

УДК 681.3 (045)

О. І. Чумаченко, канд. техн. наук, доц.

РОЗРОБЛЕННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ СИСТЕМИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ВЕБ-ДАНИХ

Факультет інформатики та обчислювальної техніки НТУУ «КПІ»

Розглянуто програмне забезпечення інтелектуального аналізу веб-даних з використанням нейронної мережі ART2. Запропоновано розв'язання задачі кластеризації.

Ключові слова: інтелектуальний аналіз веб-даних, кластеризація, програмне забезпечення.

Вступ. Головна особливість аналізу даних (Data Mining) – це поєднання широкого математичного інструментарію (від класичного статистичного аналізу до нових кібернетичних методів) і останніх досягнень у сфері інформаційних технологій. У технології Data Mining гармонійно поєдналися строго формалізовані методи та методи неформального аналізу, тобто кількісний і якісний аналіз даних.

До методів і алгоритмів аналізу даних (Data Mining) належать такі [1]: штучні нейронні мережі, дерева рішень, символічні правила, методи найближчого сусіда і k -найближчого сусіда, метод опорних векторів, байєсові мережі, лінійна регресія, кореляційно-регресійний аналіз; ієрархічні методи кластерного аналізу, неієрархічні методи кластерного аналізу, у тому числі алгоритми k -середніх і k -медіани; методи пошуку асоціативних правил, зокрема алгоритм Apriori; метод обмеженого перебирання, еволюційне програмування і генетичні алгоритми, різні методи візуалізації даних і безліч інших методів.

У роботі розглядається завдання кластеризації веб-даних. Результатом кластеризації є розбиття об'єктів на групи.

Технологія веб-даних охоплює методи, які здатні на підставі даних сайту знайти нові, раніше невідомі знання, які надалі можна буде використовувати на практиці. Інакше кажучи, технологія веб-даних застосовує технологію інтелектуального аналізу даних для аналізу неструктурованої, неоднорідної, розподіленої і значної за обсягом інформації, що міститься на веб-вузлах.

Існує дві топології нейронних мереж, найбільш придатних для задач кластеризації користувачів сайту, зокрема карта ознак самоорганізації Кохонена (KOCK), та нейронна мережа на основі теорії адаптивного резонансу (ART), оскільки, вони належать до некерованих штучних нейронних мереж (ШНМ) (навчання без учителя).

Архітектура KOCK має один істотний недолік, а саме: досить складно визначити кількість кластерів за картою ознак, окрім випадків, коли топологічні карти чітко розрізняються.

Архітектура нейронних мереж на основі ART [2] не потребує наперед заданої кількості кластерів, а також має додаткові переваги:

- точне і автоматичне визначення кількості кластерів без візуальної оцінки;
- швидке навчання і пристосування до нестабільного середовища;
- стабільність і пластичність;
- навчання без учителя.

Привабливою особливістю нейронних мереж з адаптивним резонансом є те, що вони зберігають пластичність при запам'ятовуванні нових образів і водночас запобігають модифікації старої пам'яті.

З огляду на ці переваги нейронних мереж типу ART у розроблюваному програмному забезпеченні використовуватиметься саме така ШНМ.

Структура програмного забезпечення. Умовно розроблене програмне забезпечення можна поділити на три частини (рис. 1):

- 1) попереднє оброблення даних, яке передбачає фільтрацію та підготовку даних активності користувачів сайту електронної торгівлі для використання нейронною мережею типу ART2;
- 2) реалізація (структурно-параметричний синтез) нейронної мережі для виконання кластеризації запитів користувачів за товарами сайту інтернет-магазину;
- 3) графічне відображення результатів роботи нейронної мережі для користувача.



Рис. 1. Загальна структурна схема програмного забезпечення

Попереднє оброблення даних. У розробленому програмному забезпеченні попереднє оброблення даних полягає у виконанні таких необхідних дій:

- перевірка даних веб-логу на коректність;
- очищення даних, отриманих з веб-логу сервера від надлишкової та непотрібної інформації;
- підготовка вхідних даних для нейронної мережі типу ART2.

Реалізація структурно-параметричного синтезу нейронної мережі. Програмна реалізація синтезу нейронної мережі здійснюється за допомогою додаткового програмного модуля на основі фреймворку ANN for PHP5 [3]. Значення параметрів нейронної мережі налаштовуються згідно з вимогами оригінальної архітектури ART2, описаної у працях [4] і [5].

Нейронна мережа, що була спроектована у цій роботі, має таку структуру:

- вхідний шар нейронів;
- один прихований шар нейронів;
- вихідний шар нейронів.

Опис вхідних даних. На вхід нейронної мережі подаються URL адреси, відвіданих користувачами сторінок (отримуються після завершення попереднього етапу оброблення даних веб-логу). Адреси обробляються у їх звичайному текстовому форматі (такому, у якому подані користувачу сайту).

Опис вихідних даних. Після завершення роботи нейромережі отримується набір кластерів розподілу запитів користувачів за товарами електронного магазину. Тип виходів нейронної мережі бінарний, що пов'язано з використанням фреймворку ANN for PHP5.

Прихований шар являє собою набір вагових коефіцієнтів, які обчислюються згідно з працею [6]. Вагові коефіцієнти коригуються на підставі методу зворотного поширення помилки, який є у згаданому фреймворку.

Алгоритм роботи програми. Алгоритм роботи програмного забезпечення починається із проходження етапу збирання даних на веб-сервер. Далі виконується перевірка на наявність та коректність зібраних даних.

Вхідний вектор подається до нейромережі. Для того щоб мати змогу користуватися уже наявними бібліотеками класів нейромереж, до системи підключається фреймворк ANN for PHP5 і створюється екземпляр нейронної мережі. Після цього настає важливий етап – навчання нейронної мережі [7]. Результати навчання зберігаються та використовуються при роботі нейронної мережі.

Після етапу навчання виконується ще одна перевірка вхідних даних на можливість їх використання нейронною мережею цього типу (у разі помилки виводиться повідомлення «Network cannot be loaded»).

Для всіх відвідувань користувачів у нейронній мережі запитується, до якого розділу належить кожна сторінка і через процедуру get Outputs отримується відповідь.

Далі завантажується бібліотека для формування діаграми результатів на основі опрацьованих даних та для користувача розробленого програмного забезпечення.

Загальну блок-схему алгоритму роботи програмного забезпечення показано на рис. 2.

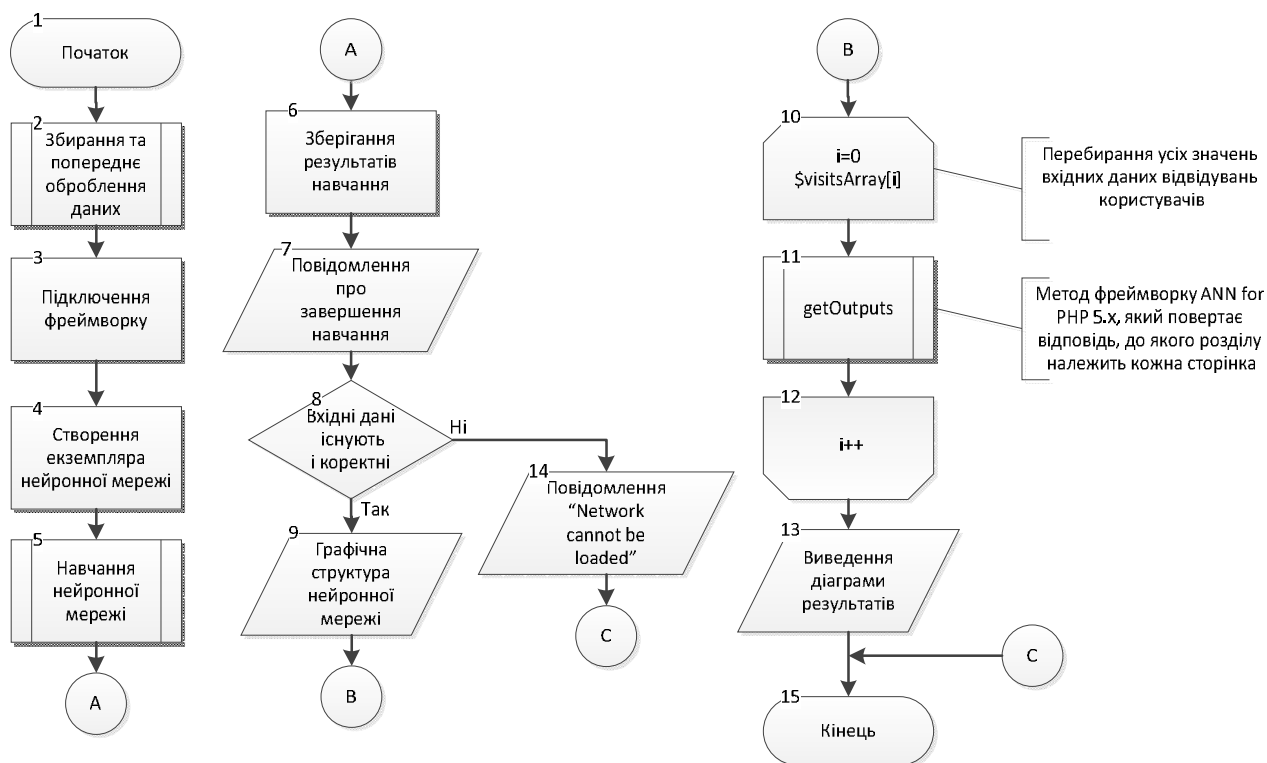


Рис. 2. Блок-схема алгоритму роботи програмного забезпечення

Інтерфейс користувача. Інтерфейс користувача програмного забезпечення складається з двох частин:

- табличного подання результатів;
- подання результатів у вигляді діаграми.

Табличне подання являє собою перелік відвіданих сторінок користувачами сайту, кількість відвідувань кожної сторінки та найменування позицій, що їм відповідають.

Подання результатів у вигляді діаграми є більш наочним і містить відсоткове співвідношення кількості переглядів по кожній із позицій і тому є зручнішим для сприйняття відвідувачами сайту.

Вимоги до обчислювальної системи. Розроблений програмний продукт написаний веб-орієнтованою мовою програмування PHP версії 5.3.10 і працює на платформах Windows,

UNIX, Linux, Free BSD, Mac OS. Для його роботи необхідно, щоб на комп'ютері, який виконуватиме роль сервера збирання та оброблення даних (персональний комп'ютер або ж виділений сервер хостинг-провайдера), було встановлено таке програмне забезпечення:

- HTTP-сервер (наприклад, Nginx);
- PHP 5.x (рекомендується версія 5.3.10 або вище);
- необхідні установлення PHP (php.ini):
 - file_uploads = On;
 - magic_quotes_gpc = On;
- система керування базами даних MySQL.

Наведені програмні продукти належать до вільно поширюваного програмного забезпечення, а тому не потребують ліцензій на їх використання.

Контрольний приклад. З метою перевірки якості кластеризації записів веб-логу нейронною мережею ART2 проведено експеримент.

Використано файли веб-логу за п'ятиденний період. Джерелом файлів веб-логу був сайт <http://top-trade.lviv.ua>, налаштований вести цілодобовий журнал активності за допомогою написаного скрипту ga.php. Виконання скрипту, крім збирання даних, їх обробляє та готує вхідний вектор для нейронної мережі ART2. Отже, формування навчальної вибірки для нейронної мережі збігається із завершенням етапу попереднього оброблення даних.

Після цього проведено навчання нейронної мережі з критерієм подібності $\rho = 0,7$.

Результати навчання показано на рис. 3.

Detected output type	Binary
Backpropagation algorithm	Back propagation
Activation function	Sigmoid
Momentum	0.95
Learning rate	0.7
Network error	0
Output error tolerance	+/- 0.02
Total loops	3,135 loops
Total trainings	2,893 trainings
Total activations	3,360 activations
Total activation requests	6,253 activation requests
Epoch	5 loops
Training time	11 seconds = 0.2 minutes
Loops / second	285 loops / second
Training finished	Yes

Рис. 3. Результати навчання нейромережі

Як видно з рис. 3, у результаті навчання нейронної мережі пройдено 3135 циклів та 2893 тренувань. Загальний час навчання склав 11 с (285 циклів/с).

Сформовано 60 входів та відкориговано вагові коефіцієнти за методом зворотного поширення помилки.

На виході отримано 5 кластерів бінарного типу, які виражають розподіл відвіданих користувачами сторінок за позиціями, до яких ці сторінки належать (рис. 4).

[illegible]

Software Design of Web Data Analysis Intelligent Systems
The intellectual analysis software of web-datas on basis of use neuro networks ART2 is considered.
The solution of clasterization problem is proposed.