Харківський національний університет радіоелектроніки

Волкова Валентина Володимирівна

УДК 004.912:004.8

МЕТОДИ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ ПОЛІТЕМАТИЧНИХ ТЕКСТОВИХ ДОКУМЕНТІВ

05.13.23 – системи та засоби штучного інтелекту

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2010

Дисертацією є рукопис.

Роботу виконано у Харківському національному університеті радіоелектроніки Міністерства освіти і науки України.

Науковий керівник кандидат технічних наук, доцент
Рябова Наталія Володимирівна,
Харківський національний університет радіоелектроніки, доцент, виконуючий обов'язки завідувача кафедри штучного інтелекту.

Офіційні опоненти: доктор технічних наук, професор
Єрохін Андрій Леонідович,
Харківський національний університет внутрішніх справ, начальник навчально-наукового інституту психології, менеджменту та інформаційних технологій, м. Харків;

доктор технічних наук, професор
Асєєв Георгій Георгійович,
Харківська державна академія культури, завідувач кафедри інформаційних технологій, м. Харків.

Захист відбудеться «____» _____ 2010 р. о ____ годині на засіданні спеціалізованої вченої ради Д 64.052.01 у Харківському національному університеті радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна, 14.

З дисертацією можна ознайомитися у бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна, 14.

Автореферат розісланий «____» _____ 2010 р.

Вчений секретар
спеціалізованої вченої ради С.Ф. Чалий

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. В умовах стрімкого розвитку web-простору та збільшення об'єму інформації, що в ньому зберігається, з'являються нові проблеми, пов'язані з обробкою такої великої кількості даних. Зокрема, це проблема релевантного пошуку інформації в Web; категоризація та рубрикація документів (для каталогових служб); визначення відстані між персональними профілями.

Вирішення таких задач можливе за допомогою методів кластеризації, що дозволяють розбивати множину web-сторінок на категорії залежно від характеристик даних документів. Слід зазначити, що більшість web-сторінок є політематичними текстовими документами, тобто одночасно відносяться до декількох тематик, що ускладнює процес їх кластеризації.

На сьогодні існує велика кількість методів кластеризації текстових документів, але більшість з них не враховують повною мірою наявності кластерів, що перетинаються при класифікації без учителя, тобто такі ситуації, коли один і той самий документ може одночасно належати до декількох категорій. Також такі методи обмежені за засобами обробки вхідних даних, тобто не можуть послідовно кластеризувати вхідні дані. Це є істотним недоліком. Таким чином, розробка засобів, що дозволяють вирішити ці проблеми, є важливим завданням.

У зв'язку з цим актуальною є задача розробки методів нечіткої адаптивної кластеризації політематичних текстових документів.

Вагомий внесок у розвиток методів кластеризації текстових документів, у тому числі нечіткої, та видобування з них знань внесли такі вчені, як: Айвазян С.А., Загоруйко М.Г., Bezdek J.C., Kohonen T., Höppner F., Vuorimaa P., Klawonn F., Kruse R., Krishnapuram R., Keller J., Desjardins G., Zhang C., Курейчик В.М., Рутковська Д. та інші.

Незважаючи на істотні досягнення в галузі кластеризації, залишається ряд задач, які ще далекі від свого остаточного вирішення. До таких задач і відноситься задача кластеризації політематичних текстових документів з урахуванням кластерів, що перетинаються, та послідовної подачі документів на обробку. Ця задача може бути вирішена за допомогою створення методів адаптивної нечіткої кластеризації на основі методів обчислювального інтелекту.

Зв'язок роботи з науковими програмами, планами, темами. Робота виконана на кафедрі штучного інтелекту Харківського національного університету радіоелектроніки відповідно до плану науково-дослідних робіт у межах держбюджетних тем: № 195 «Розробка теоретичних засад, методів та моделей інтелектуальної обробки інформації та менеджменту знань у системах розподіленого штучного інтелекту» (№ ДР 0106U003286), №219 «Розробка Web-орієнтованої системи для підтримки процедур акредитації та ліцензування вищих навчальних закладів України» (№ ДР 0108U010139), № 214 «Синтез методів обробки інформації за умов невизначеності на основі самонавчання та м'яких обчислень» (№ ДР 0107U003028), в яких здобувачка взяла участь як виконавець.

В межах зазначених НДР автором введено поняття політематичного текстового документа, запропоновано метод кластеризації політематичних тестових документів на основі генетичного алгоритму зі штучним відбором; розроблено модель адаптивної нечіткої нейронної мережі, що самоорганізується, та методи її навчання; розроблено модель системи кластеризації політематичних текстових документів на основі запропонованої нейронної мережі.

Мета і завдання дослідження. Метою дисертаційної роботи є розробка методів адаптивної нечіткої кластеризації політематичних текстових документів для поліпшення

якості кластеризації за наявності кластерів, що перетинаються, та забезпечення можливості обробки документів в режимі реального часу. Досягнення поставленої мети здійснюється розв’язанням таких основних задач:

– аналіз основних методів кластеризації політематичних текстових документів;

– розробка моделі адаптивної нечіткої нейронної мережі, що самоорганізується, яка б дозволяла виконувати кластеризацію документів у послідовному режимі;

– розробка моделі системи кластеризації політематичних текстових документів на основі запропонованої нейронної мережі для врахування наявності кластерів, що перетинаються, а також послідовної обробки політематичних текстових документів;

– розробка методів навчання для адаптивної нечіткої нейронної мережі, що самоорганізується;

– розробка методу навчання для нейронних мереж, що самоорганізуються, який дозволяє підвищити швидкість обробки інформації, поліпшити якість кластеризації за наявності кластерів, що перетинаються, шляхом використання нечіткого виведення.

– розробка методу автоматичної кластеризації політематичних текстових документів на основі генетичного алгоритму зі штучним відбором;

– розробка структури та функцій інструментальних засобів вирішення прикладних задач.

Об’єктом дослідження є процес класифікації та кластеризації політематичних текстових документів у інтелектуальних системах обробки документів.

Предметом дослідження є методи адаптивної нечіткої кластеризації політематичних текстових документів.

Методи дослідження. Основними методами дослідження є методи штучного інтелекту: теорія штучних нейронних мереж, за допомогою якої були синтезовані нова модель і методи її навчання, що дозволяють виконувати нечітку кластеризацію політематичних текстових документів; процедури генетичної оптимізації, а також послідовного комплекс-методу пошуку екстремуму функцій багатьох змінних, що дозволили створити метод кластеризації великого обсягу даних; принципи обробки природно-мовної інформації, які дозволили подати політематичні текстові документи у придатному для машинної обробки вигляді. Експериментальні дослідження проводилися в лабораторних умовах і на реальних об’єктах.

Наукова новизна отриманих результатів. У процесі вирішення поставлених завдань отримано такі наукові результати:

1. Вперше запропоновано модель адаптивної нечіткої нейронної мережі, що самоорганізується, яка відрізняється від інших нейронних мереж використанням спеціальних нелінійних обчислювачів, які дозволяють знаходити рівні належності вхідних образів документів до кластерів, що дає можливість підвищити якість кластеризації у режимі послідовної обробки даних.

2. Вперше розроблені рекурентні ймовірнісний та можливісний методи навчання для запропонованої адаптивної нечіткої нейронної мережі, що самоорганізується, які відрізняються від інших методів навчання нейронних мереж наявністю фаззіфікатора. Це дозволило в процесі роботи методів виявляти нові кластери, а також коректно оцінювати викиди та накопичення вибірки в реальному часі, по мірі надходження.

3. Вперше запропоновано модель системи кластеризації політематичних текстових документів, яка відрізняється від існуючих моделей наявністю двох паралельно працюючих адаптивних нечітких нейронних мереж, що самоорганізуються. Це дозволило в процесі послідовної обробки політематичних текстових документів виявляти нові кластери та відновлювати такі, що перетинаються.

4. Вперше розроблено метод автоматичної кластеризації політематичних текстових документів на основі генетичного алгоритму зі штучним відбором. На відміну від існуючих еволюційних алгоритмів, в даному методі в процедури генетичної оптимізації вводяться оператори комплекс-пошуку. Це дозволило знаходити екстремум довільних функцій великої кількості аргументів в умовах істотної невизначеності про характер цих функцій та використати запропонований метод в Genetic Mining великих масивів текстових документів.

5. Набув подальшого розвитку метод навчання для нейронних мереж, що самоорганізуються, який на відміну від правила навчання «переможець отримує все», містить оцінку рівня належності спостережень кожному з наявних кластерів на основі функції належності. Це дозволило підвищити швидкість обробки інформації, поліпшити якість кластеризації за наявності кластерів, що перетинаються, шляхом використання нечіткого виведення.

Практичне значення отриманих результатів полягає в тому, що розроблені методи нечіткої кластеризації політематичних текстових документів були використані під час дослідження ефективності відображення та обробки результатів пошуку в інформаційно-пошуковій системі, а також створення модуля кластеризації результатів її роботи.

На основі розробленого в дисертаційні роботі методу кластеризації політематичних текстових документів, що базується на адаптивній нечіткій нейронній мережі, що самоорганізується, з рекурентними ймовірнісним та можливісним методами навчання, створено модуль кластеризації результатів роботи інформаційно-пошукової системи, які були впроваджені в науковій бібліотеці ХНУРЕ та науково-технічній бібліотеці Національного аерокосмічного університету ім. М.Є. Жуковського «Харківський авіаційний інститут» (акти впровадження відповідно від 15.09.2009 р. та 30.09.2009 р.).

Результати дисертаційної роботи також впроваджено у навчальному процесі на кафедрі штучного інтелекту ХНУРЕ у дисципліни «Штучні нейронні мережі: архітектури, навчання, застосування», «Нейромережеві методи обчислювального інтелекту», «Системи обробки природно-мовної інформації» (акт впровадження від 21.10.09 р.), а також у науково-дослідних роботах Харківського національного університету радіоелектроніки, що підтверджено актом від 07.09.2009 р.

Запропоновані методи можуть бути використані для спрощення роботи багатьох сервісів Інтернет, при створенні інформаційно-пошукових систем нового рівня, здатних у процесі послідовної обробки політематичних текстових документів розбивати їх на категорії, причому один і той самий документ може одночасно належати до декількох категорій. Також розроблені методи найбільш актуально й ефективно використовувати в таких областях:

- розміщення текстових документів за категоріями на комп'ютері;
- створення онтологій предметних галузей (виділення концептів);
- пошук експертів з певної предметної галузі у web-просторі.

Особистий внесок здобувача. Всі результати дисертації отримано автором особисто. У роботах, опублікованих зі співавторами, здобувачці належать: у [2,8,9] – модель адаптивної нечіткої нейронної мережі, що самоорганізується, призначеної для кластеризації політематичних текстових документів у послідовному режимі, рекурентні ймовірнісний та можливісний методи її навчання; [1,8] – модель системи адаптивної нечіткої кластеризації політематичних текстових документів на основі двох паралельно працюючих адаптивних нечітких нейронних мереж, що самоорганізуються; [4] – комбінований метод навчання нейронних мереж, що самоорганізуються, з нечітким виведенням; [3] – метод автоматичної кластеризації текстових документів на основі генетичного алгоритму зі штучним відбором; [6] – аналіз основних методів побудови онтологій предметних галузей, аналіз методів та моделей інтелектуальної обробки текстів у задачах онтологічного інжинірингу.

Апробація результатів дисертації. Результати дисертаційної роботи доповідалися і обговорювалися на 9-му Міжнародному молодіжному форумі «Радіоелектроніка і молодь у XXI столітті» (м. Харків, 2005); на 2-му Міжнародному радіоелектронному форумі «Прикладна радіоелектроніка. Стан та перспективи розвитку» (м. Харків, 2005); на 2-й Міжнародній конференції «Сучасні інформаційні системи. Проблеми та тенденції розвитку» (м. Харків, 2007); на 6-й Міжнародній науково-практичній конференції «Математичне та програмне забезпечення інтелектуальних систем» (м. Дніпропетровськ, 2008); на першій Факультетській науково-практичній молодіжній школі-семінарі студентів, аспірантів та молодих вчених «Інформаційні інтелектуальні системи» (м. Харків, 2008).

Публікації. За результатами досліджень опубліковано 9 робіт, з них 4 статті у виданнях, що входять до переліку ВАК України та 5 публікацій у збірниках праць міжнародних наукових конференцій.

Структура й обсяг дисертаційної роботи. Дисертація складається зі вступу, чотирьох розділів, висновків, додатків та списку використаних джерел. Загальний обсяг роботи складає 157 сторінок, у тому числі 15 рисунків за текстом, 4 додатки на 5 сторінках, список використаних джерел зі 150 найменувань на 17 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність дисертаційної роботи, сформульовано основну мету і задачі дослідження, наведено відомості щодо наукової новизни отриманих у дисертації результатів, визначено їх практичну цінність, наведено відомості про апробацію та впровадження результатів.

Перший розділ містить огляд предметної галузі і постановку задач дисертаційного дослідження. Проведено аналіз поняття «політематичний текстовий документ», введено його визначення, основні властивості та приклади політематичних текстових документів. Розглянуто питання попередньої обробки політематичних текстових документів для подальшої їх обробки програмними засобами. Розглянуто процес формування простору ознак та основні методи його скорочення.

Політематичним називається текстовий документ довільного розміру, який одночасно відноситься до декількох тематик. Монотематичним текстовим документом називається текстовий документ довільного розміру, який відноситься до однієї тематики. Тематикою текстового документа називається деяке суб'єктивне уявлення людини, користувача пошукової системи про предметну галузь, що розглядається в тексті, про його основний зміст. Політематичні документи містять в собі багатоаспектні цілі, відображуючи цим інформативні потреби різних користувачів. Текстові репозиторії, в яких зазвичай зберігаються такі документи, як правило характеризуються декількома тематиками (наприклад, в біомедичних статтях часто використовуються математичні і статистичні методи, технології штучного інтелекту, хімії, біології тощо).

Політематичний текстовий документ може бути поданий у вигляді документа, що складається з менших текстових модулів, кожен з яких відноситься до однієї або декількох тематик документа. Отже, текстовий документ розбивається на неподільні ділянки, які можуть бути розпізнані на різних рівнях в логічній структурі документа (наприклад, розділ, параграф).

На сьогодні існує небагато методів, які дозволяють обробляти політематичні текстові документи – виконувати їх класифікацію чи кластеризацію, вони мають ряд недоліків, пов'язаних з обчислювальною складністю та неможливістю знаходження в процесі своєї роботи нових класів/кластерів. Метою класифікації політематичних текстових документів є віднесення одного й того ж текстового документа до більш ніж однієї тематики. Мета кластеризації політематичних текстових документів – кластеризувати текстові документи так, щоб кожний документ відносився до більш ніж до одного кластеру.

Також розглянуто основні методи кластеризації політематичних документів, проведено їх оцінку та оцінку результатів кластеризації, виконаної на їх основі.

У **другому розділі** запропоновано модель адаптивної нечіткої нейронної мережі, що самоорганізується, а також рекурентні ймовірнісний та можливісний методи навчання для неї; модель системи кластеризації політематичних текстових документів на основі розробленої моделі адаптивної нечіткої нейронної мережі, що самоорганізується; метод навчання для нейронних мереж, що самоорганізуються.

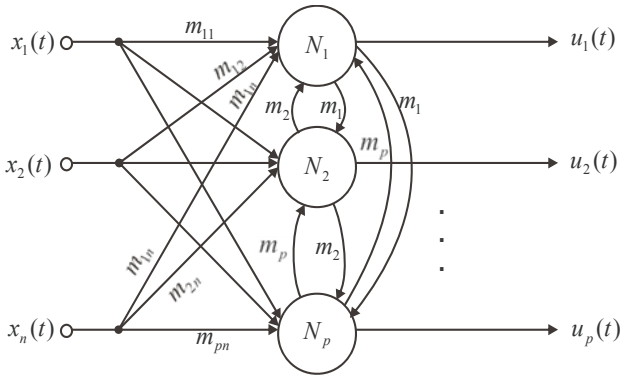

Рис.1. Адаптивна нечітка нейронна мережа

Модель адаптивної нечіткої нейронної мережі, що самоорганізується, наведена на рис. 1 та містить єдиний

прихований шар нейронів N_i, $i = 1.2\ldots, p,$ що відрізняються від традиційних лінійних асоціаторів, утво-рюючих карту Кохонена. На рецепторний шар мережі надходять образи документів, що підлягають кластеризації, у вигляді $(n \times 1)$-векторів ознак $x(t) = (x_1(t), x_2(t), ..., x_n(t))^T$, де $t = 1,2,...,V$ має зміст номера образу в навчальній вибірці або поточного дискретного часу. При цьому самі вектори ознак $x(t)$ формуються на основі зрізаних гістограм частот зустрічаності окремих слів у політематичних текстових документах, що підлягають обробці.

Синаптичні ваги $m_{ij}, i = 1,2,..., p; j = 1,2,..., n$ визначають координати центроїдів p кластерів $m_i(t)$, що взаємно перетинаються, а виходом мережі, на відміну від карти Кохонена, що самоорганізується, вихідний сигнал якої визначається лише нейроном-переможцем, є $(p \times 1)$-вектор $u(t) = (u_1(t), u_2(t), ..., u_p(t))^T$, що визначає рівень належності образу $x(t)$ до кожного з p кластерів, що формуються, та обчислюваний нейронами N_i. По латеральним зв'язкам нейрони обмінюються координатами $m_i(t)$, необхідними для обчислення належностей $u_i(t)$.

В основі самонавчання лежить рекурентний ймовірнісний метод кластеризації, що базується на оптимізації цільової функції виду:

$$E(u_i, m_i) = \sum_{t=1}^{V} \sum_{i=1}^{p} u_i^{\beta}(t) \left\| x(t) - m_i \right\|^2 \tag{1}$$

за обмежень:

$$\sum_{i=1}^{p} u_i(t) = 1,\ t = 1,2,...,V, \tag{2}$$

$$0 \le \sum_{t=1}^{V} u_i(t) \le V,\ i = 1,2,..., p, \tag{3}$$

де $u_i(t) \in [0,1]$; V – кількість образів документів, що обробляються; β – додатний параметр («фаззіфікатор»), який визначає нечітку межу між кластерами, та впливає на рівень нечіткості у кінцевому розбитті даних по кластерах.

Рекурентний ймовірнісний метод самонавчання адаптивної нечіткої нейронної мережі, що самоорганізується, може бути записаний у формі процедури стохастичної апроксимації:

$$\begin{cases} m_i^{PR}(t+1) = m_i^{PR}(t) + \dfrac{(u_i^{PR})^{\beta}(t)}{t+1}(x(t+1) - m_i^{PR}(t)), i = 1,2,..., p, \\ u_i^{PR}(t+1) = \dfrac{(\left\| x(t+1) - m_i^{PR}(t+1) \right\|^2)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (\left\| x(t+1) - m_l^{PR}(t+1) \right\|^2)^{\frac{1}{1-\beta}}}, \end{cases} \tag{4}$$

що є простою в обчисленні та дає можливість послідовної обробки даних і нечітке розбиття на кластери.

Як альтернатива методу (4) запропоновано рекурентний можливісний метод навчання адаптивної нечіткої нейронної мережі, що самоорганізується. В основі цього правила самонавчання лежить оптимізація локальної цільової можливісної функції:

$$E_t(u_i(t), m_i(t)) = \sum_{i=1}^{p} u_i^{\beta}(t) d^2(x(t), m_i) + \sum_{i=1}^{p} \mu_i (1 - u_i(t))^{\beta}, \quad (5)$$

де $u_i(t) \in [0,1], \beta$ – фаззіфікатор, $d^2(x(t), m_i) = \left\| x(t) - m^i \right\|^2$ – квадрат евклідової відстані між образом та центроїдом, $\mu_i > 0$ – скалярний параметр, що визначає відстань, на якій рівень належності приймає значення 0,5, тобто якщо $d^2(x(t), m_i) = \mu_i$, то $u_i(t) = 0{,}5$.

Рекурентний можливісний метод самонавчання адаптивної нечіткої нейронної мережі, що самоорганізується, може бути записаний у вигляді

$$\begin{cases} m_i^{POS}(t+1) = m_i^{POS}(t) + \alpha(t)\left(u_i^{POS}(t)\right)^{\beta}\left(x(t+1) - m_i^{POS}(t)\right), i = 1,2,..., p, \\ u_i^{POS}(t+1) = \left(1 + \left(\dfrac{d^2\left(x(t), m_i^{POS}(t+1)\right)}{\mu_i(t)}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \\ \mu_i(t+1) = \dfrac{\sum\limits_{p=1}^{t+1}\left(u_i^{POS}(p)\right)^{\beta} d^2\left(x(p), m_i^{POS}(t+1)\right)}{\sum\limits_{p=1}^{t+1}\left(u_i^{POS}(p)\right)^{\beta}}, \end{cases} \quad (6)$$

де $\alpha(t)$ – параметр кроку пошуку, що впливає на швидкість збіжності, та обирається відповідно до умов А.Дворецького.

Слід відзначити, що важливою властивістю цього методу самонавчання є умова $\sum_{i=1}^{p} u_i^{POS}(t) \neq 1,$ яка на відміну від процедур ймовірнісної нечіткої кластеризації дозволяє знаходити нові кластери в процесі навчання нейронної мережі, а також коректно оцінювати викиди та накопичення вибірки в реальному часі, по мірі надходження.

Паралельне застосування адаптивних ймовірнісного та можливісного методів веде до об’єднаної процедури (для $\beta = 2$):

$$
\begin{cases}
m_i^{PR}(t+1) = m_i^{POS}(t) + \alpha(t)\left(u_i^{POS}(t)\right)^2 \times \left(x(t+1) - m_i^{POS}(t)\right), \\
u_i^{PR}(t+1) = \dfrac{\left\|x(t+1) - m_i^{PR}(t+1)\right\|^{-2}}{\sum\limits_{l=1}^{p}\left\|x(t+1) - m_l^{PR}(t+1)\right\|^{-2}}, \\
m_i^{POS}(t+1) = m_i^{PR}(t+1) + \alpha(t)\left(u_i^{POS}(t)\right)^2 \times \left(x(t+1) - m_i^{PR}(t+1)\right), \\
u_i^{POS}(t+1) = \dfrac{\mu_i(t)}{\mu_i(t) + \left\|x(t+1) - m_i^{POS}(t+1)\right\|^2}, \\
\mu_i(t+1) = \dfrac{\sum\limits_{p=1}^{t+1}\left(u_i^{POS}(p)\right)^2\left\|x(p) - m_i^{POS}(t+1)\right\|^2}{\sum\limits_{p=1}^{t+1}\left(u_i^{POS}(p)\right)^2}.
\end{cases}
\tag{7}
$$

Процедура (7) є методом самонавчання нейро-нечіткої системи кластеризації політематичних текстових документів, утвореної двома паралельно працюючими адаптивними нечіткими нейронними мережами (AFSONN), що самоорганізуються, які обмінюються між собою інформацією, як це показано на рис. 2.

Ознакою коректного відновлення прототипів за допомогою методу (7) є виконання нерівності

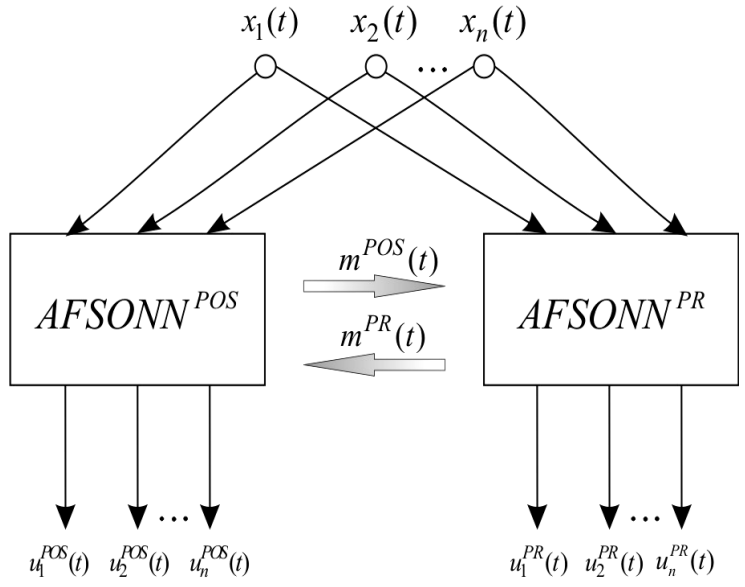

Рис.2. Система кластеризації політематичних текстових документів

$$
\sum_{l=1}^{p} d^2\left(m_l^{PR}(t), m_l^{POS}(t)\right) \le \varepsilon,
$$

де параметр ε визначає точність кластеризації.

Вводячи деяке порогове значення ξ ($\xi \approx 0{,}1 \div 0{,}2$) та контролюючи виконання умови $\sum\limits_{i=1}^{p} u_i^{POS}(t) \le \xi$, можна говорити про появу нового кластера. Слід відзначити, що якість кластеризації залежить від вибраного простору ознак та обсягу даних.

Навчена адаптивна нечітка нейронна мережа, що самоорганізується, може бути використана для кластеризації спостережень, що надходять на її входи $x(q)$, $q = N+1, N+2, \ldots$ У випадку кластерів, що перетинаються, рішення типу «переможець отримує все» може виявитися некоректним. У

зв'язку з цим запропоновано метод навчання для нейронних мереж, що самоорганізуються, який містить оцінку рівня належності спостережень кожному з наявних кластерів на основі функції належності

$$\mu_{\hat{w}l}(x(q)) = \frac{1+\cos(\hat{w}_l, x(q))}{2}, \tag{8}$$

обмеженої інтервалом $0 \le \mu_{\hat{w}l}(x) \le 1$.

Для оцінки належності поданого вектора $x(q)$ кожному з кластерів використовується нормований вираз

$$\mu_{\hat{w}l}(x(q)) = \frac{y_l(x(q))}{\sum_{l=1}^{m} y_l(x(q))}. \tag{9}$$

Можна побачити, що функція належності $\mu_{\hat{w}l}(x)$ є окремим випадком квадратичної радіально-базисної функції

$$\phi_l(x) = \max\left\{0{,}1 - \frac{\left\|x-\hat{w}_l\right\|^2}{\sigma_l^{\,2}}\right\} = \max\left\{0{,}1 - 2\frac{1-\hat{w}_l^{\,T}x}{\sigma_l^{\,2}}\right\}, \tag{10}$$

що набуває під час виконання умови нормування вигляду

$$\phi_l(x) = \frac{1+\hat{w}_l^{\,T}x}{2}. \tag{11}$$

Отже, рецепторним полем функції (9) є гіперсфера. Крім того, у випадку, якщо в навчальній вибірці апріорі присутня класифікація спостережень, то для подальшого більш точного налаштування радіусів кластерів може бути використана проста градієнтна процедура оптимізації.

Весь процес самоорганізації має дві часові фази: початкова фаза впорядкування, в якій відбувається топологічне розбиття вхідного простору, і подальша фаза збіжності, в якій здійснюється точне налаштування синаптичних ваг. Після закінчення цього процесу нейронна мережа в принципі може вирішувати поставлені задачі без уточнення ваг, проте, якщо з'явиться вхідний образ, який не буде віднесений ні до одного зі сформованих кластерів, картою має бути утворений додатковий нейрон в шарі Кохонена, що несе інформацію про цей образ, при цьому бажано, щоб знову розпочався процес самонавчання.

У **третьому розділі** запропоновано метод автоматичної кластеризації політематичних текстових документів на основі генетичних алгоритмів. Пропонується ввести в процедури генетичної оптимізації елементи штучного відбору, відмінного від загальноприйнятих стратегій елітизму (бджолина сім'я, модель островів тощо).

В основі запропонованого генетичного методу зі штучним відбором лежить синтез холландівського генетичного алгоритму з ідеями адаптаційної оптимізації, а також послідовного комплекс-методу пошуку екстремуму функцій багатьох змінних. При цьому в кожен момент часу поточна популяція ототожнюється з «хмарою» – комплексом точок у просторі змінних-факторів, а крім традиційних генетичних операторів мутації, кросовера та інверсії додатково вводяться оператори комплекс-пошуку такі, як відображення, розтягнення

та стиснення. При цьому на відміну від традиційного комплекс-методу пропонується відображати не одну найгіршу вершину комплексу, а цілу множину найгірших особин популяції.

При використанні генетичного методу зі штучним відбором для вирішення задачі кластеризації колекції політематичних текстових документів хромосома символізує можливе розбиття документів по кластерах. Кожен ген відповідає документу; значення гена (алель) дорівнює номеру кластера. Гени (документи) з однаковими алелями відносяться до одного кластера. Множина хромосом утворює популяцію зі змінним числом особин, а як функція пристосованості використовується, як правило, міра ефективності рішення, що кодується хромосомою.

У загальному випадку процедура оптимізації на основі звичайного послідовного комплекс-методу виглядає так: необхідно знайти мінімум деякої функції

$$E(x) \to \min_{x \in R^n} \tag{12}$$

досить загального вигляду, при цьому про характер цієї функції не робиться практично жодних апріорних припущень. Робота методу починається з формування початкового комплексу

$$x_i(0) = (x_{i1}(0), x_{i2}(0), ..., x_{ij}(0), ..., x_{in}(0))^T,\ i = 1,2,..., N \geq n+1, \tag{13}$$

що є «хмарою» (популяцією) точок (векторів), досить довільно розташованих в n-вимірному просторі факторів. Серед множини цих точок знаходиться «найгірша» $x_H(0)$, в якій значення функції $E(x_H(0))$ максимальне, після чого ця точка відображується через центр тяжіння решти вершин-точок, формуючи новий комплекс $x_i(1),\ i = 1,2,..., N.$ Таке відображення разом із розтягненням і стисненням забезпечує рух комплексу до екстремуму функції, при цьому, завдяки досить випадковому розподілу точок «хмари», пошук має глобальний характер.

Розглянемо процес оптимізації на t-й ітерації пошуку, коли сформований комплекс, $x_i(t)$, $i = 1,2,..., N.$ Серед множини точок $x_i(t)$ знаходиться «найгірша» така, що

$$E(x_H(t)) = \max_i \{E(x_1(t)), ..., E(x_H(t))\}, \tag{14}$$

після чого визначається центр тяжіння «хмари» без найгіршої точки:

$$x_C(t) = \frac{1}{N-1}\left(\sum_{i=1}^{N} x_i(t) - x_H(t)\right). \tag{15}$$

Далі $x_H(t)$ відображується через центр тяжіння $x_C(t)$, формуючи нову вершину комплексу $x_R(t)$, що теоретично розташована ближче до екстремуму, ніж $x_H(t)$ та $x_C(t)$, тобто

$$E(x_R(t)) < E(x_C(t)) < E(x_H(t)). \tag{16}$$

Операція відображення формально має вигляд:

$$x_R(t) = x_C(t) + \eta_R(x_C(t) - x_H(t)) = X(t)R, \tag{17}$$

де η_R – параметр кроку відображення, що часто приймається рівним одиниці, $X(t) = (x_H(t), x_1(t), ..., x_{N-1}(t))$ – $(n \times N)$-матриця координат вершин комплексу, $R = (-\eta_R, \frac{1+\eta_R}{N-1}, ..., \frac{1+\eta_R}{N-1})^T$ – $(N \times 1)$-вектор.

У випадку, якщо відображена вершина $x_R(t)$ виявиться «найкращою» серед решти точок комплексу, тобто

$$E(x_R(t)) < E(x_i(t)) < E(x_H(t)),\ i = 1,2,..., N-1, \tag{18}$$

виконується операція розтягнення комплексу в напрямку від центра тяжіння $x_C(t)$ до $x_R(t)$ відповідно до виразу

$$x_E(t) = x_C(t) + \eta_E(x_R(t) - x_C(t)) = X(t)E, \tag{19}$$

де η_E – параметр кроку розтягнення, часто приймається рівним двом, $E = (-\eta_E\eta_R, \frac{1-\eta_E(1-\eta_R)}{N-1}, ..., \frac{1-\eta_E(1-\eta_R)}{N-1})^T$. Якщо $x_R(t)$ виявиться найгіршою серед всіх $x_i(t)$, комплекс стискується відповідно до виразу:

$$x_S(t) = x_C(t) + \eta_S(x_R(t) - x_C(t)) = X(t)S, \tag{20}$$

де η_S – параметр кроку стиснення, зазвичай приймається рівним 0,5, $S = (-\eta_S\eta_R, \frac{1-\eta_S(1-\eta_R)}{N-1}, ..., \frac{1-\eta_S(1-\eta_R)}{N-1})^T$.

Отже, в процесі свого руху до екстремуму функції, що оптимізує комплекс, на кожній ітерації втрачається одна найгірша вершина та отримується одна нова точка так, що на $(t+1)$-й ітерації новий комплекс також має N точок-вершин.

На відміну від комплекс-методу в генетичних алгоритмах в результаті селекції з популяції одночасно виключається декілька особин з найгіршими (максимальними) значеннями функції пристосованості. У зв'язку з цим є доцільним ввести модифікацію комплекс-методу з відображенням, розтягненням і стисненням одразу декількох вершин.

Нехай на t-й ітерації процесу оптимізації є комплекс $x_i(t)$, $i = 1,2,..., N$ з $P < N$ найгіршими вершинами $x_{H_p}(t)$, $p = 1,2,..., P$. Тоді координати центра тяжіння комплексу без вершин $x_{H_p}(t)$ задаються виразом

$$x_C(t) = \frac{1}{N-P}(\sum_{i=1}^{N} x_i(t) - \sum_{p=1}^{P} x_{H_p}(t)), \tag{21}$$

а процедура відображення описується системою рівнянь

$$\begin{cases} x_{R_1}(t) = x_C(t) + \eta_R(x_C(t) - x_{H_1}(t)), \\ \vdots \\ x_{R_P}(t) = x_C(t) + \eta_R(x_C(t) - x_{H_P}(t)), \end{cases} \tag{22}$$

або

$$\begin{cases} x_{R_1}(t) = (1+\eta_R)x_C(t) - \eta_R x_{H_1}(t), \\ \vdots \\ x_{R_P}(t) = (1+\eta_R)x_C(t) - \eta_R x_{H_P}(t). \end{cases} \tag{23}$$

У матричній формі ці системи рівнянь можуть бути записані більш компактно $X_R(t) = X(t)R_P$, де $X(t) = (\underbrace{x_{H_1}(t),...,x_{H_P}(t)}_{(n\times P)}, \underbrace{x_1(t),...,x_{N-P}(t)}_{(n\times(N-P))})$ – $(n \times N)$-матриця,

$X_R(t) = (x_{R_1}(t),...,x_{R_P}(t))$ – $(n \times P)$-матриця,

$R_P = \begin{matrix} P\{ \\ N-P\{ \end{matrix} \underbrace{\begin{pmatrix} -\eta_R I_P \\ ------ \\ \dfrac{1+\eta_R}{N-P} I_{N-P,P} \end{pmatrix}}_{P}$ – $(N \times P)$-матриця, I_p – $(P \times P)$-одинична матриця,

$I_{N-P,P}$ – $((N-P)\times P)$-матриця, що утворена одиницями. Якщо серед відображених вершин виявиться $Q \le P$ найкращих, комплекс розтягується в їх напрямі відповідно до рівнянь

$$\begin{cases} x_{E_1}(t) = x_C(t) + \eta_E(x_{R_1}(t) - x_C(t)), \\ \vdots \\ x_{E_Q}(t) = x_C(t) + \eta_E(x_{R_Q}(t) - x_C(t)), \end{cases} \tag{24}$$

або

$$\begin{cases} x_{E_1}(t) = (1-\eta_E)x_C(t) + \eta_E x_{R_1}(t), \\ \vdots \\ x_{E_Q}(t) = (1-\eta_E)x_C(t) + \eta_E x_{R_Q}(t), \end{cases} \tag{25}$$

або

$$X_E(t) = X(t)E_Q, \tag{26}$$

де $X_E(t) = (x_{E_1}(t),...,x_{E_Q}(t))$ – $(N \times Q)$-матриця,

$$E_Q = \begin{matrix} Q\Big\{ \\ N-Q\Big\{ \end{matrix} \underbrace{\begin{pmatrix} -\eta_E\eta_R I_Q \\ ------ \\ (1-\frac{\eta_E(1-\eta_R)}{N-P} I_{N-Q,Q} \end{pmatrix}}_{Q} - (N \times Q)\text{-матриця.}$$

Якщо далі серед відображених вершин виявиться $U \le P$ найгірших, комплекс стискується в їх напрямі відповідно до рівнянь

$$\begin{cases} x_{S_1}(t) = x_c(t) + \eta_S(x_{R_1}(t) - x_C(t)), \\ \vdots \\ x_{S_U}(t) = x_C(t) + \eta_S(x_{R_U}(t) - x_C(t)), \end{cases} \tag{27}$$

або

$$\begin{cases} x_{S_1}(t) = (1-\eta_S)x_C(t) + \eta_S x_{R_1}(t), \\ \vdots \\ x_{S_U}(t) = (1-\eta_S)x_C(t) + \eta_S x_{R_U}(t), \end{cases} \tag{28}$$

або

$$X_S(t) = X(t)S_U, \tag{29}$$

де $X_S(t) = (x_{S_1}(t),...,x_{S_U}(t))$ – $(n \times U)$-матриця,

$$S_U = \begin{matrix} U\Big\{ \\ N-U\Big\{ \end{matrix} \underbrace{\begin{pmatrix} -\eta_S\eta_R I_U \\ ------ \\ (1-\frac{\eta_S(1-\eta_R)}{N-U} I_{N-U,U} \end{pmatrix}}_{Q} - (N \times U)\text{-матриця.}$$

Отже, комплекс-метод набуває риси генетичного алгоритму, у якого в результаті селекції на кожній ітерації з популяції вилучається декілька найгірших особин.

Об’єднуючи введену модифікацію комплекс-методу з холландівською генетичною процедурою, приходимо до алгоритму, що реалізує ідею штучного відбору – з популяції не лише вилучаються найгірші особини, але й одночасно створюються їх «антиподи», що мають покращені властивості.

Робота такого методу утворена послідовністю таких кроків:

- створення початкової популяції, утвореної $P(0)$ особинами хромосомами – вершинами комплексу;
- операція кросоверу зі збільшенням популяції $P_{CR}(0) > P(0)$;
- операція мутації $P_M(0) > P_{CR}(0)$;
- операція інверсії $P_I(0) > P_M(0)$;
- перша селекція (визначення найгірших особин) без скорочення популяції $P_{SEL1}(0) = P_I(0)$;

- операція відображення з вилученням P найгірших особин $P_R(0) < P_{SEL1}(0)$;
- операція розтягнення без збільшення популяції $P_E(0) = P_R(0)$;
- операція стиснення без збільшення популяції $P_I(0) = P_E(0)$;
- друга селекція з вилученням $P_W(0)$ найгірших особин $P_{SEL2}(0) = P_I(0) - P_W(0) = P(1)$ та формування популяції $P(1)$ для наступної ітерації методу.

Порівняно з традиційними генетичними процедурами, запропонований генетичний метод зі штучним відбором на основі комплекс-методу адаптаційної оптимізації, має покращені характеристики, простий в реалізації та призначений для використання в Genetic Mining великих масивів текстових документів.

Четвертий розділ присвячено вирішенню практичної задачі кластеризації результатів пошуку в інформаційно-пошуковій системі наукової бібліотеки, створенню модуля кластеризації результатів роботи цієї системи та дослідженню ефективності відображення та обробки результатів пошуку в ній.

Аналіз отриманих результатів показав, що розроблений у дисертаційній роботі метод кластеризації політематичних текстових документів, що базується на адаптивній нечіткій нейронній мережі, що самоорганізується, з рекурентними ймовірнісним та можливісним методами навчання, а також модуль кластеризації, створений на основі даного методу, забезпечують спрощення обробки результатів пошуку та скорочення часу пошуку релевантної інформації шляхом автоматичного розбиття результатів інформаційного пошуку на категорії.

У **висновках** сформульовано теоретичні та практичні результати роботи.

У **додатках** наведено акти впровадження отриманих теоретичних та прикладних результатів.

ВИСНОВКИ

У дисертаційній роботі наведено результати, які відповідно до поставленої мети є вирішенням науково-технічної задачі розробки інтелектуальних методів кластеризації політематичних текстових документів, що базуються на методах обчислювального інтелекту. Отримані результати мають важливе наукове та практичне значення для створення систем кластеризації політематичних текстових документів, що послідовно надходять до них.

Протягом наукових досліджень отримано такі результати:

1. У результаті аналізу сучасного стану проблеми кластеризації політематичних текстових документів відзначено ряд недоліків основних методів, які знижують ефективність їх застосування та ряд вирішуваних задач. Так більшість методів кластеризації не дозволяють враховувати наявність кластерів, що перетинаються. Також вони обмежені методами обробки вхідних даних (зазвичай дані надходять у пакетному режимі), тобто такі методи не дозволяють послідовно обробляти дані, що обмежує область їх застосування. В цьому випадку доцільною є розробка методів кластеризації політематичних текстових документів, які враховують перелічені вище недоліки.

2. Вперше запропоновано модель адаптивної нечіткої нейронної мережі, що самоорганізується, яка відрізняється від інших нейронних мереж використанням спеціальних нелінійних обчислювачів, які дозволяють знаходити рівні належності вхідних образів документів до кластерів, що дає можливість підвищити якість кластеризації у режимі послідовної обробки даних. Синаптичні ваги мережі визначають координати центроїдів кластерів, що перетинаються, по латеральним зв’язкам нейрони обмінюються координатами,

необхідними для обчислення належностей, а виходом мережі є вектор, який визначає рівень належності вхідного образу до кожного з кластерів.

3. Вперше розроблені рекурентні ймовірнісний та можливісний методи навчання для запропонованої адаптивної нечіткої нейронної мережі, що самоорганізується, які відрізняються від інших методів навчання нейронних мереж наявністю фаззіфікатора. Це дозволило в процесі роботи методів виявляти нові кластери, а також коректно оцінювати викиди та накопичення вибірки в реальному часі, по мірі надходження. Запропоновані методи характеризуються високою швидкодією та незначною обчислювальною складністю.

4. Вперше запропоновано модель системи кластеризації політематичних текстових документів, яка відрізняється від існуючих моделей наявністю двох паралельно працюючих адаптивних нечітких нейронних мереж, що самоорганізуються. В основі навчання цих мереж лежить комбінований метод, що базується на одночасному використанні рекурентних ймовірнісного та можливісного методів самонавчання. Це дозволило в процесі обробки політематичних текстових документів, що послідовно до неї надходять, виявляти нові кластери та відновлювати такі, що перетинаються.

5. Вперше розроблено метод автоматичної кластеризації політематичних текстових документів на основі генетичного алгоритму зі штучним відбором. На відміну від існуючих еволюційних алгоритмів, в даному методі в процедури генетичної оптимізації вводяться оператори комплекс-пошуку, такі як відображення, розтягнення та стиснення. Це дозволило знаходити екстремум довільних функцій великої кількості аргументів в умовах істотної невизначеності про характер цих функцій та використати запропонований метод в Genetic Mining великих масивів текстових документів.

6. Набув подальшого розвитку метод навчання для нейронних мереж, що самоорганізуються, який на відміну від правила навчання «переможець отримує все», містить оцінку рівня належності спостережень кожному з наявних кластерів на основі функції належності. Це дозволило підвищити швидкість обробки інформації, поліпшити якість кластеризації за наявністю кластерів, що перетинаються, шляхом використання нечіткого виведення.

7. Розроблені в дисертаційній роботі методи використано при створенні модулю кластеризації результатів роботи інформаційно-пошукової системи наукової бібліотеки ХНУРЕ та науково-технічної бібліотеки Національного аерокосмічного університету ім. М.Є. Жуковського «Харківський авіаційний інститут», а також на кафедрі штучного інтелекту при підготовці дисциплін «Штучні нейронні мережі: архітектури, навчання, застосування», «Нейромережеві методи обчислювального інтелекту», «Системи обробки природно-мовної інформації», що підтверджено відповідними актами впровадження.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Бодянский Е.В. Самообучающаяся нейро-фаззи система для адаптивной кластеризации текстовых документов / Е.В. Бодянский, В.В. Волкова, Б.В. Колчигин // Бионика интеллекта. – 2008. – Вып. 1(70). – С.34–38.

2. Бодянский Е.В. Кластеризация массивов текстовых документов на основе адаптивной нечеткой самоорганизующейся нейронной сети / Е.В. Бодянский, В.В. Волкова, А.С. Егоров // Радиоэлектроника. Информатика. Управление. – 2009. – Вып. 1(20). – С.113–117.

3. Бодянский Е.В. Автоматическая кластеризация текстовых документов на основе генетического алгоритма с искусственным отбором / Е.В. Бодянский, В.В. Волкова, К.В. Коваль // Радиоэлектроника. Информатика. Управление. – 2009. – Вып. 2(21). – С.91–96.

4. Бодянский Е.В. Нейронная сеть Т.Кохонена с нечетким выводом и алгоритм ее самообучения / Е.В. Бодянский, В.В. Волкова, Е.В. Махиборода // Сборник трудов Харьковского университета Воздушных Сил. – 2009. – Вып. 2(20). – С.74–77.

5. Волкова В.В. Исследование основных подходов к построению онтологий предметных областей / В.В. Волкова // Радиоэлектроника и молодежь в 21 веке: 9-й Международный молодежный форум: материалы форума. – Харьков: ХНУРЭ. – 2005. – С. 341.

6. Рябова Н.В. Методы и модели интеллектуальной обработки текстов в задачах онтологического инжиниринга / Н.В. Рябова, В.В. Волкова, Я.В. Дыдыкина // Международный радиоэлектронный форум «Прикладная радиоэлектроника. Состояние и перспективы развития»: тезисы докл. – Х., 2005. – С. 87–90.

7. Волкова В.В. Применение нейронных сетей в задачах онтологического инжиниринга / В.В. Волкова // Международная конференция «Современные информационные системы. Проблемы и тенденции развития»: тезисы докл. – Х., 2007. – С. 351.

8. Бодянский Е.В. Нечеткая возможностная кластеризация текстовых документов на основе самоорганизующейся карты Кохонена / Е.В. Бодянский, В.В. Волкова, Б.В. Колчигин // 6-я Международная научно-практическая конференция «Математическое и программное обеспечение интеллектуальных систем» (MPZIS – 2008): тезисы докл. – Д., 2008. – С.47–48.

9. Волкова В.В. Возможностная фаззи-кластеризация текстовых массивов в реальном времени на основе самообучающейся нейронной сети / В.В. Волкова, Б.В. Колчигин // Факультетская научно-практическая молодежная школа-семинар студентов, аспирантов и молодых ученых «Информационные интеллектуальные системы»: тезисы докл. – Харьков: ХНУРЭ., 2008. – С.22–25.

АНОТАЦІЯ

Волкова В.В. Методи нечіткої кластеризації політематичних текстових документів. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту. – Харківський національний університет радіоелектроніки, Харків, 2010.

Дисертацію присвячено розробці методів кластеризації політематичних текстових документів у режимі послідовної обробки даних та наявності кластерів, що перетинаються. Розглянуто задачу кластеризації політематичних текстових документів, основні методи обробки документів та існуючі методи їх кластеризації, визначено основні недоліки та переваги розглянутих методів. Вперше запропоновано адаптивну нечітку нейронну мережу, що самоорганізується, та ймовірнісний і можливісний методи її навчання, які дозволяють виконувати нечітку кластеризацію політематичних текстових документів, що в послідовному режимі надходять на вхід мережі, а також у процесі навчання знаходити нові кластери. Запропоновані методи навчання відрізняються своєю швидкодією та незначною обчислювальною складністю. Вперше запропоновано модель нейро-нечіткої системи кластеризації політематичних текстових документів з нечітким виведенням на основі комбінованого методу навчання. Набув подальшого розвитку метод навчання для нейронних мереж, що самоорганізуються, який дозволяє підвищити швидкість обробки інформації, поліпшити якість кластеризації за наявності кластерів, що перетинаються, шляхом використання нечіткого виведення.

Вперше запропоновано метод автоматичної кластеризації політематичних текстових документів на основі генетичного алгоритму зі штучним відбором, який базується на комплекс-методі адаптаційної оптимізації та дозволяє знаходити екстремум довільних функцій великої кількості аргументів в умовах істотної невизначеності про характер цих функцій.

Ключові слова: політематичні текстові документи, нечітка кластеризація, штучні нейронні мережі, методи навчання, генетичні алгоритми, нечітке виведення.

АННОТАЦИЯ

Волкова В.В. Методы нечеткой кластеризации политематических текстовых документов. – Рукопись.

Диссертация на соискание научной степени кандидата технических наук по специальности 05.13.23 – системы и средства искусственного интеллекта. – Харьковский национальный университет радиоэлектроники, Харьков, 2010.

Диссертационная работа посвящена разработке интеллектуальных методов нечеткой кластеризации политематических текстовых документов в режиме последовательной обработки данных.

В первом разделе рассматривается понятие «политематический текстовый документ» и основные проблемы, которые возникают при обработке такого типа документов. Рассмотрены методы обработки документов и существующие методы кластеризации текстовых документов, в том числе такой, которая выполняется в последовательном режиме, определены их основные недостатки и преимущества. Обоснована целесообразность разработки методов кластеризации политематических текстовых документов на основе технологий вычислительного интеллекта, которые позволяют выполнять кластеризацию в последовательном режиме, учитывать наличие пересекающихся кластеров, а также обнаруживать новые.

Во втором разделе впервые предложена адаптивная нечеткая самоорганизующаяся нейронная сеть, нейроны которой отличаются от традиционных линейных ассоциаторов, образующих самоорганизующуюся карту Т. Кохонена. Синаптические веса определяют координаты центроидов перекрывающихся кластеров, по латеральным связям нейроны обмениваются координатами, необходимыми для вычисления принадлежностей, а выходом нейронной сети является вектор, определяющий уровень принадлежности входного образа к каждому из кластеров. Предложенная архитектура сети предназначена для решения задачи последовательной кластеризации политематических текстовых документов с учетом наличия пересекающихся кластеров.

Адаптивная нечеткая самоорганизующаяся нейронная сеть настраивается с помощью предложенных во втором разделе рекуррентных вероятностного и возможностного методов самообучения. Данные методы позволяют выполнять нечеткую кластеризацию политематических текстовых документов, находить в режиме реального времени прототипы формируемых кластеров, а также оценивать уровни принадлежности каждого образа документа, поступающего на вход нейронной сети, к конкретному кластеру. Также они позволяют обнаруживать в процессе обучения новые кластеры. Предложенные методы отличаются высоким быстродействием и незначительной вычислительной сложностью.

Впервые предложена модель нейро-нечеткой системы кластеризации политематических текстовых документов с нечетким выводом, в основе обучения которой лежит комбинированный метод, базирующийся на одновременном использовании вероятностного и возможностного рекуррентных методов самообучения. Данная система учитывает нечеткие кластеры при кластеризации последовательно поступающих политематических текстовых документов.

Получил дальнейшее развитие метод обучения самоорганизующихся нейронных сетей, который позволяет повысить скорость обработки информации, улучшить качество кластеризации при наличии пересекающихся кластеров путем использования нечеткого вывода.

В третьем разделе диссертационной работы предложен метод автоматической кластеризации политематических текстовых документов на основе генетического алгоритма с искусственным отбором. Данный метод базируется на комплекс-методе адаптационной оптимизации и позволяет находить экстремум произвольных функций большого числа аргументов в условиях существенной неопределенности о характере этих функций. Метод имеет улучшенные характеристики по сравнению с традиционными генетическими процедурами, прост в реализации и предназначен для использования в Genetic Mining больших массивов текстовых документов.

Проведено имитационное моделирование разработанной модели нейро-нечеткой системы на основе предложенной адаптивной нечеткой самоорганизующейся нейронной сети, генетического алгоритма с искусственным отбором. Показаны их преимущества перед известными архитектурами и методами обучения как по точности, методу обработки, так и по быстродействию в задачах последовательной нечеткой кластеризации политематических текстовых документов. Решена практическая задача нечеткой кластеризации результатов поиска информационно-поисковой системы научной библиотеки. Результаты показали, что разработанные методы существенно улучшают процесс обработки пользователем результатов работы информационно-поисковой системы, а также сокращают время обработки результатов запросов.

Ключевые слова: политематические текстовые документы, нечеткая кластеризация, искусственные нейронные сети, методы обучения, генетические алгоритмы, нечеткий вывод.

ABSTRACT

Volkova V.V. The fuzzy clustering methods for multi-topic text documents. – Manuscript.

The thesis for the candidate's degree in technical sciences, specialty 05.13.23 – systems and tools of artificial intelligence. – Kharkiv National University of Radio Electronics, Kharkiv, 2010.

The thesis is devoted to developing of multi-topic texts clustering methods in the real time using the adaptive fuzzy self-organizing neural network and genetic algorithm with the artificial selection. I consider the existent methods of text documents processing and their clusterization. Basic advantages and disadvantages have been revealed. For the first time, the adaptive fuzzy self-organizing neural network has been developed. The probabilistic and possibilistic methods of the self-organization for this neural network are first proposed. These methods allow to execute the fuzzy clusterization of multi-topic text documents entering on the entrance of the network in the real time. As a result, methods find new clusters. The proposed methods of learning differ from other ones by the fast operation and low requirements to computational recourses. The model of the neuro-fuzzy system of clusterization of multi-topic text documents is first developed with an fuzzy inference applying the combined method of learning. This is based on simultaneous usage of the probabilistic and possibilistic methods of self-organization, and takes into account fuzzy clusters during the clusterization process of the multi-topic text documents. The method of learning of self-organizing maps have got further development allowing to increase the speed of the information processing, improve quality of the clusterization in the presence of intersecting clusters using of fuzzy inference.

For the first time, the method of the automatic clusterization of the multi-topic text documents have been developed using the genetic algorithm with the artificial selection. The method is simple in realization and intended for the applications in Genetic Mining of large collections of text documents in the mode of the sequential processing.

Keywords: multi-topic text documents, fuzzy clusterization, artificial neural networks, learning procedures, genetic algorithms, fuzzy inference.