з кінцевою множиною елементів.*

*Ключові слова: інформація, синергетика, Text Mining, порядок та хаос.*

**Вступ.** При аналізі дискретних систем із скінченною множиною елементів можливо розглядати елементи цієї системи в площині таких понять, як порядок та хаос, тобто, аналізувати ступінь упорядкованості елементів в даній системі [1, 2]. Для прикладу розглянемо рисунок 1. На ньому зображена система з 16 елементів та її стани з різним рівнем впорядкованості елементів. На 1a елементи повністю впорядковані, на 1d порядок елементів повністю відсутній. 1b і 1c – проміжні стани системи, де наявні і хаос, і порядок.

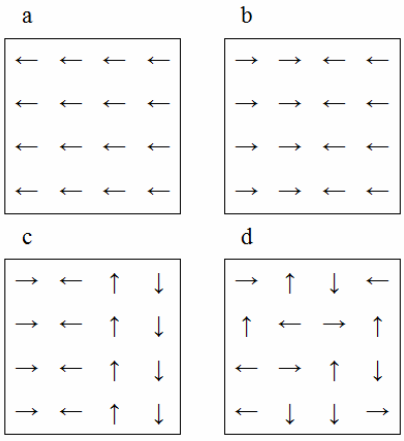


Рисунок 1 – Приклади систем з різним рівнем порядку та хаосу [1]



 ISSN 1562-9945

Кількісна сторона хаосу і порядку в структурі дискретних систем зазвичай характеризується за допомогою статистичної термодинаміки або традиційної теорії інформації [4, 6]. В обох випадках за міру хаосу приймається ентропія. Причому в термодинаміці – це ентропія Л.Больцмана, що статистично виражає другий закон термодінаміки і має теоретичне значення при аналізі молекулярних множин, а в теорії інформації – ентропія (information entropy) множини ймовірностей К. Шеннона, яка служить також і мірою кількості інформації (мікроінформація), переданої у вигляді символьних повідомлень по технічних каналах зв'язку [3, 4].

Відомо [5], що диференційовану оцінку структурного хаосу і порядку в теперішній час ефективно проводити за допомогою синергетичної теорії інформації, предметом пізнання якої є інформаційно-кількісні аспекти відображення дискретних систем в площині ознак їх опису. У даній теорії, при аналізі відображення системи через сукупність своїх частин, отримані функції відображеної і відображеної інформації, за допомогою яких в структурі системи можна індивідуально оцінювати порядок і хаос відповідно. При цьому відображена інформація іменується як адитивна негентропія (мікроінформація) відображення, а невідображена – як ентропія відображення. Також наголошується, що кожна із цих інформаційних функцій має свій безпосередній взаємозв'язок з ентропією Больцмана, а ентропія відображення, крім того, математично тотожна ентропії Шеннона, але, на відміну від останньої, отримана аналітичним шляхом.

У світлі синергетичної теорії інформації мірами структурного порядку і хаосу є адитивна негентропія (мікроінформація) і ентропія відображення, а відношення цих показників іменується R (Reflection) - функцією[1].

$$R = \frac{I_\Sigma}{S} = \frac{I_0 - S}{S} = \frac{I_0}{S} - 1 = \frac{порядок}{хаос}, \qquad (1)$$

де IΣ – адитивна негентропія (мікроінформація як міра порядку на мікрорівні її запам'ятовування), а S – ентропія відображення (міра хаосу). В свою чергу, кількісні значення негентропії и ентропії відображення можливо знайти, використовуючи наступні формули:

$$I_0 = \log_2 M = I_\Sigma + S; \qquad (2)$$

$$I_{\Sigma} = \sum_{i=1}^{N} \frac{m_i}{M} \log_2 m_i; \quad (3)$$

$$S = -\sum_{i=1}^{N} \frac{m_i}{M} \log_2 \frac{m_i}{M}, \quad (4)$$

де $I_0$ – відображена інформація, M – загальна кількість елементів у складі системи, N – кількість частин системи, mi – кількість елементів в і-й частині.

R-функція є кількісною характеристикою, що показує співвідношення хаотичності і впорядкованості в системі, тобто визначає структурну організацію даної системи в цілому. Її значення показує, що і в якій мірі переважає в даній системі – хаос або порядок.

Також виділяють функцію розвитку (D (Develop) - функція), яка є добутком один на одного функцій хаосу і порядку:

$$D = I_{\Sigma} \cdot S. \quad (5)$$

Описані інформаційно-синергетичні функції хаосу і порядку мають універсальний характер і можуть використовуватися при структурному аналізі будь-яких дискретних систем з кінцевою множиною елементів. В теперішній час вже проведено подібний аналіз таких різних за своєю природою систем, як електронні системи атомів хімічних елементів, білкові молекули, павутини павуків, поетичні твори і т.д.

Постановка завдання. В даній статті розглядається застосування методів синергетичної теорії інформації до структурного інтелектуального аналізу текстових даних (Text Mining).

Основна частина. Будь-який мовний текст, від одиничного слова до великого літературного твору, може бути представлений як система, елементами якої є окремі букви, а частини являють собою сукупності однакових літер. Відповідно, за допомогою синергетичної теорії інформації можна проводити структурний аналіз довільних текстів з боку їх хаотичності та впорядкованості за кількістю і кількістю народження окремих літер [6].

Так як Smax=log2N, то із виразу (1) випливає, що після того, як в тексті будут задіяни всі букви алфавіту, подальше збільшення його довжини приведе до того, що значення ентропії відображення S почне коливатись біля своєї верхньої границі Smax, в той час як адитивна негентропія IΣ буде нескінченно зростати. Тобто, R-функція

 ISSN 1562-9945

тексту по суті стає залежною тільки від його довжини. Щоб цього уникнути, потрібно встановити порогову довжину тексту, до досягнення якої, і при якій, ентропія відображення S і адитивна негентропії IΣ будуть знаходитися в рівних умовах. Такою пороговою величиною є довжина тексту, що знаходиться в стані інформаційно-ентропійної рівноваги при використанні всіх букв алфавіту. Таку довжину прийнято називати характеристичною довжиною тексту, і позначати символом L*.

Замінюючи в формулі (1) M на L* і, виходячи з умови $R = \frac{\log_2 L^*}{S} - 1 = 1$, отримуємо, що при інформаційно-ентропійній рівновазі $\log_2 L^* = 2S$, звідки:

$$L^* = 4^S. \qquad (6)$$

Для прикладу покажемо розраховані за отриманою формулою характеристичні довжини для різних мов. Використовуються значення ентропії Шеннона (H), що розраховані за сукупністю відносних частот зустрічаємості різних букв і пробілів в російській, англійській, німецькій, французькій, іспанській та українській мовах. Для російської та української мови даний показник з урахуванням не тільки букв і пробілів, а й знаків синтаксису (.,!?:;).

Таблиця 1

Розрахунок характеристичної довжини тексту для різних мов

| Мова | $L^*$ |
|---|---|
| *З урахуванням букв і пробілів* | |
| Російська | 416 |
| Німецька | 294 |
| Англійська | 267 |
| Іспанська | 249 |
| Французька | 242 |
| *З урахуванням букв і пробілів і розділових знаків* | |
| Російська | 449 |
| Українська | 562 |

Для аналізу текстів необхідно їх розділити на фрагменти, що дорівнюють характеристичній довжині, і досліджувати кожен фрагмент окремо, після чого знайти середні результати для функцій R і D.

Покажемо алгоритм структурного аналізу з використанням інформаційно-синергетичного підходу на прикладі вірша Ліни Костенко «Страшні слова, коли вони мовчать»:

*Страшні слова, коли вони мовчать,*
*коли вони зненацька причаїлись,*
*коли не знаєш, з чого їх почать,*
*бо всі слова були уже чиїмись.*

*Хтось ними плакав, мучився, болів,*
*із них почав і ними ж і завершив.*
*Людей мільярди і мільярди слів,*
*а ти їх маєш вимовити вперше!*

*Все повторялось: і краса, й потворність.*
*Усе було: асфальти й спориші.*
*Поезія – це завжди неповторність,*
*якийсь безсмертний дотик до душі.*

Тут число букв разом з пробілом (N) дорівнює 32, а їх загальна кількість M = 423, що менше за Lукр* = 562. Частота зустрічаємості окремих букв і пробілу, а також розрахунок значень інформаційно-синергетичних функцій хаосу і порядку, наведені в таблиці 2, з якої видно, що по своїй структурній організації наведений текст близький до інформаційно-ентропійний рівноваги (R = 1,08).

Таблиця 2

Розрахунок значень інформаційно-синергетичних функцій хаосу і порядку в буквеній структурі вірша Ліни Костенко «Страшні слова, коли вони мовчать»

| № з.п. | Буква | $m$ | $m\log_2 m$ | № з.п. | Буква | $m$ | $m\log_2 m$ |
|---|---|---|---|---|---|---|---|
| 1 | пробіл | 110 | 745,95 | 17 | к | 8 | 24,00 |
| 2 | и | 29 | 140,88 | 18 | д | 7 | 19,65 |
| 3 | о | 28 | 134,61 | 19 | ч | 7 | 19,65 |
| 4 | в | 21 | 92,24 | 20 | ш | 7 | 19,65 |
| 5 | а | 20 | 86,44 | 21 | у | 6 | 15,51 |
| 6 | с | 19 | 80,71 | 22 | б | 5 | 11,61 |
| 7 | і | 16 | 64,00 | 23 | й | 5 | 11,61 |
| 8 | л | 16 | 64,00 | 24 | я | 5 | 11,61 |
| 9 | е | 14 | 53,30 | 25 | ї | 4 | 8,00 |
| 10 | н | 14 | 53,30 | 26 | х | 4 | 8,00 |
| 11 | т | 14 | 53,30 | 27 | ж | 3 | 4,75 |
| 12 | ь | 14 | 53,30 | 28 | є | 2 | 2,00 |
| 13 | р | 12 | 43,02 | 29 | ц | 2 | 2,00 |
| 14 | м | 10 | 33,22 | 30 | г | 1 | 0,00 |
| 15 | п | 10 | 33,22 | 31 | ф | 1 | 0,00 |
| 16 | з | 8 | 24,00 | 32 | ю | 1 | 0,00 |
| $M$ | | | | | | 423 | |
| $\sum m\log_2 m$ | | | | | | 1913,54 | |
| $I_\Sigma$ = 1913,54 : 423 | | | | | | 4.52 | |
| $S = \log_2 423 - 4.52$ | | | | | | 4.2 | |
| $R$ = 4.52 : 4.2 | | | | | | 1,08 | |

 ISSN 1562-9945

За таким же принципом проаналізовано ряд російськомовних і україномовних творів. Однак виникла проблема аналізу творів (фрагментів творів), довжина яких менше характеристичної довжини (див. твір Л. Костенко). Запропоновано кілька варіантів вирішення даної проблеми.

Перший підхід полягає в наступному: доповнюємо відсутні символи до характеристичної довжини, вихідним текстом, починаючи з початку твору. Таким чином, зберігаємо структурне співвідношення елементів твору і добиваємося потрібної довжини фрагмента.

Покажемо другий підхід вирішення даної проблеми. Виходячи з того, що автор завершив твір, вважаємо, що його рішення відповідає золотого перетину. Тоді:

$$\frac{L^*}{L} \cong 1.618 \cong \frac{L}{\Delta L}, \tag{7}$$

де L – довжина твору.

Так як L*=4S, отримуємо:

$$L \cong \frac{4^S}{1.618} \cong 0.618 \cdot 4^S. \tag{8}$$

В більш узагальненому вигляді слід використовувати фрактальну розмірність DL, тоді:

$$L = D_L \cdot 4^S, \tag{9}$$

де

$$D_L = \frac{\Delta L}{L} = \frac{L^* - L}{L} = \frac{L^*}{L} - 1. \tag{10}$$

Із виразу (9) знаходимо ентропію S і використовуємо її для розрахунку функцій R і D.

Третій підхід побудований за тим же принципом, що і попередній, тільки в даному випадку фрактальну розмірність розраховуємо з використанням показника Херста за формулою (11), для чого представляємо текст у вигляді сигналу по значенням індексів з таблиці символів Windows.

$$D_L = 2 - H, \tag{11}$$

де H – показник Херста, що отримується із виразу (12).

$$M\left[\frac{R}{S}\right] = \lambda n^H, \tag{12}$$

де M – математичне очікування, R – розмах значень, S – стандартне відхилення, λ – константа, що дорівнює 0,5, n – кількість значень сигналу.

Результати аналізу творів з використанням всіх підходів зібрані в таблиці 3. Аналізуючи отримані дані, було прийнято рішення для подальших досліджень використовувати другий підхід, заснований на розрахунку ентропії твору з урахуванням фрактальної розмірності. Крім того, запропоновано комплексний критерій К, який об'єднує значення функцій R і D:

$$K = \frac{D}{1 + \exp(|1 - R|)}. \qquad (13)$$

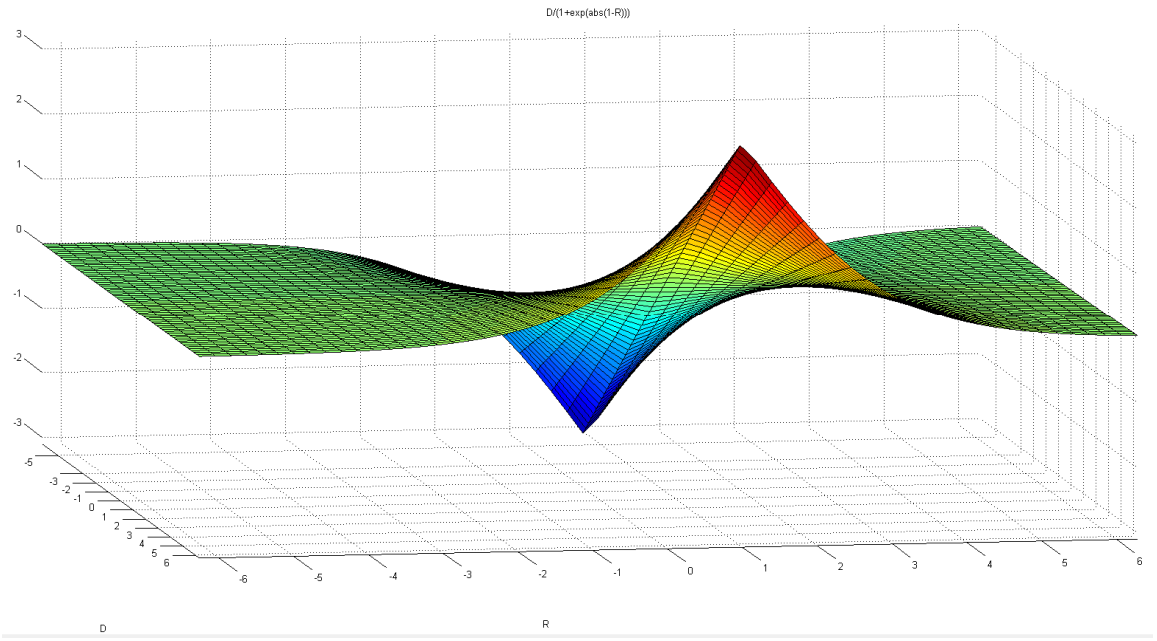


Рисунок 2 – Поверхня значень критерію К(R,D)

Таблиця 3

Зведена таблиця результатів аналізу літературних творів

| № з/п | Твір | *M* | Перший підхід | | Другий підхід | | | Третій підхід | |
|---|---|---|---|---|---|---|---|---|---|
| | | | *R* | *D* | *R* | *D* | *K* | *R* | *D* |
| *Російськомовні твори* | | | | | | | | | |
| 1 | Alekseev – «Чувствую душой» | 880 | 0,997 | 19,406 | 0,831 | 23,963 | **7,269** | 0,996 | 19,157 |
| 2 | А. Пушкин – «Евгений Онегин» | 369* | 0,957 | 19,397 | 0,751 | 21,624 | **6,934** | 0,943 | 17,221 |
| 3 | Грибы – «Тает лед» | 1771 | 1,007 | 19,398 | 0,918 | 21,648 | **6,179** | 0,996 | 19,457 |
| 4 | М. Цветаева – «Мне нравится» | 810 | 1,010 | 19,404 | 0,916 | 20,358 | **5,817** | 0,998 | 18,529 |
| 5 | В. Высоцкий – «Я из дела ушел» | 1272 | 0,991 | 19,403 | 0,916 | 20,324 | **5,810** | 0,966 | 19,112 |
| 6 | П. Гагарина – «Обезоружена» | 1281 | 1,185 | 19,198 | 1,019 | 21,803 | **5,781** | 1,101 | 19,868 |
| 7 | А. Дементьев – «Ни о чем не жалейте» | 782 | 1,002 | 19,403 | 0,911 | 19,674 | **5,643** | 0,954 | 18,740 |
| 8 | В. Маяковский – «Долг Украине» | 1357 | 0,957 | 19,381 | 0,942 | 19,249 | **5,399** | 0,944 | 19,252 |
| 9 | Ф. Достоевский – «Преступление и наказание» | 16495 | 0,967 | 19,397 | 0,963 | 19,402 | **5,362** | 0,965 | 19,348 |
| 10 | Л. Толстой – «Война и мир» | 10140 | 0,979 | 19,401 | 0,975 | 19,294 | **5,284** | 0,976 | 19,283 |
| 11 | А. Пушкин – «Дубровский» | 14050 | 0,997 | 19,403 | 0,996 | 19,299 | **5,206** | 0,993 | 19,327 |
| 12 | А. Розенбаум – «Я родину свою люблю» | 978 | 0,964 | 19,399 | 0,958 | 18,136 | **5,026** | 0,921 | 18,459 |

 ISSN 1562-9945

Продовження таблиці 3

| № з/п | Твір | *M* | Перший підхід | | Другий підхід | | | Третій підхід | |
|---|---|---|---|---|---|---|---|---|---|
| | | | *R* | *D* | *R* | *D* | *K* | *R* | *D* |
| *Україномовні твори* | | | | | | | | | |
| 13 | Т. Шевченко – «Заповіт» | 468* | 1,042 | 20,850 | 0,787 | 24,548 | **7,691** | 0,969 | 19,881 |
| 14 | С. Жадан – «Я знав священика» | 1097 | 1,011 | 20,859 | 0,85 | 25,450 | **7,621** | 0,999 | 20,868 |
| 15 | В. Симоненко – «Лебеді материнства» | 1081 | 1,021 | 20,857 | 0,873 | 24,509 | **7,220** | 1,013 | 20,562 |
| 16 | Л. Костенко – «Страшні слова, коли вони мовчать» | 409* | 1,026 | 20,856 | 0,821 | 20,906 | **6,389** | 0,950 | 17,984 |
| 17 | П. Тичина – «Гаї шумлять» | 407* | 0,953 | 20,847 | 0,791 | 20,019 | **6,245** | 0,886 | 17,863 |
| 18 | В. Стус – «Вдасться чи ні...» | 946 | 1,014 | 20,853 | 0,933 | 20,523 | **5,795** | 0,964 | 19,853 |
| 19 | Леся Українка – «Contra spem spero» | 1512 | 1,021 | 20,857 | 0,974 | 20,573 | **5,640** | 1,001 | 20,002 |
| 20 | І. Котляревський – «Енеїда» (1 ч.) | 1748 | 1,005 | 20,856 | 1,005 | 20,796 | **5,571** | 1,002 | 20,816 |
| 21 | М.Коцюбинський – «Фата моргана» (уривок) | 4921 | 1,049 | 20,845 | 1,030 | 20,949 | **5,511** | 1,044 | 20,648 |
| 22 | І. Драч – «Крила» | 1422 | 1,042 | 20,850 | 1,007 | 19,808 | **5,300** | 0,999 | 19,949 |
| 23 | В. Сосюра – «Любіть Україну, як сонце, любіть» | 1192 | 1,041 | 20,844 | 1,040 | 19,833 | **5,180** | 0,989 | 20,311 |
| 24 | П. Чубинський – «Ще не вмерла Україна» | 742 | 1,022 | 20,857 | 1,000 | 18,117 | **4,871** | 0,957 | 18,723 |
| 25 | І. Франко – «Гімн» | 1138 | 1,198 | 20,68 | 1,172 | 20,460 | **4,840** | 1,177 | 20,458 |

**Висновки.** Структурно-синергетичний аналіз текстів показав, що у всіх досліджених класичних поетичних творів значення R-функції (міра порядку відображеної в хаосі інформації) зі збільшенням довжини тексту, статистично зростають, наближаючись до одиниці. При цьому інтерес також представляє аналіз D-функції (міра розвитку системи – характеристика її прагнення до максимуму інформаційної ємності), яку для розглянутих поетичних творів класиків світового рівня можна розглядати як міру духовності – ступеня практично незмінною в часі інформаційної цінності цих унікальних літературних творів і, що особливо слід підкреслити, написаних на різних мовах. Запропоновано комплексний критерій K, який є достатньо ефективним при аналізі літературних творів (див. таблицю 3).

В цілому, як було показано на ряді прикладів, методи синергетичної теорії інформації є універсальними для структурного аналізу довільних дискретних систем з кінцевою множиною елементів. В прикладному сенсі розглянутий підхід можливо використовувати для виявлення глибинного плагіату з виявленням особливостей стилю написання конкретного автора.

## ЛІТЕРАТУРА

1. Вяткин В.Б. Синергетическая теория информации. Часть 1. Синергетический подход к определению количества информации //Научный журнал КубГАУ [Електронний ресурс]. – Краснодар: КубГАУ, 2008. – №44(10).

2. Царик В.Ю., Михалёв А.И. Методы синергетической теории информации для анализа текстовых данных // Комп'ютерне моделювання та оптимізація складних систем (КМОСС-2017): матеріали III Міжнародної науково-технічної конференції (м. Дніпро, 1-3 листопада 2017 року) / Міністерство освіти і науки України, Державний вищий навчальний заклад «Український державний хіміко-технологічний університет». – Дніпро: ДВНЗ УДХТУ, 2017. – С. 241-243.

3. Чернавский Д.С. Синергетика и информация. Динамическая теория хаоса, М.: Наука, 2001. – 105 с.

4. Шеннон К. Работы по теории информации и кибернетике. – М.: Изд. иностр. лит., 1963. – 830с.

5. Луценко Е.В. Универсальный информационный вариационный принцип развития систем. // Научный журнал КубГАУ [Електронний ресурс]. – Краснодар: КубГАУ, 2008. – №41(07).

6. Яглом А.М., Яглом И.М. Вероятность и информация. – М.: Наука, 1973. – 512с.

 ISSN 1562-9945