УДК 004.942

С.В. ГОЛУБ[*], Г.І. МАРТИНОВА[*], М.С. ГОЛУБ[**]

МОДЕЛЮВАННЯ ДІАЛЕКТНОГО ТЕКСТУ В ТЕХНОЛОГІЇ БАГАТОРІВНЕВОГО ІНФОРМАЦІЙНОГО МОНІТОРИНГУ

[*]Черкаський національний університет імені Богдана Хмельницького, Черкаси, Україна
[**]Черкаський державний технологічний університет, Черкаси, Україна

Анотація. *У статті наведено результати досліджень процесів моделювання діалектних текстів у рамках інформаційної технології багаторівневого моніторингу. Запропоновано новий метод класифікації текстових повідомлень за місцем проживання їхніх авторів. Класифікаційні ознаки отримано після декомпозиції текстів та розрахунку їхніх частотних характеристик. Для синтезу моделей використано багаторядний алгоритм МГУА. Кількість правильно класифікованих текстів складає від 78% до 100%. Перетворення текстових повідомлень у масив вхідних даних дозволяє використати переваги методів багаторівневого моделювання в технологіях інформаційного моніторингу текстових повідомлень.*
Ключові слова: *діалектні тексти, моделювання, класифікація, інформаційний моніторинг.*

Аннотация. *В статье приведены результаты исследований процессов моделирования диалектных текстов в рамках информационной технологии многоуровневого мониторинга. Предложен новый метод классификации текстовых сообщений по месту проживания их авторов. Классификационные признаки получены после декомпозиции текстов и расчета их частотных характеристик. Для синтеза моделей использовался многорядный алгоритм МГУА. Количество верно классифицированных текстов находится в пределах от 78% до 100%. Преобразование текстовых сообщений в массив входных данных позволяет использовать преимущества методов многоуровневого моделирования в технологиях мониторинга текстовых сообщений.*
Ключевые слова: *диалектные тексты, моделирование, классификация, информационный мониторинг.*

Abstract. *The results of research processes modeling of dialect texts within the multi- level information monitoring technology are regarded in the article. A new method of classifying text messages at the residence of their authors is proposed. The signs for classification got after decomposition of text and calculating their frequency characteristics. For synthesis models used GMDH. The quantity correctly classified texts from 78% to 100%. Convert text messages in an array input allows the advantage of multi-level modeling techniques in information technology monitoring text messages.*
Keywords: *dialect text, modeling, classification, information monitoring.*

1. Вступ

Процеси визначення характеристик автора друкованого тексту набувають особливої актуальності в сучасних умовах інформаційної війни. Крім того, ці завдання є традиційно актуальними в криміналістиці. Інтелектуальний аналіз діалектних текстів дозволяє виявити найбільш значимі властивості авторів та відобразити їх у структурі багатопараметричних моделей. Ці моделі розв’язують слабоформалізовані завдання класифікації текстів за властивостями авторів, виконуючи функції вирішуючих правил.

Потреба в обробці великих обсягів текстової інформації спричиняє застосування моніторингових інформаційних технологій. Складність цих завдань зумовила створення методів та засобів багаторівневого моделювання [1].

Технологія багаторівневого моделювання [1] передбачає можливість консолідації інформації на вищих рівнях глобальної функціональної залежності (ГФЗ), отриманої не тільки

за результатами моніторингу текстів, але й з інших різнорідних джерел нижніх рівнів. Зокрема, ієрархічно поєднані у ГФЗ також моделі, синтезовані на основі таблиць чисельних характеристик економічного, соціоекологічного, медичного та інших станів об'єкта. Разом з іншими виявленими характеристиками особи результати таких досліджень містять важливу інформацію, використану у процесі підтримки прийняття необхідних рішень.

Для синтезу моделей, що відображають у собі властивості об'єктів моніторингу, чисельні характеристики перетворюють у таблиці Бази даних первинного опису (ПО), а потім, після оцінки інформативності цих характеристик, ПО перетворюють у масиви вхідних даних (МВД). На основі МВД відбувається синтез моделей об'єктів моніторингу. Набір алгоритмів синтезу моделей (АСМ) та правила їх використання утворюють окрему підсистему, яка отримала назву «Синтезатор» [1]. Основою АСМ стали індуктивні методи [2], нейромережі різноманітних типологій, генетичні та гібридні алгоритми. Необхідним атрибутом наших синтезаторів є технологія багатошарового моделювання.

Моніторингові інформаційні системи (МІС), як правило, забезпечують кілька типів технологій моніторингу, які різняться між собою процесами формування ПО та забезпечення інформативності показникам, що утворюють МВД. На етапі ж синтезу моделей розв'язані типові завдання ідентифікації функціональних залежностей, класифікації, розпізнавання образів, прогнозування та ін. Технологія багаторівневого моделювання передбачає поєднання множини моделей, здатних розв'язувати різнотипні завдання, в єдину структуру ГФЗ.

Технології інформаційного моніторингу обробляють текстові, відео- та аудіофайли. Тому методи формування МВД при реалізації цих технологій містять додаткові етапи перетворення тексту, звуку чи відео до типової форми двовимірного масиву чисельних характеристик об'єктів моніторингу.

У статті розглянуто методи інтелектуального аналізу текстових повідомлень, методи формування масиву інформативних ознак тексту, їхні зв'язки з методами синтезу моделей з метою виявлення місця проживання автора діалектного тексту.

2. Аналіз останніх досліджень і публікацій

Технології багаторівневого моніторингу використовують для забезпечення процесів прийняття рішень у тому випадку, коли складність завдань із перетворення інформації переважає можливості методів і засобів їхнього розв'язання. У такому разі [1] 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. 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 [3].

На мікрорівні моніторингу відбувається перетворення файлів із різнорідною інформацією від початкової форми тексту, відео- чи аудіофайлів 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 за класами Z, оцінюється впливовість факторів W.

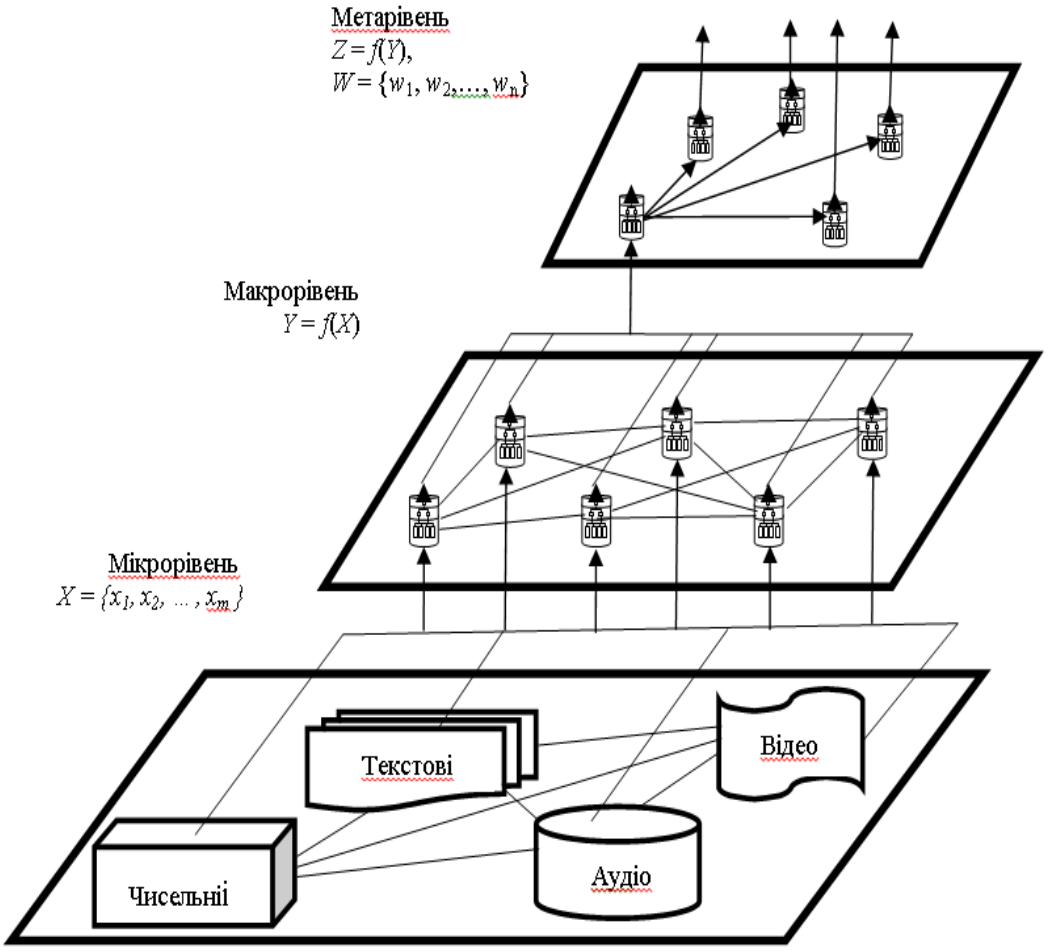

Рис. 1. Структура формування глобальної функціональної залежності системи багаторівневого інформаційного моніторингу

Для формування ПО діалектних текстів на мікрорівні доцільно використати вже існуючі методи та засоби інтелектуального аналізу текстів, зокрема, ті, що дозволяють профілювати [4] їхніх авторів. Ці методи повинні працювати в межах технології багаторівневого моніторингу [1].

Найбільш ефективними засобами автоматизації процесів виявлення характеристик авторів друкованих текстів є методики профілювання текстів, в яких використано методи статистичного моделювання. На думку авторів, одним із найбільш вдалих прикладів використання такого підходу є серія робіт Т.А. Литвинової, зокрема, робота [5]. Застосування регресійно-кореляційного аналізу дозволило отримати множину моделей, що уможливлюють виявлення статі автора, оцінку рівня самоконтролю, емоційної врівноваженості, практичності.

Модель отримано у вигляді лінійної регресії, що реалізовувала функціональну залежність

$$Y = f(x_1, x_2, \ldots, x_n), \tag{1}$$

де n – кількість факторів, що містить модель, $x_1, x_2, \ldots, x_n$ – характеристики факторів, що впливають на результат Y.

За допомогою критерію Пірсона автор виявила значимий перелік факторів. Привертає увагу й те, що, не зважаючи на постановку завдання ідентифікації функціональної залежності, автор розв'язувала завдання класифікації. До того ж відомо, що критерій кореляції Пірсона сформульовано, ґрунтуючись на гіпотезі про нормальність закону розподілу вхідних даних. Із тексту статті [5] незрозуміло, чи здійснювала автор оцінку виду закону розподілу випадкових величин, адже характеристики аналізованих текстів є величинами випадковими. З огляду на це, отримані моделі можуть бути нестійкими. До того ж Т.А. Литвинова досліджує тексти, написані російською мовою. Праць, що стосуються вивчення україномовних текстів, не виявлено.

3. Мета статті

Метою статті є розробка нового методу класифікації текстів за говірками їх авторів шляхом поєднання процедур перетворення тексту в масив чисельних характеристик та побудови вирішуючого правила у вигляді багатопараметричної моделі-класифікатора для виконання завдання виявлення місця проживання автора текстового повідомлення. Крім того, необхідно було розробити механізм підтвердження ефективності вже застосованого лінгвогеографічного методу [6] класифікації говірок, записаних від авторів текстових повідомлень, що проживають на території Черкаської області.

Таким чином необхідно автоматизувати процес класифікації текстових повідомлень. Це завдання слабоформалізоване, оскільки успішно його виконати за допомогою переліку заданих ознак із однозначно визначеними чисельними характеристиками не вдається. Математичне формулювання завдання набуває такого вигляду.

Нехай відомий початковий перелік текстів, що утворюють множину T:

$$T = f(t_1, t_2, \ldots, t_m), \tag{2}$$

де m – кількість текстів, що піддаються дослідженню, і перелік типів говірок їх авторів, що утворюють множину класів Z:

$$Z = f(z_1, z_2, \ldots, z_n), \tag{3}$$

де n – кількість говірок, якими користується населення заданого регіону.

Якою говіркою автора написаний який текст відомо для обмеженої кількості елементів навчальної підмножини T_n:

$$T_n = \{(t_1, z_1), (t_2, z_2), \ldots, (t_n, z_n)\}. \tag{4}$$

Існує невідома цільова залежність – відображення

$$z^* : T \to Z \tag{5}$$

значення якої відоме на елементах підмножини T_n. Необхідно побудувати модель

$$a : T \to Z, \tag{6}$$

що здатна вірно класифікувати невідомий текст із підмножини $\{t_{n+1}, t_{n+2}, \ldots, t_m\} \in T$, тобто вірно визначити тип говірки автора цього тексту, і, відповідно, місце його проживання.

4. Результати досліджень

Була сформульована гіпотеза про те, що вирішуюче правило необхідно будувати у вигляді індуктивної моделі за багаторядним алгоритмом МГУА. У випадку, коли завдання автоматизації процесу класифікації текстових повідомлень за допомогою багатопараметричної індуктивної моделі буде успішно виконане, то це слугуватиме підтвердженням ефективності лінгво-географічного методу, поданого в [6].

Досліджено особливості формування масиву вхідних даних (МВД) [7] та процесу синтезу багатопараметричних моделей [8], здатних класифікувати текстові повідомлення за їх належністю до різних типів говірок, притаманних населенню центральної, північної, південної, західної та східної частин Середньої Наддніпрянщини.

Підґрунтям дослідження стали діалектні тексти, наведені у збірниках [9–11]. У монографії Г.І. Мартинової [6] представлено принципи, методи та результати лінгво-географічної класифікації говірок. Вони використані для формування класів Z (2) – визначення переліку текстів, що відносяться до певного виду говірок. У результаті перетворення цих текстів у чисельні характеристики сформовані МВД для синтезу індуктивних моделей-класифікаторів. У табл. 1 поданий перелік класів досліджуваних текстів.

Таблиця 1. Перелік класів

Клас	Тип говірки (назва класу)	Опис класу
1	Західна зона середньонаддніпрянського ареалу (перехідні та східноподільські говірки)	Перехідні говірки середньонаддніпрянсько-подільського та середньонаддніпрянсько-волинського типів на межі двох наріч (південно-східного та південно-західного)
2	Центральна зона середньонаддніпрянського ареалу	Говірки з найбільш типовими для середньонаддніпрянського діалекту особливостями
3	Північна зона середньонаддніпрянського ареалу	Говірки з найбільш типовими для середньонаддніпрянського діалекту особливостями, що мають окремі вкраплення ознак середньо- і східнополіського діалектів північного наріччя

З метою оцінки коректності поєднання текстів у класи у табл. 2 поданий перелік населених пунктів, в яких записані текстові повідомлення, та їхня класифікація за типами говірок [11].

Таблиця 2. Перелік населених пунктів, в яких записані діалектні тексти

№ з/п	Клас	Район	Село
1	1	Звенигородський	Багачівка, Княжа, Моринці, Стебне, Боровикове
2	1	Катеринопільський	Вікнине, Пальчик, Петраківка, Ямпіль, Ярошівка
3	1	Лисянський	Вотилівка, Порадівка, Боярка
4	1	Маньківський	Іваньки, Кинашівка, Кривець, Чорна Кам’янка, Багва
5	1	Монастирищенський	Княжики, Попудня, Халаїдове
6	1	Тальнівський	Гордашівка, Зеленьків, Колодисте, Криві Коліна, Онопріївка, Білашки
7	1	Уманський	Доброводи, Дубова, Ладижинка, Острівець, Ропотуха, Ятранівка

Продовж. табл. 2

8	1	Христинівський	Шукайвода
9	1	Шполянський	Кавунівка, Кримки, Лозуватка, Соболівка
10	2	Золотоніський	Богуславець, Вознесенське, Гельмязів, Домантове, Ковтуни, Коробівка, Кропивна, Піщане, Скориківка, Хвилівка, Хутори Каврайські, Деньги
11	2	Чорнобаївський	Богодухівка, Васютинці, Велика Бурімка, Великі Канівці, Вереміївка, Воронинці, Ірклiїв, Кліщинці, Комінтерн, Ленінське, Москаленки, Тимченки, Хрестителеве
12	3	Драбівський	Бирлівка, Білоусівка, Великий Хутір, Золотоношка, Кантакузівка, Кононівка, Мехедівка, Нехайки
13	3	Золотоніський	Зорівка, Каленики, Підставки
14	3	Канівський	Бобриця, Литвинець, Сушки

При формуванні МВД у таблицю поєднані значення частотних характеристик показників тексту, перелік яких поданий у [6]. Частотні характеристики були розраховані на окремих вікнах – ділянках тексту, які містили по 5000 знаків. У результаті для синтезу моделей використано 119 точок спостережень первинного опису. Їх розбито на послідовності *A* і *B* для формування зовнішнього критерію селекції моделей. Ще 11 точок утворювали послідовність *C*, їх використано для випробувань готових моделей, але у процесі синтезу цих моделей вони участі не брали. У процесі синтезу моделі розв'язано завдання класифікації точок спостереження. Модель навчалась зараховувати тексти із табл. 2, описані точками спостереження в масиві даних ПО, до конкретних класів, поданих у табл. 2. Після навчання моделі отримували назви, що збігаються з населеними пунктами, які були об'єктами для моделювання. У табл. 3 подані результати випробувань отриманих моделей.

Таблиця 3. Результати випробування моделей

№ з/п	Назва моделі	Клас	Кількість правильно класифікованих точок спостереження, %
1	Західні_Моринці	1	92,86
2	Західні_Чорна Кам'янка	1	82,14
3	Західні_Соболівка Шполянського	1	96,43
4	Західні_Ладижинка Уманського	1	92,86
5	Центральні_Гельмязів	2	82,14
6	Центральні_Воронинці	2	82,14
7	Центральні_Москаленки	2	85,71
8	Центральні_Богодухівка	2	82,14
9	Центральні_Ірклiїв	2	89,29
10	Північні_Зорівка	3	100,00
11	Північна_Кононівка	3	89,29

Результати випробування моделей, подані в табл. 3, свідчать, що вдалось синтезувати корисні моделі, здатні виконувати функції класифікатора. Це означає, що МВД інформативні і різноманітність методів і засобів синтезу моделі є достатньою для побудови корисних моде-

лей. Отримано експериментальне підтвердження достовірності принципів та методів класифікації говірок, поданих у монографії [6].

5. Висновки

Отримання задовільних результатів розв'язання завдання інтелектуального аналізу текстів методом індуктивного моделювання дозволяє розширити можливості інформаційної технології багаторівневого моніторингу і застосувати інформаційний моніторинг.

Доведено, що декомпозиція текстів на вікна в 5000 знаків дозволяє отримати стійкі й задовільні результати класифікації говірок при застосуванні синтезатора моделей МІС, тобто МІС набуває можливості визначати місце проживання досліджуваного об'єкта.

Отримано експериментальне підтвердження гіпотези про можливість використання методів індуктивного моделювання для побудови вирішуючого правила та успішного виконання завдання класифікації текстів за місцем проживання їх авторів. Доведена здатність виконання завдань із інтелектуального аналізу текстів засобами моделювання моніторингових інформаційних систем. Кількість правильно розпізнаних точок спостереження в досліджуваних умовах перебуває в межах від 78 % до 100 %.

Запропоновано підхід до забезпечення процесу підтвердження ефективності лінгвістичних методів класифікації говірок. Експериментально підтверджено його ефективність.

У перспективі необхідно виявити мінімальний обсяг вікна, який дозволяє надійно забезпечувати класифікацію текстів за типами говірок із урахуванням обробки 2–3 точок спостереження.

СПИСОК ЛІТЕРАТУРИ

1. Голуб С.В. Багаторівневе моделювання в технологіях моніторингу оточуючого середовища / Голуб С.В. – Черкаси: Вид. від. ЧНУ імені Богдана Хмельницького, 2007. – 220 с.
2. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем / Ивахненко А.Г. – К.: Наукова думка, 1981. – 296 с.
3. Голуб С.В. Відображення консолідованої інформації економічних показників регіону у структурі багаторівневих моделей / С.В. Голуб, Н.О. Химиця // Вісник Східноукраїнського національного університету імені Володимира Даля. – 2012. – № 8 (179), Ч. 1. – С. 122 – 128.
4. Pennebaker J.W. Secret life of pronouns: what our words say about us / Pennebaker J.W. – N.Y.: Blumberry Press, 2011. – 352 p.
5. Литвинова Т.А. Формально-грамматические корреляты личностных особенностей автора письменного текста / Т.А. Литвинова // Филологические науки. Вопросы теории и практики. – 2013. – № 12 (30), Ч. 1. – С. 132 – 135.
6. Мартинова Г. Середньонаддніпрянський діалект. Фонологія і фонетика / Мартинова Г. – Черкаси: Тясмин, 2003. – 356 с.
7. 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: зб. наук. праць. – Х.: Харківський університет повітряних сил імені Івана Кожедуба, 2014. – Вип. 2 (118). – С. 89 – 92.
8. Голуб С.В. Відображення властивостей автора тексту в структурі багатопараметричної моделі / С.В. Голуб, О.В. Константиновська, М.С. Голуб // Системи обробки інформації: зб. наук. праць. – Х.: Харківський університет повітряних сил імені Івана Кожедуба, 2014. – Вип. 9 (125). – С. 82 – 87.
9. Говірки Південної Київщини: зб. діалектних текстів / Упорядники Г.І. Мартинова, З.М. Денисенко, Т.В. Щербина. – Черкаси: ПП Чабаненко Ю.А., 2008. – 370 с.
10. Говірки Західної Полтавщини: зб. діалектних текстів / Упорядник Г.І. Мартинова. – Черкаси: ПП Чабаненко Ю.А., 2012. – 325 с.

11. Говірки Черкащини: збірник діалектних текстів / Упорядники Г.І. Мартинова, Т.В. Щербина, А.А. Таран. – Черкаси: ПП Чабаненко Ю.А., 2013. – 870 с.

Стаття надійшла до редакції 17.10.2016