Формування Масиву Вхідних Даних для Класифікації Друкованих Текстів в Технології Багаторівневого Інтелектуального Моніторингу

Марія Голуб
кафедра інформаційної безпеки та комп'ютерної інженерії
Черкаський державний технологічний університет
Черкаси, Україна
mas-golub@yandex.ua

Formation of Input Data Array for Classification of Printed Texts in Multilevel Tintelligent Monitoring Echnolog

Maria Holub
dept. of Information Security and Computer Engineering
Cherkasy State Technological University
Cherkasy, Ukraine
mas-golub@yandex.ua

***Анотація* — В роботі описано процес розв'язання однієї із задач інтелектуального аналізу даних - класифікації текстових повідомлень. Подані результати досліджень процесу перетворення друкованого тексту до типової форми масиву вхідних даних синтезатора моделей. Синтезатор моделей є елементом моніторингової інтелектуальної системи. Він використовується для синтезу моделей-класифікаторів. Вони дозволяють згрупувати тексти за заданими вимогами. За кількістю вірно класифікованих текстів оцінюють ефективність результатів удосконалення процесу формування ознак та побудови точок спостереження. Запропоновано використати в якості показника інформативності ознаки ймовірність її викоирстання у окремому вікні тексту. Експериментально підтверджено доцільність використання такого показника інформативності ознаки.**

***Abstract* — The paper describes the process of solving one of the tasks of the intellectual analysis of data - the classification of text messages. The results of researches of the process of conversion of the printed text into the typical form of the input array of model synthesizer are given. The simulator of the models is an element of the monitoring intellectual system. It is used for the synthesis of modeling classifiers. They allow you to group texts according to the given requirements. The number of correctly classified texts evaluates the effectiveness of the results of improving the process of forming signs and building points of observation. It is suggested to use as an indicator of the informativi of a sign the likelihood of its occurrence in a separate text box. The expediency of the use of such indicator of the informative nature of the sign has been experimentally confirmed.**

Ключові слова— класифікація, текст, модель, МГУА

Keywords— classification, text, model, GMDH

I. Вступ

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 залежить від потреб 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. Одним із потужних інструментів, що здатні розв'язувати подібні задачі, є технології інформаційного моніторингу. Значна популярність 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 [1]. 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 [2] 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. На сьогодні не до кінця усвідомлені можливості використання методів прогнозування в процесі інтелектуального аналізу текстів.

Багаторівневий інформаційний моніторинг реалізується шляхом формування глобальної функціональної залежності – функціонала, який перетворює результати різнорідних спостережень у відомості про властивості об'єктів. Властивості описують здатність об'єкта реагувати на зовнішні впливи. Ці властивості враховуються в процесі формування управлінських рішень — послідовності керуючих впливів, що забезпечують перехід об'єкта із даного стану в запланований. Природньо, що кожна МІС будується під індивідуальні потреби особи, що приймає рішення (ОПР).

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].

Результати класифікації текстів залежать від інформативності МВД. Носієм інформації є ознаки. Вдале формування словника ознак забезпечує можливість існуючими інструментами синтезатора моделей побудувати вирішуюче правило у вигляді поліноміальної або іншої подібної моделі, яка дозволяє максимально адекватно класифікувати тексти. При недостатній інформативності МВД збільшують кількість інформативних ознак шляхом обробки текстів до їх перетворень або після них. Одним із засобів підвищення інформативності МВД є перетворення текстів за словником ознак і побудова точок спостереження шляхом обробки частотних характеристик показників із словника текстових повідомлень. Недостатньо дослідженими залишаються процеси побудови точок спостережень. Невирішеними залишаються завдання виявлення впливу процесу побудови точок спостереження на результати класифікації текстів.

В цій роботі подані результати досліджень, метою яких є доведення ефективності нового підходу до побудови технологій багаторівневого інформаційного моніторингу – поєднання багаторівневого моделювання моніторингових інформаційних систем (МІС) та методів глибинної декомпозиції друкованих текстів для формування словника інформативних ознак та розрахунку їх чисельних характеристик, що входять до масиву вхідних даних (МВД) [4].

II. ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

Для виявлення властивостей автора текстового повідомлення необхідно розв'язати задачу класифікації. У випадку Text Mining математична постановка завдання має такий вигляд.

Нехай відомий початковий перелік текстів, що утворюють множину T:

$$T = f(t_1, t_2, \ldots, t_m) \quad (1)$$

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

$$Z = f(z_1, z_2, \ldots, z_n) \quad (2)$$

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

$$T^n = \{(t_1, z_1), (t_2, z_2), \ldots, (t_n, z_n)\} \quad (3)$$

Існує невідома цільова залежність – відображення

$$z^* : T \to Z \quad (5)$$

значення якої відоме на елементах підмножини Tn. Необхідно побудувати модель

$$a : T \to Z, \quad (6)$$

що здатна вірно класифікувати невідомий текст із підмножини $\{t_{n+1}, t_{n+2}, \ldots, t_m\} \in T$.

III. РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ

Досліджувався процес класифікації текстів, який передбачає:

1) декомпозицію текстів на ділянки однакової довжини, що звуться вікнами;

2) формування словника ознак;

3) розрахунок показника інформативності ознак;

4) формування точок спостереження у багатовимірному просторі ознак. Точка спостереження має вигляд строчки у матриці чисельних характеристик (ознак) тексту;

5) формування масиву вхідних даних шляхом поєднання окремих точок спостереження різних текстів у класи;

6) побудова правила групування точок спостереження у класи. Правило групування має вигляд моделі-класифікатора, що отримується шляхом машинного навчання нейромереж, синтезу поліноміальних моделей МГУА та інших еволюційних методів;

7) побудови інформаційної технології класифікації текстів шляхом поєднання методів визначення інформативності ознак, відбір та поєднання у словники сукупності інформативних ознак, методів формування точок спостереження різних текстів та поєднання їх в масив вхідних даних, методів синтезу моделі-класифікатора, методів формування сукупності моделей-класифікаторів та одночасного їх використання в процесі інтелектуального аналізу текстів.

Крім того інформаційна технологія класифікації текстів повинна містити результати розв'язання задачі координації – узгодження взаємодій різних методів, що забезпечує максимальну кількість вірно класифікованих текстів та їх частин.

Для кожного із етапів цієї ІТ необхідно знайти спільну стратегію розв'язку задач параметричної оптимізації локальних процесів перетворення інформації. Параметрична оптимізація кожного із локальних процесів повинна покращувати результати глобального процесу класифікації тексту – збільшувати кількість вірно класифікованих текстів.

Була сформульована гіпотеза про те, що адаптивність МВД забезпечується шляхом оптимізації довжини вектору ознак $\vec{x}$, довжини вікна, виявлення переліку ознак необхідної інформативності, підвищення інформативності масиву вхідних даних вцілому шляхом побудови точок спостереження,. Інформативність ознаки тим вища, чим частіше ця ознака використовується у тексті.

Для розрахунку інформативності окремої ознаки в цій роботі застосовувався імовірнісний критерій [5]:

$$K_i = \frac{\gamma_i}{\sum_{i=1}^{n} \gamma_i} 100\%, \qquad (7)$$

де K_i – показник інформативності і-ї ознаки; γ_i- частість і-ї ознаки (кількість разів, що використана і-та ознака у окремому вікні), n – кількість ознак у МВД.

Для експериментальної перевірки цієї гіпотези розв'язувалась задача класифікації текстів за гендерною ознакою автора. Було задано 2 класи: 1- жінки; 2- чоловіки. В якості АСМ використовувався багаторядний алгоритм МГУА [6].

В процесі планування експерименту за критерій якості моделі використовувався показник кількості вірно розпізнаних вікон у тексті. Було сплановано двофакторний експеримент. Досліджувався вплив зміни розміру вікна та мінімальної інформативності ознак, що були відібрані із словника, на результати класифікації текстів. Частість застосування кожної ознаки у одному вікні утворюють строчку в МВД – точку спостереження у багатовимірному просторі ознак. Кількість вікон, перелік ознак та їх частість дозволяють сформувати МВД.

Досліджувались 20 текстів, отриманих із журналістських інтернет-публікацій.

За результатами цього дослідження виявлені залежності зміни кількісті вірно класифікованих точок спостереження при зростанні інформативності ознаки, що розрахована за критерієм (7) при декомпозиції тексту на вікна різної довжини. Закономірності зміни кількості вірно класифікованих точок спостереження для вікон різної довжини різні.

Зростання значення показника інформативності приводить до збільшення кількості вірно класифікованих точок тільки на окремих ділянках. Результати класифікації точок спостереження, отримані для значень показника інформативності 5% і 7% дозволяють стверджувати, що підвищення індивідуальної інформативності ознаки не завжди дозволяє отримати підвищення інформативності всього масиву. При зростанні інформативності ознак кількість вірно класифікованих точок зменшується. Це може бути спричинено зростанням впливу на результат моделювання суміщених ознак при зростанні їх інформативності [7] та впливом факторів, які не ввійшли до плану експерименту.

При значенні показника інформативності ознак 1,5% і 3% в умовах експерименту вдалось отримати максимальну кількість вірно розпізнаних точок за умови, що довжина вікна буде 2000 знаків. Але при цьому при застосуванні переліку ознак, що мають імовірність застосування у вікні 1,5 % їх кількість – 151, а при застосуванні переліку ознак із 3% імовірністю застосування, їх кількість 34, тобто зменшується більше ніж в 4 рази. Це дозволяє зменшити кількість комп'ютерних ресурсів, зокрема часу, на побудову окремої моделі, підвищивши таким чином ефективність методу. Оскільки параметрична оптимізація в цій технології реалізується шляхом багаторазового синтезу та випробування моделей, зменшення часу синтезу окремої моделі є показником значимим.

Аналізувались залежності зміни кількості вірно класифікованих точок від довжини вікна тексту при різних показниках імовірності застосування ознаки у вікні, що є показником інформативності цієї ознаки. Виявлено, що оптимальною довжиною вікна для розв'язку задачі класифікації текстів за гендерною ознакою в цих умовах доцільно вважати 2000 знаків. Отриманий результат у 100% вірно класифікованих точок спостереження дозволяє класифікувати текст за однією точкою.

Для підвищення інформативності МВД була сформульована гіпотеза, що закономірність зміни значення дисперсії ознаки при зростання кількості елементів, за якою ця дисперсія розрахована є індивідуальною характеристикою кожного класу.

Для перевірки цієї гіпотези був проведений експеримент.

Точка спостереження формувалась шляхом розрахунку дисперсії ознаки у фіксованій кількості вікон. ПО тексту формувався шляхом розрахунку частості використання ознаки у кожному вікні. Послідовне поєднання вікон утворює масив чисельних характеристик. Запропоновано розбити масив даних ПО на ділянки по z вікон (строчок) і розрахувати дисперсію кожної ознаки в цих z вікнах. Результати розрахунків утворюють точку спостереження МВД. Кількість точок спостереження у МВД буде в z разів менше в порівнянні з первинним описом. Після цього розв'язується типова задача синтезу моделі-класифікатора, визначення розділяючої поверхні для класів, та його випробування на тестовій послідовності точок. Тестова послідовність точок спостереження не використовувалась в процесі синтезу моделі-класифікатора.

Для формування МВД була проведена декомпозиція текстів на вікна з розміром 500 знаків. Всі тексти були розділені на 2 класи: 1 – Жінки, 2- Чоловіки. Для дослідження вибрані тексти авторів, що були задіяні в попередніх дослідженнях.. Моделі-класифікатори синтезувались за багаторядним алгоритмом МГУА [6].

Перевага нового методу формування точок спостереження в таблиці 2 розраховувалась за формулою:

$$x_3 = \frac{(x_2 - x_1)}{x_1} 100\%, \qquad (6)$$

де x_3 – відносна перевага нового методу, x_2 – нормована до 100% кількість вірно класифікованих вікон із існуючим методом формування точок, x_1 – нормована до 100% кількість вірно класифікованих вікон із новим методом формування точок.

Результати досліджень дозволяють стверджувати, що перевага нового методу побудови точок спостереження забезпечується на кожному рівні експерименту і складає від 19,29% до 161,51%. Це означає, що сформульована гіпотеза експериментально підтверджена.

ВИСНОВКИ

Підвищення інформативності масиву вхідних даних при розв'язання задачі класифікації текстів досягається шляхом параметричної оптимізації процесу формування МВД та обробкою результатів моделювання кількох ділянок текстів.

Запропоновано новий метод класифікації текстових повідомлень, що передбачає формування словника інформативних ознак, декомпозицію тексту на ділянки однакової довжини, перетворення тексту на масив його характеристик, побудову моделей-класифікаторів, випробування цих моделей на текстах, що не використовувались при їх створенні. На відміну від існуючих методів запропоновано для кожної задачі формувати індивідуальний перелік інформативних ознак і індивідуально підбирати довжину вікон – ділянок, на які розбиваються тексти перед перетворенням.

Для масивів вхідних даних із недостатньою інформативністю запропоновано удосконалити новий метод класифікації текстів шляхом застосування процедури обробки результатів моделювання кількох вікон, на основі яких формуються точки спостереження в МВД. Новий метод побудови точок спостереження переважає існуючий в цих умовах дослідження на (19,29 – 161,51)%.

Недослідженими залишаються випадки, коли система не забезпечує надійного розв'язку поставленої задачі, наприклад на передостанньому рівні експерименту із мінімальною інформативністю ознак 5%.

Література References

[1] Голуб С.В. Консолідація моделей в процесі багаторівневого опрацювання даних / С.В. Голуб Інформація, комунікація, суспільство 2014: матеріали 3-ї Міжнар. наук. конференції ICS-2014. – Львів: Видавництво Львівської політехніки, 2014. – С. 162-163.

[2] Плескач В.Л., Затонацька Т.Г. Інформаційні системи і технології на підприємствах. К.: Знання, 2011. –718 с..

[3] Голуб С.В. Багаторівневе моделювання в технологіях моніторингу оточуючого середовища. Черкаси: Вид. від. ЧНУ імені Богдана Хмельницького, 2007. – 220 с..

[4] Голуб С.В. Відображення властивостей текстового документа в багатовимірних моделях інформаційних систем комп'ютерного моніторингу / С.В. Голуб // Інформація, комунікація, суспільство 2013: Матеріали 2-ї Міжнародної наукової конференції ICS-2013. – Львів: Видавництво Львівської політехніки, 2013. – С. 180-181.

[5] Голуб С.В. Формування показників масиву вхідних даних для ідентифікації авторства текстових повідомлень / С.В. Голуб, О.В. Константиновська, М.С. Голуб // Системи обробки інформації : збірник наукових праць. – Х.: Харківський університет Повітряних сил імені Івана Кожедуба, 2014. – Вип. 2 (118). – С. 89-92..

[6] Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем / Ивахненко А.Г. – К.: Наукова думка, 1981. – 296 с..

[7] Голуб С.В. Зниження суміщеності сигналів в методах синтезу індуктивних моделей / С.В. Голуб // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2007. – № 1(29). – С.150-152.