Застосування Частотного Критерію Інформативності Ознак в Задачах Інтелектуального Аналізу Тексту Багаторівневого Інформаційного Моніторингу

М.С. Голуб
кафедра інформаційної безпеки та комп'ютерної інженерії
Черкаський державний технологічний університет
Черкаси, Україна
Mas-golub@yandex.ru

The use of the Frequency Criterion Informative Signs in the Tasks Text Mining Multilevel Monitoring of Information

M. Holub
Department of Information Security and Computer Engineering
Cherkasy State Technological University
Cherkasy, Ukraine
Mas-golub@yandex.ru

***Анотація* — Перетворення друкованого тексту в масив чисельних характеристик дозволяє консолідувати інформацію технологією багаторівневого моніторингу. Виявлені особливості перетворення текстової інформації до форми масиву вхідних даних інформаційної системи багаторівневого моніторингу. Запропоновано новий частотний критерій інформативності ознак текстів. Розроблено метод перетворення тексту, за якого частотний критерій інформативності не залежить від розміру вікна. Експериментально доведено, що цей критерій селективний до різних задач класифікації текстів. Випробування критерію інформативності однак проведені в процесі розв'язання задач визначення місця проживання автора друкованого тесту. Експериментально підтверджено можливість та доцільність застосування описаного методу перетворення тексту в технології багаторівневого інформаційного моніторингу.**

***Abstract* — Desribed the information technology of intelligent search information technology systems. To form the deciding rules proposed use multilevel modeling techniques. Experimentally confirmed the possibility and feasibility of the technology to automate the process of finding a given text content. Offered the new frequency criterion informative signs of texts. Created a method of converting text which provided the frequency informative criterion, that does not depend on the size of the window. Experimentally proved that this criterion is selective for different classification text's problems. Criterion is informative but proved by the process of solving problems of determining residence labels of printed text. Experimentally confirmed the possibility and feasibility of the described text-tiered information technology monitoring method.**

Ключові слова—Перетворення тексту, критерій інформативності ознак, масив вхідних даних, інформаційний моніторинг, консолідація інформації, інтелектуальний аналіз тексту, багаторівневий моніторинг

Keywords—Search for text content, multilevel modeling, system engineering, information monitoring, consolidation information, text mining

I. Вступ

Процеси пошуку інформації заданого змісту, пошуку текстів, авторами яких є задані категорії осіб, виявлення текстів, що несуть в собі ворожий для України зміст та виконання багатьох інших завдань є функціями технологій інформаційного моніторингу [**3**].

Інтелектуальний аналіз друкованих текстових повідомлень є дієвим засобом забезпечення інформацією процесів прийняття рішень (ППР). З метою розширення можливостей технологій моніторингу, зокрема інформаційного моніторингу, розв'язуються традиційні завдання інтелектуального аналізу даних – класифікація, структурна та параметрична ідентифікація , прогнозування та інші. Перелік цих завдань та їх поєднання визначається потребами ППР, цілями, що досягає особа, що приймає рішення (ОПР), предметною галуззю, де використовується інформація, яка тримана із друкованих текстовий повідомлень. Найчастіше метою класифікації текстових повідомлень є їх групування за авторством, певними характеристиками стану автора (вік, стать, освіченість, фізичне та психологічне здоров'я, належність до певної спільноти).

Актуальність досліджень, пов'язаних із класифікацією текстових повідомлень та визначенням характеристик їх авторів визначається потребою у протидії методам та засобам інформаційної війни, яка зараз ведеться проти України. Технології інформаційного моніторингу є потужними інструментами, що здатні розв'язувати подібні задачі. Значна популярність моніторингових інформаційних систем (МІС) пов'язана із задоволенням інформаційних потреб військових, бізнесу, народного господарства, медицини, екології та інших галузей.

Інформаційна технологія багаторівневого моніторингу застосовується у випадках, коли необхідно забезпечити процеси прийняття рішень інформацією, яку треба здобувати «по крихтах» із багатьох різнорідних джерел [**6**]. У такому разі застосовується декомпозиція складних завдань до більш простих. Глибина декомпозиції визначає кількість рівнів перетворення інформації і зумовлена потужністю синтезатора. Формується ієрархія локальних завдань із перетворення даних [**3**].

Розв'язання кожного з них отримують у результаті синтезу багатопараметричних моделей. Ієрархічне поєднання цих моделей утворює структуру ГФЗ. На рис. 1 подана структура формування ГФЗ за технологією багаторівневого інформаційного моніторингу [**5**].

На мікрорівні моніторингу відбувається перетворення файлів із різнорідною інформацією від початкової форми тексту, відео- чи аудіо- файлів до форми масиву чисельних характеристик X. На макрорівні синтезуються моделі-класифікатори Y та відбувається їх випробування. На метарівні розробляються процедури використання цих моделей для групування вхідної інформації по класам Z, оцінюється впливовість факторів W.

Однією із складових цієї технології є методи інтелектуального аналізу текстів (Text Mining).

На сьогодні існує кілька класів задач, що розв'язуються методами інтелектуального аналізу текстів. Класифікація і кластеризація текстів [**9**] використовується для здобування ключових слів, реферування, тематичне індексування.

Поєднання в єдину технологію методів класифікації та структурно-параметричної ідентифікації дозволяє розв'язати задачі виявлення та аналізу зв'язків між поняттями, пошуку ключових фраз для навігації по текстам.

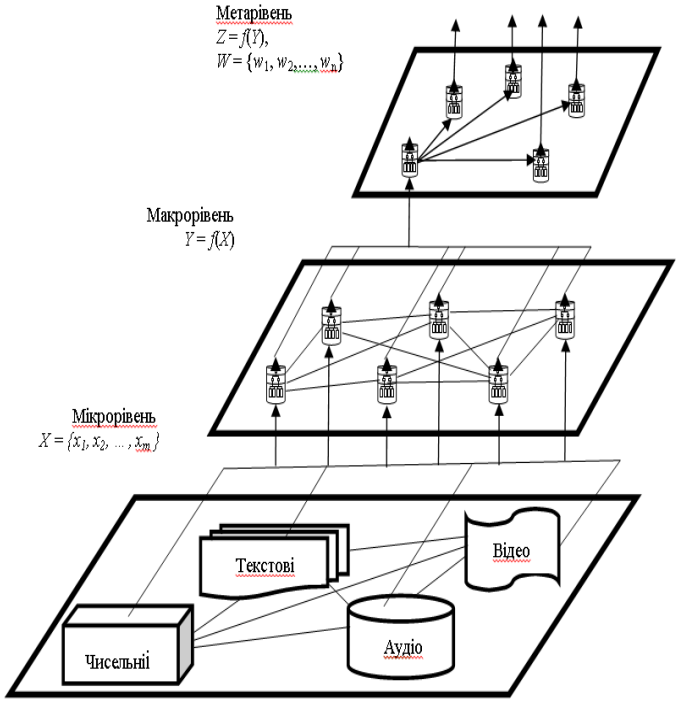

Рис. 1. Структура формування глобальної функціональної залежності системи багаторівневого інформаційного моніторингу

Використання інформаційної технології багаторівневого моніторингу для інтелектуального аналізу текстів (Text mining) дозволяє надати текстовій інформації особливої форми. Вона стає придатною для консолідації із іншими результатами моніторингу та використовується в бізнес-аналітиці, в процесі підтримки прийняття рішень та інше.

В якості характеристик тексту при автоматизованому аналізі використовуються його частотні показники: використання певних частин мови, деяких слів, розділових знаків, фразеологізмів, архаїзмів, слів, які рідко трапляються в тексті, слів іншомовного походження, довжини речення (в словах, складах, знаках) і таке інше [**1**]

II. Постановка задачі досліджень

В цій роботі досліджується застосування частотного критерію інформативності ознак в процесі розв'язання задачі класифікації тексту.

У випадку Text Mining математична постановка завдання має такий вигляд [**7**].

Нехай відомий початковий перелік текстів, що утворюють множину T:

$$T = f(t_1, t_2, \ldots, t_m) \quad (2.1)$$

і перелік властивостей їх авторів, що утворюють множину класів Z:

$$Z = f(z_1, z_2, \ldots, z_n) \quad (2.2)$$

Яка властивість автора відображена в якому тексті відомо для обмеженої кількості елементів навчальної підмножини T^n:

$$T^n = \{(t_1, z_1), (t_2, z_2), \ldots, (t_n, z_n)\} \quad (2.3)$$

Існує невідома цільова залежність – відображення

$$z^* : T \to Z \quad (2.4)$$

значення якої відоме на елементах підмножини Tn. Необхідно побудувати модель

$$a : T \to Z, \quad (2.5)$$

що здатна вірно класифікувати невідомий текст із підмножини $\{t_{n+1}, t_{n+2}, \ldots, t_m\} \in T$, тобто вірно визначити властивості автора цього тексту.

Виявляється залежність значення цього критерію від розміру вікна, відповідно до якого обраховуються числові характеристики вектору ознак МВД.

III. Дослідження частотного критерію інформативності ознак

Значна кількість завдань Text mining виконується МІС шляхом розв'язання типової задачі класифікації точок спостереження. Точка спостереження може поєднувати від одного до кількох вікон. Інколи вона поєднує всі вікна тексту, що досліджується.

Тексти перетворюються до типової форми масиву вхідних даних (МВД) – двовимірної таблиці чисельних показників, що заповнена за певними правилами. Тексти розбиваються на ділянки, що звуться вікнами. Для кожного із вікон розраховуються чисельні характеристики вектору ознак текста. Множина показників, що входять до цього вектору утворюють словник ознак.

Кожна ознака є носієм інформації – відомостей про властивості тексту. Словники ознак визначають інформативність МВД. Тому дослідження, що дозволяють підвищити інформативність МВД, є актуальними.

Наперед відомо, що існує мінімальна кількість інформації в МВД, яка дозволяє існуючими методами та засобами створити корисну модель. Ця мінімальна кількість інформації позначається як межа інформаційної достатності (МІД).

Корисність моделі визначається її здатністю розв'язувати задачі Text mining, зокрема забезпечити класифікацію текстів за заданими класами.

Для забезпечення достатньої інформативності МВД необхідно оцінити інформативність кожного із ознак тексту.

Для оцінки інформативності ознак текстів запропоновано частотний критерій (3.1):

$$K_i = \frac{\gamma_i}{\sum_{i=1}^{n} \gamma_i} \quad (3.1)$$

де K_i – критерій інформативності i-ї ознаки творів автора, що ідентифікується, γ_i – частота використання і-ї ознаки, n – кількість ознак первинного опису

$$\gamma_i = \frac{N_i}{R} \quad (3.2)$$

де N_i – частість використання i-ї ознаки у вікні; R – розмір вікна.

На підставі аналізу виразу (3,1) була сформульована гіпотеза: частотний критерій інформативності ознак (ЧКІО) критерій не повинен залежати від розміру вікна R, оскільки частота використання тексту γ_i присутня і в чисельнику, і в знаменнику. Це може стати однією із вагомих переваг цього критерію інформативності перед дисперсним критерієм.

Для перевірки цієї гіпотези був проведений модельний експеримент.

Досліджувались особливості формування МВД [**8**] та процесу синтезу багатопараметричних моделей [**4**], здатних класифікувати текстові повідомлення за їх належністю до різних типів говірок, що притаманні населенню центральної, північної, південної, західної та східної частин Середньої Наддніпрянщини [**7**].

Розв'язувалась задача класифікації тестів за говірками на території Черкаської області, які використовуються в селах Макіївка та Гарбузин. Підґрунтям дослідження стали діалектні тексти, наведені у збірнику [**2**].

Основною вимогою до МВД є його інформативність. Він повинен містити достатньо інформації для того, щоб побудувати існуючими засобами синтезатора МІС корисну модель а, яка забезпечить необхідні перетворення відповідно до вимоги (6). До ПО заносяться результати першого етапу моніторингу текстового повідомлення – чисельні характеристики інформативних ознак тектсів.

Нижче запропоновано новий метод інтелектуального аналізу тексту з метою виявлення властивостей їх авторів: 1) визначення обсягу вікна; 2) визначення переліку ознак тексту; 3) визначення переліку чисельних характеристик ознак тексту; 4) розрахунок чисельних значень характеристик ознак текстів; 5) задання чисельного значення критерію класу, що поділяє точки спостереження на класи «Свій» та «Чужий»; 6) формування масиву вхідних даних. 7) синтез моделі; 8) дослідження моделі. Визначення розділяючої поверхні. 9) випробування моделі; 10) оцінка впливовості факторів.

Одними із головних завдань, що визначає успішність Text Mining – це декомпозиція тексту і визначення переліку його інформативних характеристик – ознак, що дозволять синтезатору адекватно і точно відобразити властивості автора в моделі текстового повідомлення.

В процесі дослідження застосовувати декомпозицію тексту на окремі ділянки – вікна та на окремі елементи – букви, буквосполучення, речення та інші, які на наступних етапах використовуються для розрахунку значень показників. [4].

Досліджувались значення критерію інформативності (3.1) при зміні розміру вікна в напряму від найбільшого 5000 знаків, до найменшого – 50 знаків.

Тексти перетворювались в первинний опис об'єкта (ПО) досліджень шляхом адаптивного формування переліку ознак із МІД = 1. Таким чином формувався словник із повного переліку показників, що потенційно можуть бути інформативними, і які хоча б один раз зустрічались в тексті.

Після цього ПО піддавався перетворенню шляхом розрахунку значення критерію (3.1) для вікон різного розміру. Після заміни частотних характеристик тексту на значення частотного критерію інформативності проводилось їх усереднення Таким чином ПО перетворюється в МВД. На рис. 1 і 2 подані результати досліджень.

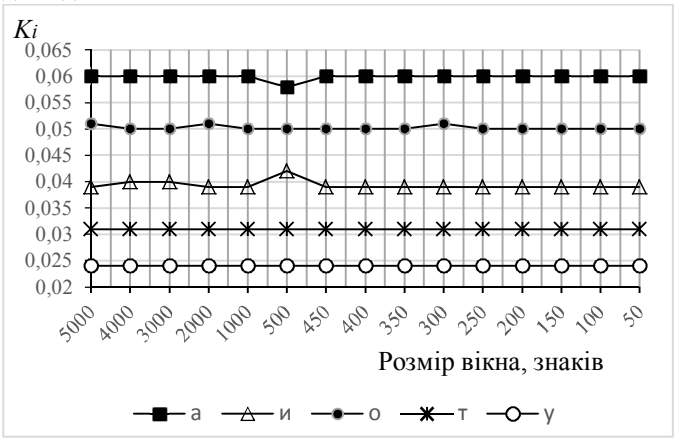

Рис. 2. Інформативність ознак текстів мешканців с. Гарбузин

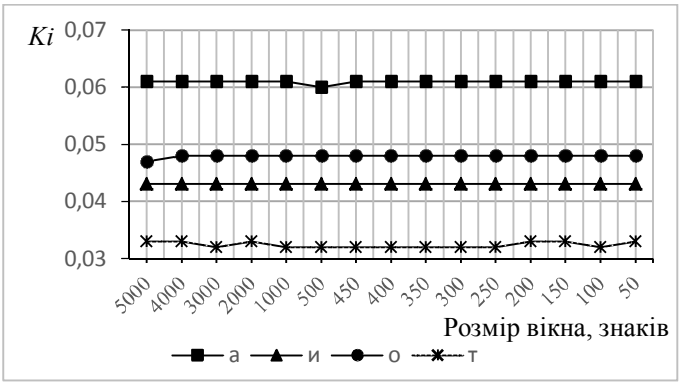

Рис. 3. Інформативність ознак текстів мешканців с. Макіївка

Значення частотного критерію інформативності (3.1) не залежить від розміру вікна після їх усереднення як для текстів с. Гарбузин, так і для текстів с. Макіївка. При цьому словник текстів с. Гарбузин на 1 інформативну ознаку більше в порівнянні із текстами с. Макіївка. Значення одних і тих же інформативних ознак для різних сіл майже завжди різняться, що доводить їх селективність.

IV. Висновки

Частотний критерій інформативності ознак друкованого тексту може бути використаний для перетворення первинного опису об'єкта в масив вхідних даних. При цьому експериментально підтверджено гіпотезу про незалежність значення критерію інформативності від розміру вікна за умови усереднення результатів. Доведено також, що ЧКІО селективний до класів. При аналізі інформативності ознак текстів повідомлень мешканців різних селищ значення ЧКІО різняться. Таким чином частотний критерій може бути використано в процесі розв'язанні задач інтелектуального аналізу текстів в технологіях багаторівневого інформаційного моніторингу.

Література References

[1] E.Y. Lynhvystycheskaya bezopasnost' rechevoy kommunykatsyy // HLƏDYS. URL: http://www.rusexpert.ru/maga¬zine/034.htm

[2] Hovirky Cherkashchyny: zbirnyk dialektnykh tekstiv / Uporyadnyky H.I. Martynova, T.V. Shcherbyna, A.A. Taran. – Cherkasy: PP Chabanenko Yu.A., 2013. – 870 s.

[3] S.V.Holub, Bahatorivneve modelyuvannya v tekhnolohiyakh monitorynhu otochuyuchoho seredovyshcha. Cherkasy: Vyd. vid. ChNU imeni Bohdana Khmel'nyts'koho, 2007. – 220 s.

[4] S.V. Holub, Vidobrazhennya vlastyvostey avtora tekstu v strukturi bahatoparametrychnoyi modeli / S.V. Holub, O.V. Konstantynovs'ka, M.S. Holub // Systemy obrobky informatsiyi: Zbirnyk naukovykh prats'. – Kh.: Kharkivs'kyy universytet povitryanykh syl imeni Ivana Kozheduba, 2014. – Vyp. 9 (125). – S. 82-87

[5] S.V. Holub, Vidobrazhennya konsolidovanoyi informatsiyi ekonomichnykh pokaznykiv rehionu u strukturi bahatorivnevykh modeley / S.V. Holub, N.O. Khymytsya // Visnyk Skhidnoukrayins'koho natsional'noho universytetu imeni Volodymyra Dalya. – 2012. – # 8 (179). – Chastyna 1. – S. 122-128.

[6] S.V.Holub, Konsolidatsiya modeley v protsesi bahatorivnevoho opratsyuvannya danykh / S.V. Holub Informatsiya, komunikatsiya, suspil'stvo 2014: materialy 3-yi Mizhnar. nauk. konferentsiyi ICS-2014. – L'viv: Vydavnytstvo L'vivs'koyi politekhniky, 2014. – S. 162-163.

[7] S.V. Holub, Modelyuvannya dialektnoho tekstu v tekhnolohiyi bahatorivnevoho informatsiynoho monitorynhu / S.V. Holub, H.I. Martynova, M.S. Holub // Matematychni mashyny i systemy. – 2016. – # 4. – S. 76-83

[8] S.V. Holub, Formuvannya pokaznykiv masyvu vkhidnykh danykh dlya identyfikatsiyi avtorstva tekstovykh povidomlen' / S.V. Holub, O.V. Konstantynovs'ka, M.S. Holub // Systemy obrobky informatsiyi: zb. nauk. prats'. – Kh.: Kharkivs'kyy universytet Povitryanykh syl imeni Ivana Kozheduba, 2014. – Vyp. 2 (118). – S. 89 – 92

[9] V.L. Pleskach, T.H. Zatonats'ka, Informatsiyni systemy i tekhnolohiyi na pidpryyemstvakh. K.: Znannya, 2011. – 718 s.