СИСТЕМИ ТА МЕТОДИ ОБРОБКИ ІНФОРМАЦІЇ

УДК 861.325

В.О. Хорошко, доктор технічних наук, професор
В.Ю. Артемов, кандидат юридичних наук
Я.В. Невойт

АНАЛІЗ СИМВОЛЬНИХ ІНФОРМАЦІЙНИХ ПОТОКІВ ДЛЯ ВИЗНАЧЕННЯ ІНДИВІДУАЛЬНИХ ОСОБЛИВОСТЕЙ ЛЮДИНИ

Проведено аналіз інформаційних потоків для визначення індивідуальних особливостей людини, на підставі якого можна визначити й ідентифікувати автора, визначити його рідну мову, а отже, і національну приналежність.

Ключові слова: *інформаційні потоки, індивідуальні особливості людини, літера, теорія ергодичності.*

В данной работе проведен анализ информационных потоков для определения индивидуальных особенностей человека, на основании которого можно определить и идентифицировать автора, определить его родной язык, а следовательно, и национальную принадлежность.

Ключевые слова: *информационные потоки, индивидуальные особенности человека, буква, теория эргодичности.*

In this paper the analysis of information flows for the determination of personalities of a man on which you can define and identify the author, his own language and therefore nationality is carried out.

Keywords: *information flow, personalities, letter, theory of stability.*

Сучасні інформаційні технології оперують великим обсягом символьних даних. При цьому процес створення, обробки, передачі та зберігання інформації вимагає в певних випадках проведення її аналізу, діагностики, ідентифікації та визначення авторства. Через розмаїття методів трансформації даних, обсягу потоку і можливостей оператора щодо виділення тієї чи іншої цільової ознаки проведення подібних заходів украй складне. Це обумовлює необхідність удосконалення та пошуку нових шляхів вирішення завдань, які пов'язані з аналізом і обробкою циркулюючого в мережах комунікацій символьного потоку.

Дослідження в цьому напрямі займають особливе місце серед проблем штучного інтелекту, тому що результати в області обробки та аналізу символьного

потоку застосовуються в безлічі експертних систем. На цей час ведуться прикладні дослідження з розпізнавання образів, спрямовані на розвиток і вдосконалення або апаратної частини техніки обробки, або автоматизації сучасних процесів ідентифікації авторства. Однак, спираючись тільки на прикладні аспекти, не завжди можна комплексно і оптимально оцінити всю проблему в цілому. Проблеми передачі, зберігання, обробки та аналізу потоку висувають завдання, метою яких є побудова стисненого подання цілого класу потоків з урахуванням того, що індивідуальні особливості людини вимагають у всьому компактності, інформативності і новизни.

У зв'язку з цим виникає цілий ряд завдань в області лінгвістичного аналізу та закритті інформаційних потоків (криптографії, стеганографії і т.д.):

– формування описів (образів) потоку відповідно до необхідної точності (повноти);

– дослідження однорідності потоків предметної області;

– розробка методів ідентифікації потоку на основі аналізу складових його образів та методів ідентифікації потоків предметної області;

– дослідження і формування критеріїв кластеризації потоків;

– розробка методів і засобів ідентифікації прихованої в потоці інформації в залежності від завдань подальшого перетворення потоку;

– розробка методів і засобів визначення авторства інформації, переданої в потоці;

– аналіз потокових перетворень, які можуть призвести до втрати чи спотворення інформації в предметній області.

Розглядаючи процес інформаційної взаємодії, необхідно відзначити його відмітну особливість – доменну структуру існування замкнутих циклів інформаційного простору, його природну структурованість і дуже точне відображення індивідуальних особливостей людини [1, 2].

Формалізація основних принципів механізму аналізу інформаційного потоку використовує визначення джерела інформації і його закономірностей.

Визначення 1. Джерелом вихідного повідомлення τ називається суб'єкт α інформаційного домену, що генерує скінченну або нескінченну послідовність змінних a_n з алфавіту $A_{\Sigma_1 N}$:

$$\alpha_n \in A_{\Sigma_1 N}, \quad \alpha_n \in A_{\Sigma_1 N},$$

де Σ – множина елементарних одиниць (для символьного потоку – буква).

Визначення 2. Закономірностями суб'єкта α називаються характеристики генеруючої їм інформаційної послідовності τ – алфавітний склад A_α, безліч правил структурування повідомлень S_α і безлічі правил композиції структурних одиниць повідомлень C_α:

$$\alpha=(A_\alpha, S_\alpha, C_\alpha).$$

Упорядкований вираз називається мовою (в окремому випадку – лексикон) суб'єкта α, Σ_α, що є індивідуальною особливістю людини.

У своїх роботах А.А. Марков довів, що в загальній схемі досліджень вони пов'язані в ланцюг. На цій схемі він встановив ряд закономірностей, що поклали початок сучасної теорії марковських процесів.

А.А. Марков розглянув деякі приклади пов'язаних досліджень і показав, що якщо розглядати букви літературного тексту, то ймовірність голосного і приголосного залежить від однієї або двох попередніх літер, які характеризують авторство. Найбільш відомий результат вченого – теорія ергодичності. Проведений ним статистичний аналіз літературних текстів підтвердив наявність у них властивостей однорідності і ергодичності для пов'язаних ланцюгів, що незаперечно визначає автора цих творів. Розглядаємо необмежений ряд досліджень або вимірювань, які відзначаються по порядку номерами

$$1, 2, ..., k, k+1, ...$$

Твердження. Випробування пов'язані в однорідний ланцюг: якщо для одного з них з'явилася подія Г (голосна буква) або протилежна йому П (приголосна буква), то наступні за ним дослідження залежать від цього результату, але не залежать від результатів попередніх його досліджень (у загальній схемі дослідження можуть залежати від кількох попередніх результатів).

Для всього ланцюга визначені одні й ті ж два числа: p_1 і p_2. Число p_1 означає ймовірність події Г при k +1-му дослідженні, якщо дано, що Г з'явилося при k-му дослідженні. Число p_2 означає ймовірність події Г при k +1-му дослідженні, якщо k-п дослідження викликала подія П. Щоб забезпечити висновкам повну визначеність і вміти обчислювати ймовірності подій Г і П при будь-якому дослідженні, слід ввести ще число p^1, що представляють ймовірність події Г при першому дослідженні.

Зауважимо, що у двох літерах алфавіту однорідний ланцюг визначається трьома ймовірностями, тому що ймовірність пар ГП і ПП відповідно дорівнюють $1-p_1$ і $1-p_2$.

Згідно з А.А. Мартиновим [3], який першим вивчив ергодичні властивості ланцюгів, досліджувався ряд чисел

$$p^1, p^2, ..., p^k, p^{k+1}, ...$$

відповідно, представляють ймовірності події Г (бути букві голосною) при кожному з досліджень 1, 2, ..., k, k +1 .... Вводячи додаткові параметри δ і ρ, визначаючи рівності $\delta = p1-p2$, $p_2 = p(1-\delta)$ і вважаючи $|\delta| < 1$, він вивів загальну формулу

$$p^k = p + (p^1 - p)\,\delta^{k-1}, \qquad (1)$$

звідки видно, що число

$$p = \frac{p_2}{1 + p_2 - p^k} \qquad (2)$$

служить межею $p^k \to p$, а $k \to \infty$.

Зауважимо, що збіжність в (1) геометрична і межа р залежить від початкової ймовірності p^1. З теореми ергодичності [3] випливає, що однорідний ланцюг є слабо залежною послідовністю, оскільки вплив значення початкової літери на наступні швидко зменшується при видаленні від початку послідовності.

У своїх дослідженнях А.А. Марков розглянув послідовність двадцяти тисяч літер у романі А.С. Пушкіна "Євгеній Онєгін" (не рахуючи твердого м'якого знаку, м'якого знаку та ком). Вона становить двадцять тисяч залежних досліджень,

кожне з яких дає голосну або приголосну літеру. Проведемо дослідження невеликого фрагменту початку поеми, щоб переконатися в індивідуальних особливостях Пушкінського тексту:

"Мой дядя самых честных правил
Когда не в шутку занемог
Он уважать себя заставил
И лучше выдумать не мог
Его пример другим наука"

Літературний текст поеми А.С. Пушкіна є осмисленою і залежною послідовністю 30 літер російського алфавіту. Двобуквена послідовність голосних (г) і приголосних (п), виділена з цього фрагменту літературного тексту, як ми бачимо, ніякого явного сенсу не має:

пгг пгпг пгпгп пгпппгп ппгпгп
пгпг пг п пгппг пгпгпгп
гп гпгпгп пгпг пгппгпгп
г пгппг пгпгпгп пг пгп
гпг ппгпгп ппгпгп пггпг

Було допущено існування невідомої постійної ймовірності p – бути літері голосною і наближену величину числа p було знайдено зі спостережень, вважаючи число появ голосних літер у 200 послідовностях по 100 літер. Отримане таким чином значення p виявилося рівним 0,4319. Відповідно до методу найменших квадратів підраховано коефіцієнт дисперсії серії досліджень та виявлено, що він значно відрізняється від одиниці, що переконливо підтверджує залежність двобуквеного тексту, і доведено, що математичне очікування коефіцієнта дисперсії дорівнює одиниці.

Крім числа p, було знайдено також зі спостережень наближені величини двох чисел (p_1 і p_2), що представляють ймовірність:

p_1 – голосній літері слідувати за голосною;

p_2 – голосній літері слідувати за приголосною.

При обчисленні ймовірностей p_1 і p_2 підраховується число пар (приголосна–голосна), які при діленні на число всіх голосних в тексті дає для p_1 наближену величину $p_1 = \frac{1104}{8636} \approx 0{,}128$. Подібним чином виходить, що $p_2 = \frac{7534}{11362} \approx 0{,}663$.

Звідси випливає, що ймовірність літері бути голосною чи приголосною залежить від того, яка попередня літера. Знаменник прогресії в (1) дорівнює $\delta = 0{,}535$ і $|\delta| \approx 0{,}001$.

Підставивши значення p_1 і p_2 в (2), отримаємо число 0,4319, що збігається з раніше отриманими наближеним значенням для ймовірності p. Такий збіг двох підрахованих p пояснюється тим, що двобуквений текст відповідає моделі однорідного пов'язаного ланцюга і для неї виконується властивість ергодичності. Через швидку збіжність (1), ймовірність букви бути голосною стає постійною і рівною p по всій довжині тексту, тому підрахована частота голосних літер також виявилася рівною p в (2). У цьому сенсі двобуквений текст є регулярною послідовністю.

Для ергодичного ланцюга справедливий закон великих чисел [4], з якого випливає, що частота голосної букви довгого ланцюжка тексту мало відрізняється від значення p в (2). Закон великих чисел – загальний принцип, через який сумісна дія випадкових факторів призводить при досить загальних умовах до результату, майже незалежного від випадку [4]. Для двобуквеного тексту збіг підрахованої частоти голосної букви з ймовірністю p служить переконливим прикладом дії цього принципу. Іншими словами, двобуквений варіант тексту успадковує властивості зв'язності від свого попередника – 30-літерного осмисленого літературного тексту. Легко помітити, що якщо однорідний ланцюг відразу стартує з початковою імовірністю голосної $p^1 = p$, то ланцюг стаціонарний, тобто ймовірність букви бути голосною дорівнює p в кожній позиції ланцюжка тексту, що є індивідуальною особливістю тієї чи іншої людини.

Зробимо ще одне зауваження. Нехай текст починається з фіксованої букви, наприклад із приголосної, як у поемі А.С. Пушкіна. Тоді з теореми ергодичності випливає, що приблизно через 15 позицій залежність поточної літери тексту від початкової літери зовсім зникає. Така ж ситуація виникає, коли текст починається не з приголосної літери, а з голосної. Подібне дослідження виконано над творами прози в інших авторів, що і підтвердилося.

У літературі застосовується поняття інформаційного потоку, який поділяється на відкритий і закритий потоки.

Відкритий потік підпорядковується семантиці мови суб'єкта Σ_a в поняттях частотної повторюваності елементарних одиниць мови. Для символьного потоку відправною точкою служить алфавіт.

Для прикладу використовуємо чотири найбільш уживані в тексті літери у трьох мовах:

українській – А-0,086; Н-0,064; Р-0,047; Т-0,047;

російській – А-0,062; Н-0,053; Р-0,040; Т-0,053;

англійській – А-0,063; N-0,059; R-0,054; T-0,072.

Закритий потік текстової інформації забезпечений захистом при її зберіганні і передачі комунікаційними мережами. Закриття потоку даних здійснюється криптографічними і стеганографічними методами захисту інформації. Середня частота повторення букв алфавіту в повідомленні російської мови, закритому інформатором алгоритму гамування, становить 0,0303. Проте після його роз'єднання поновлюються мовні особливості відкритого потоку.

Тому на підставі цих досліджень можна визначити національну приналежність і рідну мову автора.

Поведінка однорідного пов'язаного ланцюга відповідно до індивідуальних особливостей людини (ІОЛ) визначається початковим розподілом ймовірностей чотирьох букв: p (A), p (N), p (R), p (T), що становлять у сумі 1, і перехідних ймовірностей, записаних у вигляді матриці:

$$P = \begin{Vmatrix} p(AA) & p(AN) & p(AR) & p(AT) \\ p(NA) & p(NN) & p(NR) & p(NT) \\ p(RA) & p(RN) & p(RR) & p(NT) \\ p(TA) & p(TN) & p(TR) & p(TT) \end{Vmatrix},$$

де суми ймовірностей за рядками також дорівнюють одиниці. Перехідні ймовірності $p(ij)$, $i, j \in \{A, N, R, T\}$ мають такий самий зміст, як і в [5].

Оцінки перехідних ймовірностей p (ij) обчислюються за формулами:

$$\hat{p}(ij) = \frac{m(ij)}{\sum_j m(ij)} \quad (3),$$

де m (ij) – число пар (ij).

Основні труднощі дослідження оцінок полягають в тому, що знаменник у (3) не фіксований, як у випадку бернулієвих незалежних випробувань, і оцінки $\hat{p}(ij)$ є зміщеними. У роботах [5, 6] проведено статистичний аналіз $\hat{p}(ij)$ при $m \to \infty$, де m – довжина ланцюга. Величини $\hat{p}(ij)$ – p (ij) мають граничне нормальне розподілення із середнім 0, дисперсіями

$$\frac{p(ij)[1-p(ij)]}{mp(i)},$$

і для чотирьох різних значень $i \in \{A, N, R, T\}$ ці величини асимптотично незалежні.

Мета А.А. Маркова виявляється простою та економічною моделлю дослідження ІОЛ. Для визначення ймовірностей індивідуальних особливостей необхідно завдання 15 ймовірностей (3 – для початкового стану і 12 – для перехідних ймовірностей). Обчислення показали, що оцінки цих ймовірностей побудовані у вигляді частот стабільні на 23 ІОЛ (обчислювалися також оцінки перехідних ймовірностей в залежності від однієї до чотирьох попередніх літер). Частоти букв А і Т становлять число 0,29, а для букв N і R – 0,21. Дивно, що природа використовувала одну і ту ж схему запису інформації для ІОЛ. Цілком можливо, що окремі особливості могли записуватися з різними перехідними ймовірностями.

У табл. 1 наведено частоти літер і частоти пар букв (оцінки перехідних ймовірностей у ІОЛ).

Таблиця 1

Частоти букв	Частоти пар букв			
A-0,29137	AA-0,32685	AN-0,17234	AR-0,24438	AT-0,25643
N-0,20838	NA-0,34876	NN-0,26056	NR-0,04823	NT-0,34245
R-0,20834	RA-0,28666	RN-0,21137	RR-0,26047	RT-0,24150
T-0,29190	TA-0,21836	TN-0,20497	TR-0,24945	TT-0,32722

Оцінки перехідних ймовірностей для 23 особливостей людини є позитивними і знаходяться в діапазоні (0,005–0,34). Як і у двох літер алфавіту, цей факт означає, що такий ланцюг є ергодичним [3]: при $n \to \infty$ $\hat{p}(ij)$ сходяться до граничних значень (π_A, π_N, π_R, π_T,), незалежних від *i*, та утворюють розподіл ймовірностей $\pi_A + \pi_N + \pi_R + \pi_T = 1$, тут $\hat{p}(ij)^n$, $I, j \in \{A,N,R,T\}$ – елементи матриці $\hat{p}^n$, складеної з оцінок перехідних ймовірностей, причому збіжність також відбувається з геометричною швидкістю.

Обчислення на ЕОМ показали, що для всіх особливостей має місце збіжність величин $\hat{p}(ij)^n$ до граничного вектору (π_A, π_N, π_R, π_T,) приблизно за 50 ітерацій, і що цей вектор мало відрізняється від значення частот букв *A, N , R, T*. Зрозуміло, що така ситуація можлива лише в тому випадку, коли всі ІОЛ описуються однорідним ланцюгом з однаковими перехідними ймовірностями. Тим самим можливість окремої літери з множини {A, N, R, T} досить швидко стає постійною при видаленні від початку ланцюга, тобто текст ІОЛ, як і двобуквений текст Маркова, є слабкозазалежною послідовністю, і для нього виконується закон великих чисел, як і для марківських ланцюгів.

Статистичний аналіз і розрахунки на ЕОМ показують, що ІОЛ записана у вигляді однорідного та ергодичного ланцюга Маркова і для неї виконується закон великих чисел. У чотирибуквеному алфавіті для дослідження ІОЛ потрібно задати 15 параметрів ймовірностей. З такої моделі випливає, що ймовірності окремих літер досить швидко стають постійними при видаленні від початку тексту, що характеризує індивідуальні особливості автора, тобто на підставі цих досліджень можна визначити й ідентифікувати автора і визначити його рідну мову, а отже, і національну приналежність.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *Манухин А.В.* Построение образа сообщения в пределах выбранной языковой пары / А.В. Манухин, С.В. Попов, В.А. Хорошко, Д.В. Чирков // Захист інформації. – 2005. – № 4. – С. 42–51.

2. *Манухин А.В.* Методика образного представления символьного потока потока языков славянской языковой группы / А.В. Манухин, В.А. Хорошко, С.В. Попов // Сб. науч. трудов НАУ, 2005. – С. 226–236.

3. *Марков А.А.* Исчисление вероятностей / А.А. Марков. – М. : Госиздат, 1924. – 592 с.

4. Больших чисел закон. Вероятность и математическая статистика. – М. : Научн. изд-во "Большая Российская Энциклопедия", 1999. – 912 с.

5. *Гупал А.М.* Статистическое оценивание марковской процедуры распознавания / А.М. Гупал, А.А. Вагис // Проблемы управления и информатики. – 2001. – № 2. – С. 62–71.

6. *Манухин А.В.* Методика образного представления символьного потока германской языковой группы / А.В. Манухин, С.В. Попов, В.А. Хорошко, Д.В. Чирков // Вісник ДУІКТ. – 2005. – Т. 3. – № 3–4. – С. 190–193.

Отримано 22.03.2011