НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
МІЖНАРОДНИЙ НАУКОВО-НАВЧАЛЬНИЙ ЦЕНТР ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА СИСТЕМ

СОКОЛОВ Артем Михайлович

УДК 004.8+004.032.26

МЕТОДИ НЕЙРОМЕРЕЖЕВОГО РОЗПОДІЛЕНОГО ПРЕДСТАВЛЕННЯ ТА ПОШУКУ СХОЖИХ СИМВОЛЬНИХ ПОСЛІДОВНОСТЕЙ В ЗАДАЧАХ КЛАСИФІКАЦІЇ НА ОСНОВІ МІРКУВАНЬ ЗА ПРИКЛАДАМИ

05.13.23 – системи та засоби штучного інтелекту

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Київ – 2008

Дисертацією є рукопис.

Робота виконана у Міжнародному науково-навчальному центрі інформаційних технологій та систем НАН і МОН України.

Науковий керівник: доктор технічних наук
Рачковський Дмитро Андрійович,
Міжнародний науково-навчальний центр
інформаційних технологій та систем
НАН і МОН України,
старший науковий співробітник.

Офіційні опоненти: доктор технічних наук, професор
Вінцюк Тарас Климович,
Міжнародний науково-навчальний центр
інформаційних технологій та систем
НАН і МОН України,
завідувач відділу;

кандидат фізико-математичних наук
Новицький Дмитро Вадимович,
Інститут проблем математичних машин і систем
НАН України,
старший науковий співробітник.

Захист відбудеться «28» жовтня 2008 року о 15 годині на засіданні спеціалізованої вченої ради Д 26.171.01 в Міжнародному науково-навчальному центрі інформаційних технологій та систем НАН і МОН України за адресою: 03680, Київ, проспект акад. Глушкова, 40.

З дисертацією можна ознайомитися в бібліотеці Міжнародного науково-навчального центру інформаційних технологій та систем НАН і МОН України: 03680, Київ, проспект акад. Глушкова, 40.

Автореферат розісланий «25» вересня 2008 року.

Вчений секретар
спеціалізованої вченої ради Бабак О. В.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. При розв'язанні широкого спектру задач пошуку, класифікації і розпізнавання якість аналізу великих масивів даних істотно залежить від можливості врахування не тільки наявності, а й послідовності інформаційних елементів. Прикладами актуальних задач, які вимагають роботи з послідовностями, є пошук текстових дублікатів у пошукових машинах, боротьба з несанкціонованими розсилками електронної пошти (спамом), пошук генетичних послідовностей, виявлення вторгнень у комп'ютерних системах для забезпечення інформаційної безпеки, розпізнавання й ідентифікація акустичних сигналів тощо. Для розв'язання такого класу задач перспективним є використання підходу, що базується на моделюванні інтелектуальної діяльності людини і передбачає реалізацію механізму міркувань за прикладами.

При розв'язанні задач на основі прикладів у базі прикладів запам'ятовуються дані з супутньою інформацією, наприклад відомими текстами, спам-повідомленнями, розміченими генетичними послідовностями, "почерк" користувачів комп'ютерних систем з відомими ідентифікаторами тощо. Для нової вхідної інформації система знаходить у базі один чи кілька схожих збережених прикладів і приймає рішення, а також робить прогнози і висновки про вхідні дані, адаптуючи до них знання про відомі приклади. (J.G. Carbonell, K. Forbus, D. Gentner, J. Kolodner, C.K. Riesbeck, R.C. Shank, В.П. Гладун, Н.Г. Загоруйко, Д.А. Поспєлов, А.И. Уємов та ін.).

Етап пошуку схожих прикладів є центральним у методі міркувань за прикладами. Для оцінки схожості прикладів, що мають послідовну структуру, часто використовують відстань Левенштейна (класичну відстань редагування). Широко відомим є класичний алгоритм обчислення відстані Левенштейна, складність якого квадратична відносно довжини послідовностей. Однак при великих довжинах і великій кількості послідовностей у базах і у вхідних потоках безпосереднє застосування цього алгоритму вимагає великих обчислювальних витрат.

Для підвищення ефективності знаходження схожості прикладів використовують трансформації форм їх представлення. Низка підходів до такої трансформації пов'язана з моделюванням нейромережевих механізмів структурно-функціональної організації мозку (П.І. Бідюк, Є.В. Бодянський, О.М. Касаткін, Л.М. Касаткіна, Н.Н. Куссуль, О.М. Різник, А.А. Фролов, S. Amari, J. Hopfield, S. Grossberg, T. Kohonen, B. Widrow, D. Willshaw та інші) і розподіленого представлення інформації в мозку (D. Hebb, G. Hinton, P. Kanerva, J. McClelland, G. Palm, T. Plate, J. Pollack, D. Rumelhart, P. Smolensky, D. Touretzky, Е.М. Куссуль, Д.А. Рачковський та ін.). Розподілене представлення – форма векторного представлення інформації, де кожен об'єкт (символ, підпослідовність, ознака, фізичний об'єкт, їх сукупність тощо) представлений множиною елементів вектора, а окремий елемент вектора може належати представленням різних об'єктів. Цей

підхід тісно пов’язаний із підходом вкладень просторів (P. Indyk, R. Motwani, A. Broder, C. Sahinalp, G. Cormode, T. Batu, M. Charikar та ін.).

Розподілені представлення забезпечують високу інформаційну ємність, обчислювально ефективну оцінку схожості векторними мірами, а також дають змогу використовувати відомі методи обробки векторної інформації. Однак запропоновані підходи до нейромережевого розподіленого представлення послідовностей потребують теоретичного обґрунтування, мають бути створені та реалізовані відповідні методи, ефективність яких потрібно дослідити у практичних задачах.

Таким чином, в умовах зростання обсягів і складності оброблюваної інформації, яка містить символьні послідовності (тексти, Інтернет, електронна пошта, геноми, аудит-послідовності комп’ютерних систем), існує практична потреба у підвищенні ефективності обробки послідовностей на основі реалізації міркувань за прикладами за допомогою розподілених представлень. У той же час рівень розвитку теоретичної бази розподіленого представлення послідовностей є недостатнім для ефективного розв'язання прикладних задач, пов’язаних із пошуком схожих прикладів, що мають структуру послідовності елементів.

Це обумовлює актуальність наукової задачі розробки методів нейромережевого розподіленого представлення послідовностей, а також їх пошуку і класифікації, для ефективної оцінки схожості і використання інформації про послідовності в системах штучного інтелекту, де застосовують моделі міркувань людини на основі прикладів. Практичні аспекти роботи вимагають створення програмних інструментальних засобів і прикладних систем, у яких реалізують і використовують розроблені методи, а також дослідження цих засобів і систем у застосуваннях.

Зв’язок роботи з науковими програмами, планами, темами. Робота виконувалася в рамках таких НДР: "Розробка та дослідження нейромережевих методів моделювання когнітивних процесів", (№ ДР 0101U002685, 2001-2003); "Дослідження та розроблення нових інтелектуальних інформаційних технологій на основі використання високоефективних нейромережевих методів та алгоритмів" (№ ДР 0102U002070, 2002-2006); "Розробка та дослідження нейромережевих інформаційних технологій роботи з базами знань" (№ ДР 0104U003191, 2004-2006); "Створити дослідні зразки нейрокомп’ютерів нових поколінь" (№ ДР 0101U006718, 2000-2001), "Розробити методи та створити способи інтелектуалізації інформаційних технологій широкого використання" (№ ДР 0101U007953, 2001); "Створити засоби автоматичної обробки інформації із застосуванням міркувань за аналогіями" (№ ДР 0103U008280, 2003-2006), ДНТП "Образний комп’ютер": "Розробка технології для створення систем смислової інтерпретації текстової інформації та смислового перекладу текстів з однієї мови на іншу" (№ ДР 0102U005512, 2002); "Розробити комп’ютерну технологію цілеспрямованої обробки текстової і аудіо-інформації" (№ ДР 0103U005770, 2003); "Розробити інтелектуальні інформаційні технології розпізнавання та ідентифікації аудіо-відеоінформації на основі нейромережевих технологій " (№ ДР 0104U008324, 2004).

Мета роботи: підвищення ефективності оцінки схожості інформації з послідовною структурою для розв'язання прикладних задач класифікації і пошуку за рахунок розробки і реалізації методів та засобів нейромережевого векторного розподіленого представлення послідовностей.

Для досягнення цієї мети було поставлено й розв'язано такі **задачі**:

1. Розробити методи векторного представлення символьних послідовностей, що зберігають схожість за відстанню редагування.
2. Розробити і теоретично проаналізувати методи рандомізованого розподіленого представлення послідовної інформації.
3. Розробити і дослідити методи пошуку схожих символьних послідовностей за допомогою розподілених представлень.
4. Розробити програмні засоби, що реалізують запропоновані методи представлення, пошуку і класифікації приблизно найближчих послідовностей.
5. Експериментально дослідити обчислювану ефективність і якість розроблених методів у задачах пошуку і класифікації інформації з послідовною структурою на основі міркувань за прикладами.

*Об'єкт дослідження***:** представлення й обробка символьних послідовностей в системах пошуку і класифікації.

Предмет дослідження: нейромережеві розподілені представлення інформації, методи їх формування й обробки, методи оцінки схожості символьних рядків, методи пошуку і класифікації схожих послідовностей.

Методи дослідження. При розробці та дослідженні методів формування розподілених представлень, а також пошуку і класифікації послідовної інформації використовувалися методи математичного та імітаційного моделювання, дискретної математики, теорії ймовірностей і математичної статистики. Для експериментальної перевірки розроблених методів і програмних систем застосовувалися методи статистичної обробки результатів масових експериментальних досліджень. При розробці систем і засобів реалізації запропонованих методів використовувалися методи системного й об'єктно-орієнтованого аналізу, проектування і функціонального програмування.

Наукова новизна роботи. Основні результати роботи, які визначають наукову новизну та виносяться на захист, такі:

1. Розроблено новий метод вкладення простору послідовностей з відстанню редагування Левенштейна у векторний простір з манхетеновою відстанню, який відрізняється врахуванням підпослідовностей змінної довжини й забезпечує підвищення точності апроксимації відстані редагування.
2. Уперше запропоновано локально-чуттєву функцію, яка відрізняється врахуванням відстані редагування між послідовностями та продукує розподілене представлення послідовностей, яке зберігає їх схожість за відстанню редагування.
3. Уперше отримано оцінки витрат пам'яті, обчислювальної складності й точності апроксимації відстані редагування на основі одержаних розподілених

представлень за рахунок використання методів метричних вкладень просторів й методів аналізу рандомізованих алгоритмів.

4. Удосконалено методи пошуку приблизних найближчих послідовностей за рахунок застосування розроблених локально-чуттєвих хеш-функцій, що дало змогу забезпечити сублінійний відносно розміру бази прикладів час пошуку схожих послідовностей.

5. Отримали подальший розвиток методи розв'язання задач пошуку та класифікації послідовностей на базі міркувань за прикладами за рахунок використання розробленого методу хешування і загальної базової операції пошуку найближчих сусідів у базах прикладів-послідовностей, що мають різну довжину.

Практична значимість отриманих результатів.

На основі розроблених оригінальних методів представлення і пошуку послідовностей створено нові програмні засоби для розв'язання прикладних задач і реалізації інформаційних технологій, пов’язаних із обробкою символьних послідовностей, а саме:

- програмна об’єктно-орієнтована бібліотека для представлення та пошуку послідовностей *LSHLibrary*, що реалізує методи представлення й пошуку приблизних найближчих символьних послідовностей.
- програмний засіб *TextInputTools*, який містить модулі форматованого вводу для баз генетичних послідовностей, електронних листів, ряду популярних текстових корпусів, аудит-послідовностей UNIX-систем.
- програмні макети *DuplClassifier*, *EmailClassifier*, *NuclClassifier*, SessionClassifier для пошуку текстових дублікатів, спаму, кодуючих ділянок генетичних послідовностей, а також класифікації сесій користувачів *UNIX*-системи.
- модулі програмного нейрокомп’ютера *SNC*:
- *KNN*, який реалізує алгоритм пошуку *K* найближчих сусідів за заданою метрикою;
- *VectorComparer* – уніфікований засіб порівняння векторів за множиною стандартних метрик і мір;
- блоки обробки: форматування, передобробки, кодування тощо.

Розроблене алгоритмічне й програмне забезпечення використовується в наукових і практичних цілях, що підтверджується відповідними актами: Міністерства промислової політики України (від 26.10.05); Інституту інформатики АН Чеської Республіки (від 21.11.2005); ТЕЛКО ЛІМІТЕД (від 17.10.2007).

Особистий внесок здобувача. В роботах, написаних у співавторстві й опублікованих у профільних виданнях, внесок здобувача такий: в [1] – метод ідентифікації користувачів за допомогою нейронних мереж зворотного розповсюдження помилки; в [4] – концепція модулів форматованого вводу інформації та класифікації в нейрокомп’ютері; в [7] – проріджувальна схема

розподіленого представлення послідовностей та ідеї її аналізу; в [5] – пошук текстової інформації з використанням *q*-грамного представлення; в [6] – модулі форматованого вводу інформації та класифікації.

Апробація результатів дисертації. Про результати дисертаційного дослідження було зроблено доповіді на: Int. Joint Conf. on Neural Networks (USA, 2003); на X, XI, XII Int. Conf. "Knowledge-Dialogue-Solution" (Bulgaria, 2003, 2005, 2006); на Міжнародній конференції "Проблеми нейрокібернетики" (Ростов-на-Дону, 2002, 2005); на Міжнародному семінарі з індуктивного моделювання (Київ, 2005); Школі-семінарі "Про проблеми образного мислення" (Жукин, 2005); на семінарах "Проблеми нейрокомп'ютерів і нейромереж" Наукової ради НАНУ з проблеми "Кібернетика" (Київ, ІПММС НАН України і МННЦ ІТС НАН України, 2001 – 2008) та "Образний комп'ютер" (Київ, МННЦ ІТС НАН України, 2008).

Публікації. Основні результати дисертації викладено у 18 друкованих роботах, з них 11 – в профільних виданнях України й інших країн, з них 6 – без співавторів.

Структура дисертації. Дисертація складається із вступу, п'яти розділів, висновків, списку використаних джерел із 150 найменувань та одного додатка. Основна частина займає 143 сторінки, ілюстрована 30 рисунками та 12 таблицями. Загальний обсяг дисертаційної роботи – 159 сторінок.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дисертаційної роботи і наукової задачі, сформульовано мету і задачі дослідження, показано наукову новизну і практичну значимість отриманих результатів, зазначено особистий внесок здобувача, наведено список публікацій.

У **першому** розділі розглянуто модель міркувань за прикладами, існуючі підходи до визначення мір схожості послідовностей, алгоритмів обчислення мір та їх апроксимації.

Для розв'язання задач пошуку і класифікації, що оперують з послідовностями, перспективним є використання підходу на основі міркувань за прикладами: у базі запам'ятовуються приклади, а для вхідної інформації система знаходить у базі схожі приклади і на їх основі приймає рішення про вхідні дані (рис. 1).

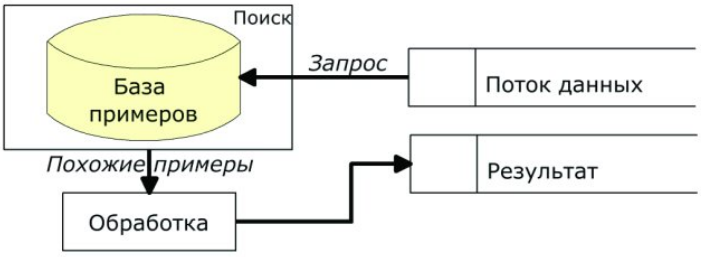

Рис. 1. Схема міркувань на основі прикладів

Приклад-послідовність розглядається як рядок символів зі скінченного алфавіту Σ. Для оцінки схожості рядків довжини n застосовують як векторні метрики (Хеммінга, відстані ℓ_p), які не враховують порядку елементів послідовностей, але швидко обчислюються, так і різні метрики редагування, які цей порядок враховують, але мають більшу обчислювальну складність. Класичну

відстань редагування $ed(x,y)$ (В. Левенштейн) між рядками $x,y\in\Sigma^n$ визначено як мінімальне число операцій видалення, вставки й заміни символів, необхідних для перетворення x в y. Класичний алгоритм обчислення такої відстані (Т. Вінцюк) використовує динамічне програмування і має складність $O(n^2)$, яка може бути зменшена до $O(n^2/\log n)$ (M. Paterson).

Швидке знаходження схожості між послідовностями може бути забезпечене використанням їх нейромережевих розподілених представлень, тобто форми представлення об'єктів різної природи сукупністю елементів векторів, а також застосування ефективних векторних операцій порівняння. Однак підходи до розподіленого представлення послідовностей потребують теоретичного обґрунтування і дослідження в практичних задачах.

Для аналізу розподілених представлень перспективним є використання підходів метричних вкладень, які виконують відображення об'єктів із простору зі складнообчислюваною метрикою, якою є класична відстань редагування, у векторний простір з метрикою, що обчислюється легко. В огляді розглянуто сучасні результати, що отримані в цьому напрямі, й обґрунтовано необхідність їх покращення за обчислювальною складністю й обсягом потрібної пам'яті.

Таким чином, при зростанні довжин послідовностей і обсягів даних у базах прикладів і у вхідних потоках складність точного обчислення відстані редагування перешкоджає його ефективній оцінці й пошуку схожих послідовностей у базах прикладів. Це зумовлює актуальність спрямованості дисертаційної роботи на розробку й дослідження нових методів представлення послідовностей і на підвищення ефективності розв'язання задач пошуку і класифікації на їх основі.

У **другому** розділі запропоновано *детермінований метод формування* векторного представлення послідовностей, який зберігає схожість за відстанню редагування, а також отримані оцінки точності розробленого вкладення.

Розглядаються *q-грами* – підрядки символьного рядка $x\in\Sigma^n$ довжиною $q\in\mathbb{N}$. Уводиться поняття *q-грамного вектора* $v_{n,q}(x)\in(\mathbb{N}\cup\{0\})^{|\Sigma|^q}$, де кожній q-грамі $\sigma\in\Sigma^q$ відповідає елемент v_i вектора v, значення якого – число входжень σ в x. Манхетенова відстань (ℓ_1) між такими векторами, тобто сума модулів різниць значень елементів векторів, є *q-грамною відстанню*.

Розроблено *детермінований метод апроксимації* схожості символьних послідовностей за відстанню редагування шляхом вкладення простору послідовностей із класичною метрикою редагування у векторний простір з метрикою ℓ_1.

Розглянемо вікно шириною w символів і два значення довжини q-грам: q_1 і q_2, $q_1<q_2$. Позначимо $x[i,j]$ підрядок рядка x, який містить символи x від i до j. Рядок x перетворюється у вектор $v(x)$ конкатенацією усіх q-грамних векторів:

$$v_{i,w,q}(x)=v_{w,q}(x[i,i+w-1]), \qquad (1)$$

для $q=q_1,\ldots,q_2$ і $i=1,\ldots,n-w+1$. За відстань між такими векторами приймемо q-

грамну відстань, тобто для рядків $x,y\in\sum^w$:

$$d^{\Sigma}_{w,q1,q2}(x,y)=\sum_{q=q1,\ldots,q2} d_q(x,y). \quad (2)$$

Позначимо $\Delta q=q_2-q_1$. Для $x,y\in\Sigma^n$ визначимо з використанням (2) відстань

$$D(x,y) = \sum_{i=1,\ldots,n-w+1} d^{\Sigma}_{w,q1,q2}(x[i,i+w-1],y[i,i+w-1])\} \,/\, ((n-w+1)(q+1)). \quad (3)$$

У дисертаційній роботі доведено таку теорему.

Теорема 1. Нехай $w\ge6$, $k_1\ge1$, $q_1=2w/3$, $\Delta q=\lfloor 1/2(-7+(57+16(w-q_1))^{1/2})\rfloor$, $Q=(\Delta q+1)(\Delta q+2)$, $t=w-\Delta q+1$, $n>w(k_1+1)+1$, тоді

якщо $ed(x,y) > k_2$, то $D(x,y) \ge Qt(k_2 / (2(\Delta q+1))-2) / ((n-w+1)(\Delta q+1)) = d_2$, (4)

якщо $ed(x,y) \le k_1$, то $D(x,y) \le 2k_1[w^2+(n+1)] / (n-w+1) = d_1$. (5)

Доведення базується на апараті графів де Брейна – направлених графах $G[\Sigma,q]$ для алфавіту Σ і параметра q, що складаються з вершин, які відповідають множині Σ^{q-1}, і дуг, які відповідають Σ^q. Дуга з'єднує вершини тоді і тільки тоді, коли вихідна вершина є її максимальним префіксом, а цільова вершина – максимальним суфіксом. Кожному рядку x, $|x|\ge q$ відповідає певний шлях π_x на графі $G[\Sigma,q]$, що складається з дуг, які послідовно з'єднують вершини, мітки яких – (q-1)-грами, що послідовно входять у рядок x. Позначимо $B[x;q]$ ($B[x,y;q]$) підграф $G[\Sigma,q]$, який складається з вершин і дуг, що входять у π_x ($\pi_x\cup\pi_y$).

Рядок $x\in\Sigma^*$ назвемо (q,w)-повторюваним, якщо у будь-якому інтервалі з w символів усі $w-q+1$ штук q-грам відмінні. У цьому випадку можлива класифікація взаємних локальних конфігурацій шляхів на графі, яка відповідає двом послідовностям, на петлі й вилки (рис. 2). Якщо рядок не є (q,w)-повторюваним, то в $B[x,q]$ існують цикли. Для петель і вилок відрізняються зміни кількості відмінних дуг при збільшенні довжини q-грам. Збільшення q на одиницю збільшує на два кількість відмінних дуг у конфігурації петля, в той час як кількість відмінних дуг у вилці не змінюється. Це дає можливість оцінки кількості операцій редагування, необхідних для перетворення одного рядка в інший, базуючись на значенні введеної відстані (2).

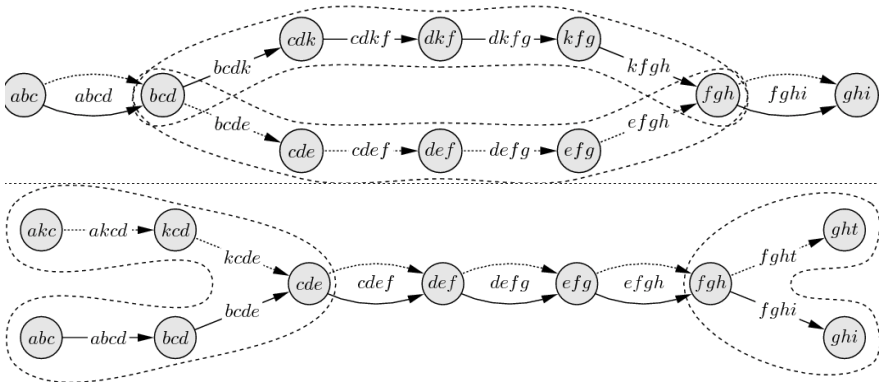

Рис. 2. Приклади конфігурацій шляхів петля і вилка

У роботі доведено такі леми.

Лема 1. Нехай $x,y\in\Sigma^w$ (q_1,w)-неповторювані, $w\ge q_2>q_1\ge3$, $\Delta q\le\lfloor(w-q_1+1)/2\rfloor$, $Q\le4(w-q_2+1)$, $d^{\Sigma}_{w,q1,q2}(x,y) < Q$, тоді $ed(x,y) < 2(\Delta q+1)$.

Лема 2. Для $x\in\Sigma^w$, $q>2w/3$, якщо існує підрядок $x'\subseteq x$, такий, що $\pi_{x'}$ утворює цикл C на $B[x,q]$, то однакові дуги поза циклом C відсутні.

Лема 3. Для $x,y\in\Sigma^w$, $q>2w/3$, якщо перші (останні) вершини підшляхів $\pi_{x'}\subset\pi_x$, $\pi_{y'}\subset\pi_y$, які утворюють спільний цикл C, не збігаються, то спільні дуги у π_x і π_y перед (після) цих вершин відсутні.

Використовуючи леми 2 та 3, доведено таку лему.

Лема 4. Нехай при $q_1>2w/3$ для рядків $x,y\in\Sigma^w$ існують такі підрядки $x'\subseteq x$, $y'\subseteq y$, $x'\neq y'$, що на графі $B[x,y;q]$ шляхи $\pi_{x'}$ і $\pi_{y'}$ складаються з дуг, які належать спільному циклу C. Нехай також $d^{\Sigma}_{w,q_1,q_2}(x,y)\le Q$. Тоді $ed(x,y)<2(\Delta q+1)$.

Використовуючи леми 1 і 4, доведено лему 5.

Лема 5. Нехай $x,y\in\Sigma^w$, $w\ge 6$, $w\ge q_2>q_1\ge 2w/3$ і $\Delta q\le\lfloor(w-q_1+1)/2\rfloor$, $Q\le 4(w-q_2+1)$. Якщо $d^{\Sigma}_{w,q_1,q_2}(x,y)<Q$, то $ed(x,y)<2(\Delta q+1)$.

Результат леми 5 розповсюджено на суміжні інтервали двох послідовностей.

Лема 6. Нехай $x,y\in\Sigma^m$. За виконання умов леми 5, якщо для всіх $i=0,\ldots,(m-w)$ виконується $d^{\Sigma}_{w,q_1,q_2}(x[it'+1,it'+w],y[it'+1,it'+w])<Q$, то $ed(x,y)<2(\Delta q+1)$.

Теорему 1 доведено на основі наступних лем, які визначають кількість інтервалів, усунення відмінностей між якими потребує (за результатами попередніх лем) невеликої кількості операцій.

Лема 7. Нехай кількість інтервалів, де $d^{\Sigma}_{w,q_1,q_2}(x[i,i+w-1],y[i,i+w-1])>Q$ для рядків x,y довжиною n фіксована й дорівнює N. Тоді $ed(x,y)\le 2(\Delta q+1)(\lceil N/t\rceil+1)$.

Лема 8. Для $x,y\in\Sigma^n$ і $q_1<q_2\le w\le(n+1)(k_1+1)$, якщо $ed(x,y)\le k_1$, то кількість інтервалів, де $d^{\Sigma}_{w,q_1,q_2}(x[i,i+w-1],y[i,i+w-1])\le 2k_1(\Delta q+1)$ більше $n-w(k_1+1)+1$.

Експериментальну перевірку обмежень (4) і (5) виконано за допомогою чисельних експериментів на штучних даних – для згенерованого набору рядків, що розташовані на різних відстанях редагування від фіксованого випадково згенерованого рядку довжиною n=1000,5000,10000,50000,100000. Показано, що експериментальні значення $D(x,y)$ (3) для кожного зі значень $ed(x,y)$ відповідають теоретичним значенням верхньої (4) і нижньої (5) границь на значення відстані (3).

Чим меншою є різниця між k_1 і k_2, тим точніше можна апроксимувати $ed(x,y)$ за відомою $D(x,y)$. Покладемо $w=n^{\gamma}$, тоді показник росту w, що забезпечує найбільшу точність апроксимації, досягається при γ=0.5. При цьому $k_2=\Omega(k_1 n^{1/2})$, час побудови векторів – $O(n^{3/2})$, а їх розмір – $O(n^{5/4})$. Розмір векторів може бути надто великим для довгих послідовностей. Тому був розроблений рандомізований метод (розд. 3), який дає змогу зменшити вимоги до ресурсів на створення й зберігання векторів, отримати ефективний сублінійний до розмірів бази метод пошуку приблизних найближчих сусідів, а також розподілену схему представлення послідовностей.

У **третьому** розділі розроблено *методи розподіленого представлення* послідовностей. Для розробки розподіленого представлення послідовностей і ефективного *методу пошуку приблизних найближчих послідовностей* запропоновано рандомізацію *детермінованого методу* вкладення відстані редагування (розд. 2), яка використовує локально-чуттєве хешування (*LSH*) P. Indyk і R. Motwani.

Нехай вектор $v^i_{w,q_1,q_2}(x)$ є конкатенацією q-грамних (від q_1 до q_2) векторів (2) підрядка $x[i,i+w-1]$ довжиною w, де i вибрано випадково й рівномірно з множини можливих позицій вікна шириною w: $i = 1,...,n-w+1$. Розмірність вектора $v^i_{w,q_1,q_2}(x)$ дорівнює $\sum_{q=q_1,...,q_2}|\Sigma|^q$. Нехай ϕ^i – випадковий вектор такої ж розмірності, як і $v^i_{w,q_1,q_2}(x)$, з елементами, що вибрані випадково з розподілу Коші $p(x)=(\pi(1+x^2))^{-1}$. Побудуємо для рядку x хеш-вектор розмірності K: $h(x)=(h_1(x),h_2(x),...,h_K(x))$, де

$$h_i(x) = \lfloor((v^i_{w,q_1,q_2}(x),\ \phi^i) + b_i) / r\rfloor, \tag{6}$$

r і b_i – дійсні числа, b_i вибрано незалежно, випадково й рівномірно з $[0, r]$.

Визначимо шар $B(y,k)=\{x\in\Sigma^n | ed(y,x)\le k\}$. Сімейство $H=\{h: \Sigma^n\to X\}$ (X – певна скінченна або зліченна множина значень) називається (k_1,k_2,p_1,p_2)-чуттєвим або просто локально-чуттєвим (у термінології P. Indyk і R. Motwani), якщо для будь-яких $x,y\in\Sigma^n$ і будь-якої незалежно і рівномірно вибраної хеш-функції $h\in H$ виконується за $k_1<k_2$ і $p_2<p_1$:

$$\text{якщо } x\in B(y,k_1), \text{ то } Prob[h(x) = h(y)] > p_1, \tag{7}$$

$$\text{якщо } x\notin B(y,k_2), \text{ то } Prob[h(x) = h(y)] < p_2. \tag{8}$$

Використовуючи властивості розробленого детермінованого вкладення (розд. 2) й розподілу Коші у застосуванні до локально-чуттєвого хешування ℓ_1, у роботі доведено, що (6) є локально-чуттєвою функцією (за $w=n^{2/3}$, $r=w$, і q_1, Δq, Q, t таких, як у детермінованому методі).

Запропонована локально-чуттєва функція генерує розподілене представлення послідовності – вектор, елементи якого є значеннями незалежно й рівномірно вибраних функцій виду (6), що зафіксовані і використовуються для отримання розподіленого представлення всіх послідовностей.

На основі запропонованої в роботі локально-чуттєвої функції і отриманих розподілених представлень побудовано *метод пошуку приблизних найближчих рядків*. Нехай із бази $P\subset\Sigma^n$ необхідно на запит $y\in\Sigma^n$ повернути приблизного найближчого сусіда – тобто, якщо існує рядок із бази в $B(y,k_1)$, то повернути будь-який рядок із кулі $B(y,k_2)$. Усі рядки $x\in P$ запам'ятовуються в комірках пам'яті наступним чином:

1. За формулою (6) для кожного рядку x бази P генеруються L штук K-мірних випадкових хеш-векторів $h^j(x)=(h^j_1(x),h^j_2(x),...,h^j_K(x))$, $j=1,...,L$.

2. Для кожного унікального хеш-вектора, отриманого з усіх рядків бази, $h^j(x)$, $j=1,...,L$, $x\in P$ утворюється комірка пам'яті.

3. Кожний рядок бази P ставиться у відповідність тим коміркам, хеш-вектори яких нею продукуються.

Пошук приблизного найближчого сусіда до рядка-запиту $y\in\Sigma^n$:

1. Формується L його хеш-векторів.

2. Переглядаються комірки, хеш-вектори яких повністю збігаються зі сформованими для y, які відповідають кожному хеш-вектору

$h^j(y)=(h^j_1(y),h^j_2(y),...,h^j_K(y)), j=1,...,L$.

3. Складається список рядків з P, що містяться в переглянутих комірках, який позначається S. Оскільки один і той самий рядок може входити в S кілька разів, S можна представити як мультимножину.

Процедура закінчується або після перегляду всіх комірок, що відповідають хеш-векторам $h^j(y)$, $j=1,...,L$, або коли розмір списку S досягає $2L$.

У дисертаційній роботі визначено значення ймовірностей p_1 і p_2 і, підставляючи знайдені значення у результаті для схеми пошуку приблизних найближчих сусідів з використанням загальних локально-чуттєвих функцій, доведено таку теорему:

Теорема 2. При $z>1$, $\rho=\log(p_1/p_2)$, $K=\log_{1/p_2}|P|$, $L=|P|^{\rho}$ алгоритм пошуку видасть у S із імовірністю більшою, ніж 1/2, рядок x, такий, що $ed(x,y)\le k_2$, де $k_2=O(zn^{1/3}\ln n)$.

У **четвертому** розділі розроблено алгоритмічну реалізацію методу пошуку послідовностей розділу 3, а також наведені результати її експериментального дослідження.

Для алгоритмічної реалізації описаного методу пошуку послідовностей за допомогою процедури *LSH* використано *LSH*-ліс (M. Bawa, T. Condie, P. Ganesan), який дає змогу знаходити найближчих сусідів без оновлення хеш-векторів при зміні розміру бази P або параметра k_2. Для кожного $j=1,...,L$ усі хеш-вектори $h^j(x)$ усіх рядків $x\in P$ зберігаються у вигляді окремого префіксного дерева T_j глибиною до K рівнів. Вузли дерева відповідають значенням елементів хеш-вектора й містять посилання на рядки, хеш-вектори яких відповідають шляхові від кореня дерева до даного вузла. Листя дерев відповідають коміркам в оригінальній процедурі. Так, якщо у хеш-векторів двох рядків збігаються перші k їх елементів, то й перші k вузлів на шляху від кореня дерева T_j до листя, що відповідає цим рядкам, також збігаються.

При надходженні рядка-запиту y:

1. Формуються L його K-мірних хеш-векторів $h^j(y)$.

2. Для кожного хеш-вектора $h^j(y)$ у T_j знаходиться вузол, який відповідає $h^j(x)$ із найбільшим числом однакових перших елементів цих векторів.

3. Починаючи зі знайдених на кроці 2 вузлів, усі L дерев синхронно переглядаються у напрямку кореня, а відповідні зазначеним вузлам рядки x додаються у результуючу мультимножину рядків S.

4. Після того як усі рядки, хеш-вектори яких збігаються на даному рівні, додані у S, процедура повторюється для наступного (більш «високого») рівня, поки не досягнуто кореня або $|S|$ не перевищить $2L$.

У результаті отримуємо мультимножину S рядків-кандидатів на приблизного найближчого сусіда до q, впорядковану в порядку зменшення глибини вузлів дерева, до яких збіглися перші елементи хеш-векторів відповідного рядка й запиту.

Як і для детермінованого вкладення (розд. 2), було перевірено потрапляння значень імовірності колізії хеш-функції (6) в межі (7) і (8). На рис. 3 наведено експериментальні значення імовірності $p_{col}=Prob[h(x)=h(y)|ed(x,y)]$ для рядків

довжиною n=1000, які відповідають теоретичним границям.

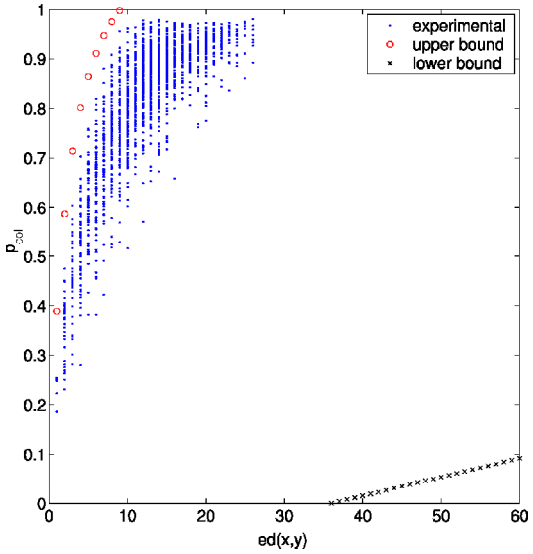

Рис. 3. Залежність імовірності колізії хеш-функції (6) від $ed(x,y)$

Оскільки необхідні ресурси ростуть лінійно від L, рекомендовані теорією значення L (теорема 2) часто завеликі для практики. Однак при використанні менших значень L теорія не гарантує, що в S потрапить як мінімум один рядок $y \in B(q,k_2)$. Експерименти виконувалися для оцінки впливу на мультимножину S вибору менших, ніж теоретичні, значень L (оцінювалася «точність на рівні $|S|$» – $|B(q,k_2) \cap P \cap S|/|S|$) і додаткової фільтрації S, якою імовірно можна позбутися рядків $y \notin B(q,k_2)$ (оцінювалась впорядкованість S відносно відомого реального впорядкування рядків за ed). Експерименти проводилися на наборі P з 2200 рядків довжиною 1000 символів, що був збережений за допомогою процедури *LSH*-ліс.

У першій серії експериментів по отриманому S обчислювалася точність на рівнях $|S|=0.5L,L,2L,3L,4L$. Її значення і дисперсію наведено в табл. 1. При малих L спостерігається максимальна точність, оскільки процедура *LSH*-ліс повертає S, впорядковане за глибиною рівня. При збільшенні L точність падає, що пояснюється більшою кількістю рядків з $P \setminus B(q,k_2)$, які потрапили в S, але в подальшому точність перестає істотно змінюватися і одночасно зменшується її дисперсія. Аналогічний ефект спостерігався при збільшенні $|S|$.

Таблиця 1

Залежність «точності на рівні $|S|$» від L і $|S|$

L	|S|=1/2L	σp	|S|=L	σp	|S|=2L	σp	|S|=3L	σp	|S|=4L	σp
1			0.950	0.048	0.945	0.029	0.927	0.025	0.893	0.031
2	0.930	0.065	0.895	0.056	0.853	0.054	0.825	0.032	0.810	0.028
3			0.885	0.050	0.816	0.035	0.784	0.030	0.770	0.025
4	0.855	0.071	0.842	0.041	0.810	0.031	0.777	0.023	0.757	0.021
5			0.824	0.039	0.786	0.021	0.759	0.014	0.735	0.014
6	0.853	0.052	0.797	0.040	0.782	0.025	0.760	0.019	0.730	0.014
8	0.846	0.038	0.791	0.024	0.755	0.016	0.729	0.013	0.724	0.010
10	0.860	0.030	0.787	0.024	0.749	0.015	0.722	0.009	0.689	0.008
20	0.811	0.024	0.762	0.014	0.708	0.006	0.682	0.005	0.658	0.004

У другій серії експериментів для порівняння впорядкованих мультимножин було взято за основу модифікацію узагальненої відстані Кендалла D_K (R. Fagin, R. Kumar, D. Sivakumar), яка дає змогу порівнювати впорядковані множини, розглядаючи різні штрафи за розташування елементів у різному відносному порядку у двох множинах.

Між отриманим S і реальним впорядкуванням рядків за відстанню редагування обчислювалася відстань D_K при різних $|S|$. Залежність D_K від L зображена на рис. 4 для $|S|$=100 і для 4 способів додаткової фільтрації: 1) «all» (без фільтрації); 2) «==max» (рядки, що збіглися з запитом y на самому глибокому для даного S рівні);

3) «==half» (рядки, що збіглися на рівні від $\lfloor K_{avg} \rfloor$ до $\lceil K_{avg} \rceil$); 4) «>=half» (рядки, що збіглися на рівні $\geq K_{avg}$), де K_{avg} – середнє значення рівня серед рядків із S). З рис. 4 видно, що D_K дещо зменшується при збільшенні L, що можна пояснити збільшенням $|S|$.

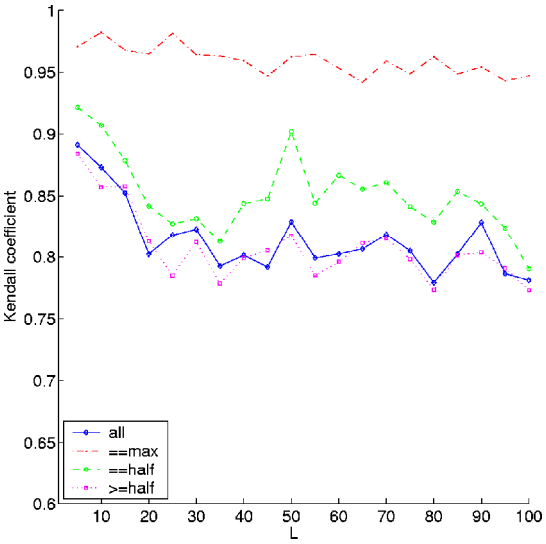

Рис. 4. Залежність D_K від L

Із обох серій експериментів можна зробити висновок про можливість на практиці фіксування L на невеликому значенні. Це дає змогу отримати прийнятний рівень точності за економії ресурсів.

У **п'ятому** розділі розглянуті розроблені й використані в роботі програмні засоби, проведено дослідження розроблених методів у реальних задачах виявлення дублікатів, кодуючих ділянок ДНК, вторгнень у комп'ютерні системи, а також оцінки кількості спаму.

Розроблені й досліджені в рамках дисертаційної роботи представлення і методи пошуку приблизних найближчих послідовностей реалізовано як комплекс програмних засобів. Методи представлення і пошуку послідовностей реалізовані в програмній об’єктно-орієнтованій бібліотеці *LSHLibrary*. Реалізовано програмні макети *DuplClassifier*, *EmailClassifier*, *NuclClassifier*, *SessionClassifier* для пошуку текстових дублікатів, спаму на основі схожості, кодуючих ділянок генетичних послідовностей і класифікації сесій користувачів *UNIX*-системи, а також програмні модулі *TextInputTools* форматованого вводу із баз даних. Частину методів реалізовано як модулі у складі колективної розробки – програмного нейрокомп’ютера *SNC*, що має модульну архітектуру і дає змогу візуально конструювати конфігурації обробки даних, використовуючи розширюваний набір блоків обробки.

Перевірка якості пошуку дублікатів здійснювалася в базах текстів *Reuters*-21578 і навчальних англомовних текстів *British National Corpus*. Для пошуку в текстових базах дублікатами вважалися рядки, у яких було хоча б один збіг хеш-векторів довжиною K. Використовувалася різна довжина n обрізки тексту та відсутність обрізки (вирівнювання за максимальною довжиною, в табл. 2 відповідає n=0). Тексти довжиною менше довжини обрізки n доповнювалися в кінці до довжини n однаковим спецсимволом. Знайдену кількість приблизних дублікатів залежно від значення K і n, для L=1 наведено в табл. 2. Ця кількість стабілізується при великих K на значеннях, які приблизно відповідають відомим результатам інших методів (M. Sanderson, 320 дублікатів). Результати пошуку дублікатів на базі *Reuters* порівнювалися з методом детермінованого вкладення BY (Z. Bar-Yossef). Як еталон було взято визначення пари дублікатів, на якій значення функції *String::Similarity* мови *PERL* (заснованої на прямому обчисленні класичної відстані

Таблиця 2

Кількість знайдених дублікатів у базах *Reuters* і *BNC*

K	*n*=100	*n*=150	*n*=250	*n*=500	*n*=750	*n*=1000	*n*=2000	*n*=0
1	21354	21324	21268	21234	21218	21198	21237	21445
2	20277	20119	19748	19350	19231	19214	19844	21429
5	8601	7635	5376	5317	6971	9783	15457	20742
10	403	361	499	1812	2983	4330	10787	19279
25	372	344	318	363	1062	1643	4047	17755
50	369	342	315	312	538	1158	3048	15702
100	369	342	315	298	264	255	469	4920
150	367	342	315	298	263	255	344	4787
200	367	342	315	298	263	254	242	2161

K	*n*=100	*n*=150	*n*=250	*n*=500	*n*=750	*n*=1000	*n*=2000	*n*=0
1	3939	3928	3921	3902	3884	3876	3833	4028
2	3567	3476	3349	3204	3027	2953	2635	4027
5	921	706	293	79	35	25	14	3559
10	10	10	9	8	7	7	7	3487
25	9	9	9	8	7	7	6	3447
50	9	9	9	8	7	7	6	2413
100	9	9	9	8	7	7	6	1267
150	9	9	9	8	7	7	6	1267
200	9	9	9	8	7	7	6	236

редагування) складало не менше 0.85. Порівняння показало відсутність істотних відмінностей в якості пошуку двома порівнюваними методами, однак порядок часу пошуку дублікатів на всій базі з врахуванням побудови дерев у методі, заснованому на застосуванні *LSH*-ліса, складає хвилини, у той час як застосування детермінованого методу BY вимагає витрат часу від кількох годин до кількох днів.

Досліджувалася також якість пошуку дублікатів на стандартній базі «Дублі Web-сторінок колекції РОМІП», яка містить понад 10 млн. пар веб-сторінок, схожість між якими за значенням функції *String::Similarity* складає не менше 0.85. Використовувалася модифікація методу пошуку дублікатів, описана вище, причому пошук дублікатів здійснювався лише серед документів, довжина яких приблизно дорівнює довжині запиту. Було отримано значення точності від 0.85 до 0.95 і повноти від 0.65 до 0.75, залежно від порогу на значення функції *String::Similarity*, що дає збільшення *F*-міри ($F=2rp/(r+p)$, де r – повнота, p – точність) порівняно з відомими результатами на 0.13-0.26.

За допомогою розроблених методів було оцінено, яка кількість спаму можна виявити лише за допомогою порівняння з раніше отриманим спамом без застосування специфічних знань і докладного аналізу спам-технологій. Було використано базу *TREC Spam Track* (за 2005 і 2006 рр.), яка містить $|P|=37822$ поштових повідомлення, розмічених на два класи: спам (66%) і неспам (34%). Кожному листу присвоювалося значення оцінки *score*. Повідомлення, які отримали *score* більше певного порогу, вважалися спамом, а решта – звичайними листами. Оцінка якості роботи фільтра проводилася за відсотком неправильно класифікованих спам-повідомлень *sm*%, а також за відсотком неправильно класифікованих неспам-повідомлень *hm*%. На рис. 5 подано залежності *sm*% від *hm*% за

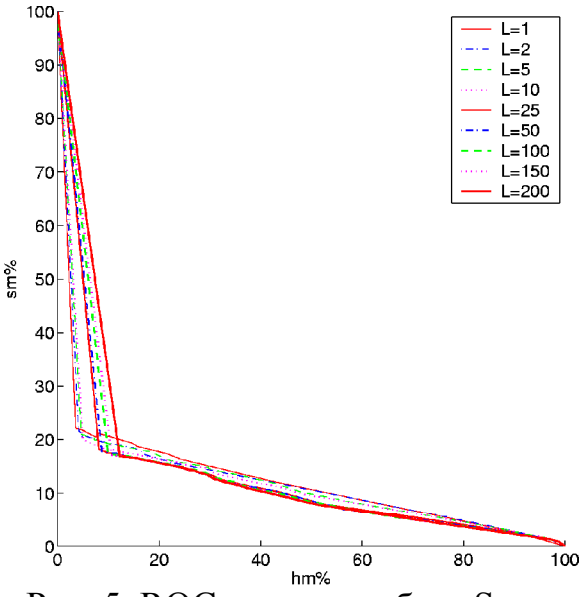

Рис. 5. ROC-крива для бази Spam Track 2006

різних значень порогу – *ROC*-криві. Видно, що при рівні *hm*%=5-10% виявляється близько 80% спаму для *Spam Track* 2006. Результати дозволяють говорити про придатність цього підходу як елементу технології виявлення спаму у великих поштових серверах.

Ефективність запропонованого методу пошуку схожих рядків було перевірено в задачі пошуку екзонів в ДНК хребетних за відомими прикладами екзонів інших організмів. За навчальну базу було вибрано набір *HMR*195 (195 послідовностей із 948 екзонами). Тестовою вибіркою була база *Burset-Guigo* (570 послідовностей, 2649 екзонів). У обох базах екзони розмічені. Навчальна база *P* складалася з нарізаних ковзним вікном підпослідовностей. Пошук екзонів здійснювався шляхом розпізнавання належності екзонам окремих нуклеотидів – пошуку в тестовій послідовності приблизних сусідів для екзонів з навчальної вибірки. Серед знайдених сусідів найближчий до запиту визначався за допомогою класичного алгоритму обчислення відстані редагування. Як інтегральна оцінка якості пошуку екзонів використовувалася стандартна міра «приблизна кореляція» (approximate correlation)

$$AC = 1/2(TP/(TP+FN) + TP/(TP+FP) + TN/(TN+FP) + TN/(TN+FN))–1,$$

де *TP* – true positives, *TN* – true negatives, *FP* – false positives, *FN* – false negatives. Запропонований метод порівнювався з підходом на основі класичного алгоритму редагування Costello (табл. 3). Метод на основі пошуку приблизних найближчих рядків за допомогою процедури *LSH*-ліс показав подібні результати, отриманим за методом Costello для невеликих значень K, але за час, менший майже в 750 раз. Застосування для задачі пошуку гіперчуттєвих сайтів (коротких рядків) дало результати на рівні відомих, що показало можливість застосування методу також в непередбаченій теорією області довжин послідовностей.

Таблиця 3

Результати пошуку екзонів

посилання	***AC***	**час на 1 *PC***
Costello	0.49	6 років (клас. алг.)
дис. робота	0.47	70 годин

Підхід на основі міркувань за прикладами було застосовано для виявлення вторгнень у комп'ютерних системах на основі класифікації належності користувальницьких сесій (послідовностей системних команд). Дослідження проводилися на даних з *UNIX*-сервера ФТІ НТУ України «КПІ» за період 3 роки – усього 717 користувачів, які виконали понад 23 млн. команд (навчальна вибірка – сесії за перші 403 дні, тестова – решта 268 днів). Проводилася нарізка навчальної і тестової сесій на ковзні вікна однакової довжини. Класифікація базувалася на визначенні користувача, вікна якого були найближчими до вікон тестової сесії. Частку правильно класифікованих сесій для різних значень ширини вікна *n*=10,20,40 і параметрів *K*=5,7, *L*=1,5,10 наведено в табл. 4. При збільшенні *L* досягається точність класифікації понад 90%. Витрати

Таблиця 4

Частка правильно класифікованих сесій

n	***K***	***L***		
		1	5	10
10	5	0.470	0.984	0.997
	7	0.431	0.945	0.987
20	5	0.440	0.942	0.983
	7	0.403	0.741	0.942
40	5	0.354	0.848	0.967
	7	0.230	0.390	0.836

часу на перевірку однієї сесії в середньому складають від 3 до 11 сек. (залежно від параметрів). Таким чином, запропонований метод є перспективним для попередньої онлайн-обробки сесій.

Розроблені програмні засоби використовуються в ряді організацій, що підтверджується відповідними актами.

ВИСНОВКИ

Отримані в дисертаційній роботі результати забезпечують розв'язання актуальної наукової задачі розробки методів нейромережевого розподіленого представлення послідовностей, а також їх пошуку та класифікації, для ефективної оцінки схожості та використання інформації про послідовності в системах штучного інтелекту, які застосовують моделі міркувань людини на основі прикладів. Розроблено, аналітично досліджено, а також програмно реалізовано методи розподіленого представлення та пошуку послідовностей. Ефективність розроблених методів підтверджено експериментальними дослідженнями на тестових і реальних даних при розв'язанні задач пошуку схожих послідовностей і класифікації інформації різного роду (тексти, ДНК, аудит-послідовності).

За результатами проведеного дослідження зроблено такі висновки:

1. Розроблений метод векторного представлення послідовностей забезпечує збереження їх схожості, лінійну (за довжиною вектора) складність апроксимації, можливість аналізу за допомогою теорії метричних вкладень. Аналітично і шляхом чисельних експериментів на штучних даних показана більш висока, порівняно з відомими результатами, точність апроксимації відстані редагування.

2. Розроблені, проаналізовані та реалізовані методи розподіленого представлення послідовностей, які за рахунок використання локально-чуттєвого хешування забезпечують малу ресурсоємність та сублінійний час пошуку приблизних найближчих послідовностей відносно розміру бази прикладів. Експериментальне дослідження якості пошуку на штучних даних показало достатність використання на практиці менших, ніж визначено аналітично, значень параметрів методу, що дозволяє зменшити ресурсоємність пошуку.

3. Запропонований метод нейромережевого розподіленого представлення послідовностей, який використовує рандомізацію векторних представлень і зв’язування елементів послідовності з їхніми позиціями, забезпечує уніфікацію формату представлення і можливість використання мір схожості векторних представлень для оцінки схожості послідовностей.

4. Розроблені методи пошуку схожих послідовностей за допомогою кластеризації за довжиною послідовностей та їх вирівнювання забезпечують пошук послідовностей різної довжини в реальних базах даних і розв'язання прикладних задач пошуку дублікатів і спаму на основі міркувань за прикладами за рахунок використання розподілених представлень і локально-чуттєвого хешування.

Ефективність і практична значимість розроблених методів підтверджені порівнянням отриманих результатів з відомими. Так, при пошуку дублікатів у базі РОМІП результат покращено на 20–40%, на базі *Reuters*-21578, – на рівні відомих. Перспективність застосування цих методів для виявлення спаму в великих поштових серверах показано на прикладі оцінки кількості спаму в колекціях електронних листів *TREC Spam Track* 2006 і 2005, де виявлено до 80% спаму при рівні неправильно класифікованих легальних повідомлень 5–10%.

5. Розроблені методи представлення і пошуку послідовностей забезпечують розв'язання прикладних задач класифікації ділянок ДНК і виявлення вторгнень, що підтверджує ефективність використання міркувань на основі прикладів для обробки послідовностей в реальних базах даних.

У задачі класифікації ділянок ДНК пошук екзонів з використанням підходу на основі міркувань за прикладами пришвидшено в 750 раз при збереженні якості на рівні відомих у цій області результатів. Розроблений метод пошуку послідовностей може застосовуватися при більш широкій області значень параметрів, ніж випливає з теоретичного аналізу, що експериментально показано на прикладі задачі пошуку некодуючих ділянок бета-глобіну при обробці коротких рядків. Запропонований метод є перспективним також для застосування в реальних системах виявлення вторгнень до комп'ютерних систем, що підтверджується результатом класифікації аудит-послідовностей, де отримано точність класифікації на рівні понад 90%.

6. Створені програмні засоби, що реалізують розроблені методи представлення і пошуку приблизних найближчих послідовностей, можуть застосовуватися як компоненти інформаційних технологій, або як самостійні модулі в системах класифікації й пошуку послідовностей. Практична значимість розробок підтверджується 3 актами впровадження.

ОСНОВНІ ПОЛОЖЕННЯ ДИСЕРТАЦІЇ ОПУБЛІКОВАНІ В ТАКИХ ПРАЦЯХ:

1. Резник А. Нейросетевая идентификация пользователей компьютерных систем / А. Резник, Н. Куссуль, А. Соколов // Кибернетика и вычислительная техника. – 1999. – Вып. 123. – С. 70–79.

2. Соколов А. Обнаружение аномалий с помощью марковских цепей переменного порядка // Искусственный интеллект. – 2002. – №4. – С. 74–83.

3. Соколов А. Современные модели обнаружения аномалий в компьютерных системах // Управляющие системы и машины. – 2004. – №5. – С. 67–73.

4. Гриценко В. Концепция и архитектура программного нейрокомпьютера SNC / В. Гриценко, И. Мисуно, Д. Рачковский, А. Соколов // Управляющие системы и машины. – 2004. – №3. – С. 3–14.

5. Мисуно И. Поиск текстовой информации с помощью векторных представлений / И. Мисуно, Д. Рачковский, С. Слипченко, А. Соколов // Проблемы программирования – 2005 – №4 – С.50–67.

6. Мисуно И. Модульный программный нейрокомпьютер SNC: реализация и применение / И. Мисуно, Д. Рачковский, Е. Ревунова, А. Соколов // Управляющие системы и машины. – 2005. – №2. – С. 74–85.

7. Sokolov A. Approaches to sequence similarity representation / A. Sokolov, D. Rachkovkij // Int. Journal of Information Theories and Applications. – 2006. – V. 13, №3. – P. 272–278.

8. Соколов А. Векторные представления для эффективного сравнения и поиска похожих строк // Кибернетика и системный анализ. – 2007. – №4. – С. 18–38.

9. Sokolov A. Searching for Nearest Strings with Neural-like String Embedding // Int. Journal of Information Theories and Applications. — 2007. — V. 14. – №3. – P. 294–299.

10. Соколов А. Исследование ускоренного поиска близких текстовых последовательностей с помощью векторных представлений // Кибернетика и системный анализ. – 2008. – №4. – С. 32–47

11. Соколов А. Рандомизированное вложение расстояния редактирования в задачах поиска генов и обнаружения вторжений // Системные технологии. – 2008. Вып. 2 (55). – С. 126–139.

12. Misuno I. SNC: The software neurocomputer with modular architecture / I. Misuno, D. Rachkovskij, E. Revunova, A. Sokolov // Междунар. конф. "Проблемы нейрокибернетики". – Ростов-на-Дону, Россия, 2002 – Т. 2. – С. 109–113.

13. Sokolov A. On Handling Replay Attacks in Intrusion Detection Systems // Proc. of X-th Int. Conf. Knowledge-Dialogue-Solution. – Varna, Bulgaria, 2003. – P. 258–265.

14. Sokolov A. An adaptive detection of anomalies in user's behavior // Proc. of the Int. Joint Conf. on Neural Networks. – Portland, USA, 2003. V. 4. – P. 2443–2447.

15. Sokolov A. Some approaches to distributed encoding of sequences / A. Sokolov, D. Rachkovskij // Proc. of XI-th Int. Conf. Knowledge-Dialogue-Solution. – Varna, Bulgaria, 2005. – V. 2. – P. 522–528.

16. Мисуно И. Обработка текстовой информации с помощью векторных представлений / И. Мисуно, Д. Рачковский, С. Слипченко, А. Соколов // Международный семинар по индуктивному моделированию МСИМ-05 (IWIM-05). – Киев: 2005. – Т. 1 –С. 230–236.

17. Рачковский Д. Концепция и методы нейросетевого распределенного представления информации в задачах искусственного интеллекта / Д. Рачковский, И. Мисун, Е. Ревунова, А. Соколов // Междунар. конф. "Проблемы нейрокибернетики". – Ростов-на-Дону, Россия: 2005. – Т. 2. – С. 30–33.

18. Sokolov A. Nearest string by neural-like encoding // Proc. of the XII-th Int. Conf. Knowledge-Dialogue-Solution. – Varna, Bulgaria, 2006. – C. 101–106.

АНОТАЦІЯ

Соколов А. М. Методи нейромережевого розподіленого представлення та пошуку схожих символьних послідовностей в задачах класифікації на основі міркувань за прикладами. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.26 – системи та засоби штучного інтелекту. – Міжнародний науково-навчальний центр інформаційних технологій та систем НАН України і МОН України, Київ, 2008.

Дисертаційна робота присвячена розробці та дослідженню методів розподіленого представлення символьних послідовностей на основі розробленого q-грамного методу вкладення простору із класичною метрикою редагування в векторний простір з метрикою ℓ_1.

Розроблено детермінований q-грамний метод вкладення простору символьних послідовностей фіксованої довжини над скінченним алфавітом з класичною метрикою редагування в векторний простір з метрикою ℓ_1. Завдяки використанню підрядків змінної довжини поліпшено якість апроксимації відстані редагування порівняно з відомими методами, що продемонстровано аналітично шляхом застосування апарату графів де Брейна і чисельно на штучних даних.

На основі розробленого детермінованого методу запропоновано локально-чуттєву функцію для класичної відстані редагування, що продукує розподілене представлення послідовностей, яке забезпеченує малу ресурсоємність і можливість створення ефективної процедури пошуку приблизних найближчих послідовностей за сублінійний до розміру бази час – базової операції підходу на основі міркувань за прикладами. Чисельні експерименти показали можливість використання менших, ніж отриманих теоретично, значень параметрів процедури. Розроблені методи пошуку схожих послідовностей за допомогою кластеризації за довжиною послідовностей та вирівнювання довжини, що дало змогу виконувати пошук приблизних найближчих послідовностей різної довжини у реальних базах даних.

Метод застосовано у ряді прикладних задач, де отримано результати кращі відомих або результати на рівні відомих, але за значно менший час. Усі методи реалізовано як програмні засоби, які можуть використовуватися в системах штучного інтелекту.

Ключові слова: представлення даних, розподілені векторні представлення послідовностей, вкладення відстані редагування, пошук дублікатів, виявлення спаму, пошук генів, виявлення вторгнень, нейронні мережі.

АННОТАЦИЯ

Соколов А. М. Методы нейросетевого распределенного представления и поиска сходных символьных последовательностей в задачах классификации на основе рассуждений по примерам. – Рукопись.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.23 – системы и средства искусственного интеллекта. – Международный научно-учебный центр информационных технологий и систем НАН Украины и МОН Украины, Киев, 2008.

Диссертационная работа посвящена разработке и исследованию методов распределенного представления символьных последовательностей на основе

разработанного q-граммного метода вложения пространства с классической метрикой редактирования в векторное пространство с метрикой ℓ_1.

Разработан детерминированный q-граммный метод вложения пространства символьных последовательностей фиксированной длины над конечным алфавитом с классической метрикой редактирования в векторное пространство с метрикой ℓ_1. Благодаря использованию подстрок переменной длины улучшено качество аппроксимации расстояния редактирования по сравнению с известными методами, что продемонстрировано аналитически путем использования аппарата графов де Брейна и численно путем экспериментов на искусственных данных.

На основе разработанного детерминированного метода предложена локально-чувствительная функция для классического расстояния редактирования, продуцирующая распределенное представление последовательностей, что обеспечило малую ресурсоемкость и возможность создания эффективной процедуры поиска приближенных ближайших последовательностей за сублинейное к размеру базы время – базовой операции подхода на основе рассуждений по аналогии. Этим достигается ускорение поиска приближенных ближайших последовательностей в больших базах. Путем численных экспериментов показана возможность использования на практике меньших, чем полученные теоретически, значений параметров процедуры, что позволяет уменьшить требования к ресурсам.

Разработаны методы поиска сходных последовательностей с помощью кластеризации по длине последовательностей и выравнивания длины, что позволило производить поиск приближенных ближайших последовательностей разной длины в реальных базах данных.

Метод применен в ряде прикладных задач. В задаче поиска дубликатов в базе РОМИП результат улучшен на 20–40%, в базе *Reuters*-21578 – получены результаты на уровне известных. Показана перспективность использования разработанного метода поиска для обнаружения спама на основе рассуждений по примерам в крупных почтовых серверах. В базе электронных писем *TREC Spam Track* 2006 обнаружено до 80% спама при уровне неправильно классифицированных легальных сообщений 5–10%. В задаче классификации участков ДНК благодаря применению разработанного метода поиск экзонов ускорен в 750 раз при сохранении качества на уровне известных результатов работ, использующих подход на основе рассуждений по примерам. Разработанный метод поиска последовательностей может применяться в более широкой области значений параметров, чем следует из теоретического анализа, что экспериментально показано в задаче поиска некодирующих участков бета-глобина. Разработанный метод перспективен для применения в системах обнаружения вторжений, что подтверждается результатом классификации аудит-сессий пользователей компьютерных систем, где получена точность классификации на уровне более 90%.

Методы представления и поиска последовательностей реализованы в виде

программных средств, которые могут быть использованы в качестве компонентов информационных технологий или как самостоятельные модули в системах искусственного интеллекта, использующих классификацию и поиск последовательностей.

*Ключевые слова***:** представление данных**,** распределенные векторные представления последовательностей, вложение расстояния редактирования, поиск дубликатов, обнаружение спама, поиск генов, обнаружение вторжений, нейронные сети.

ABSTRACT

Sokolov A. M. Neural-like distributed representations and similar sequences search for classification with case-based reasoning. – Manuscript.

Ph.D. thesis for acquiring scientific degree of Candidate of Technical Science on specialization 05.13.23 – Systems and Means of Artificial Intelligence. – International research and training center for informational technologies and systems, National Academy of Sciences of Ukraine and Ministry of Science and Education of Ukraine, Kyiv, 2008.

The dissertation is devoted to the development and investigation of methods for distributed representations of symbol sequences, based on the developed q-gram method of embedding sequence space endowed with classic edit metrics to the ℓ_1 vector space.

A deterministic q-gram method was developed, that embeds the space of symbol fixed length sequences over a finite alphabet endowed with classic edit metrics to the vector space with ℓ_1-metrics. The usage q-grams of different lengths has improved approximation quality compared to known methods. This was shown analytically by using de Bruin graphs for sequence representation.

Based on the developed deterministic method, a locality-sensitive hash-function was proposed for the classic edit distance. The function produces distributed representation of sequences, leading to a lower resource requirements and an effective approximate nearest sequence search procedure in sublinear time (on the number of sequences in the base). This search is a basic operation involved in cased-based reasoning approaches. Numeric experiments showed that it is possible to use lower than theoretically derived parameter values. Search methods for similar sequences of different length were also developed. This made possible the approximate sequence search in real-world datasets containing sequences of different length.

The method was applied to a number of real-world applications, where either results improving quality of the known ones were obtained or results that improve search time, while showing comparable performance. All methods were implemented in software that can be used in artificial intelligence systems.

Keywords: data representation, distributed vector representations of sequences, edit distance embedding, duplicates search, spam detection, gene search, intrusion detection, neural networks.