УДК 004.7(045)

Кременецький Г.М., канд. техн. наук

АСОЦІАТИВНЕ ПОДАННЯ РЕЧЕНЬ У ГРАФІЧНОМУ ВИГЛЯДІ

Інститут комп'ютерних технологій
Національного авіаційного університету

Запропоновано використання асоціацій для представлення слів та схеми класифікації отриманих асоціацій. Розглянуто метод перетворення речення у інтегровану асоціацію

Вступ

Проблема класифікації текстів є базовою для багатьох галузей використання.

Наприклад, фільтрування *SPAM*у в електронній пошті, автоматичне створення каталогів у бібліотеці, автоматичний переклад тощо [1 – 3]. Крупні корпорації розробляють інтерактивні інтерефейси взаємодії користувача з різними системами.

Дуже швидко розвивається робототехніка. Створюються нові системи які потребують нового рівня сприйняття вхідної інформації. Ці системи повинні не тільки перетворювати вхідну інформацію у цифровий еквівалент, а і певною мірою проводити попередню обробку яка забезпечить спрощення аналізу та прийняття рішень.

Постановка задачі

Загальний алгоритм класифікації можна викласти наступним чином:

$$Y = f(X), \qquad (1)$$

де Y – код класу, X – вектор вхідних параметрів, f – функція класифікації.

Фільтрування *SPAM*у потребує швидкого алгоритму, який дозволить поділити вхідний текст на 2 класи: *SPAM* або не *SPAM*. Для цього достатньо використовувати простий статистичний класифікатор, наприклад, Байєсовський. Це швидкий та достатньо надійний метод сортування електронної пошти.

$$Y = f(C,P), \qquad (2)$$

де Y – код класу, C – вектор кодів слів, P – вектор вірогідності появи слів у рекламному листі, f – функція класифікації на основі C та P.

Створення каталогів текстів у бібліотеці потребує більш високої якості розподілу на декілька різних класів. Це потребує інформативної вхідної інформації та складних алгоритмів класифікації.

Тому у цьому випадку використовують попередній морфологічний аналіз речень, що дає додаткову інформацію для якісного аналізу:

$$Y = f(C,M,P), \qquad (3)$$

де Y – код класу, C – вектор кодів слів, M – вектор кодів морфологічних конструк- цій, P – вектор вірогідності появи слова з кодом x, морфологічним кодом m у рекламному листі, f – функція класифікації на основі C та P. У цьому випадку ми маємо деяку кількість додаткової інформації для класифікації тексту завдяки морфологічному аналізу:

$$X_{C,M,P} - X_{C,P} = \Delta x, \qquad (4)$$

де Δx – кількість додаткової інформації, $X_{C,P}$, – кількість інформації у випадку фільтрації спаму, $X_{C,M,P}$, , – кількість інформації для класифікації у бібліотеці.

Як можно побачити з (4) нам потрібна додаткова вхідна інформація для створення більш потужних класифікаторів.

У випадку людського мозку, проблему підвищення кількості вхідної інформації для аналізу текстової інформації вирішила природа. Це рішення дуже просте. Ми сприймаємо оточення візуально. Тобто, кожне слово у нашому мозку асоціюється із зображенням. Але це не фотографічне відображення об'єкту, а щось дуже схематичне. У психології ця картинка має назву «асоціація».

Наприклад, слово «чоловік» викликає у нашої уяві щось схоже на:

Рис. 1. Асоціація – «чоловік»

Рис. 2. Асоціація –«жінка»

Якщо порівняти ці два зображення можна легко зазначити, що вони дуже схожі. Новий клас – «людина», це не дуже складна задача не тільки для розуму людини, але і для не складної нейтронної мережі, яка використовує алгоритми навчання без вчителя.

Тепер пропоную порівняти два слова «чоловік» та «жінка» у вигляді кодів. Наприклад, «чоловік» – 1, а «жінка» – 2. Число 1 та число 2 ні в якому разі не можуть викликати новий клас – «людина». Спробуємо додати інформацію з морфологічного аналізу. Ці два слова є іменниками. Наприклад, іменник має код 1. У цьому разі ми маємо два вектори (1, 1) – «чоловік», (2, 1) – «жінка». Але і у цьому випадку ми ні у якому разі не зможемо уявити клас – «людина» Тобто, зображення супроводжує додаткова інформація:

$$\text{Image} - X_{C,M,P} = \Delta x_1.$$

Ця додаткова інформація стане у нагоді для підвищення якості класифікації текстової інформації.

Усі попередні роздуми можна викласти у вигляді наступної схеми:

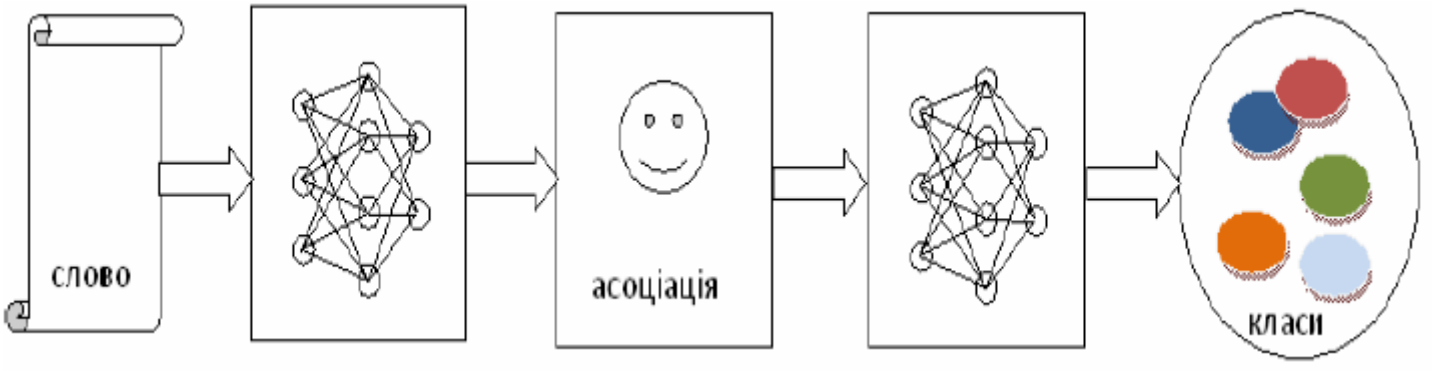

Рис. 3. Схема перетворення слова у клас

Текстова інформація перетворюється у асоціації (графічні зображення). Потім графічні зображення класифікуються. Обидва етапи вже ефективно використовуються у техніці. Перший у асоціативній пам'яті, другий у розпізнаванні зображень.

Але речення складаються у більшості випадках не лише з одного слова. Як створювати асоціацію, що включає до себе іменник, дієслово, прикметник та інші частини мови без яких не може бути сприйняття речення.

Пропоную послідовно уявити слова у реченні: «Червона людина іде». Слово «червона» викликає в уяві червоний колір. Це як червоний аркуш на якому ще нічого немає окрім червоного коліру

Рис. 4. Асоціація – «іде»

Рис. 5. Асоціація – «червона людина іде»

Далі уявимо слово «людина». Але людина в нашій уяві буде не як у попередньому випадку чорна рис. 1. Це буде вже людина червоного коліру. Тепер уявимо дієслово «іде» рис. 4.

Рис. 6. Алгоритм створення інтегрованої асоціації

Якщо уявити червону людину яка іде, то виникає асоціація рис. 4.

Асоціація на рис. 4 є інтернованою з трьох «червона» + «людина» + «іде».

Створення інтернованої асоціації буде мати наступний вигляд рис. 6 Розглянемо алгоритм:

1. Речення розкладається на окремі слова.

2. Кожне з цих слів перетворюється на асоціацію з використанням попередньо навченої нейронної мережі.

3. Наступним кроком створена асоціація передається на вхід іншої нейронної мережі яка інтегрує її з попередньою асоціацією, що сберігється у спеціальному модулі зберігання. Якщо речення ще не повністю оброблено отримана інтегрована асоціація передається до модулю зберігання.

4. У іншому випадку отримана інтегрована асоціація передається до виходу системи, а модуль зберігання очищується.

Функціональна схема створення інтегрованої асоціації рис. 7 складається за наступних компонентів:

1) модуль який виділяє з речення окремі слова – модуль розбору речення;

2) нейрона мережа перетворює кожне слово у графічну асоціацію;

3) модуль пам’яті зберігає попередню інтегровану асоціацію;

4) інтегратор створює інтегровану асоціацію базуючись на попередніх результатах, що надходять з модулів 2 та 3.

Для компоненту 2 можно використовувати не тільки нейрону мережу а і довідник асоціацій. Тобто базу даних яка буде складатися зі слів та їх асоціацій.

Але з точки зору подальшої реалізації у вигляді компонентів краще використовувати однотипну базу.

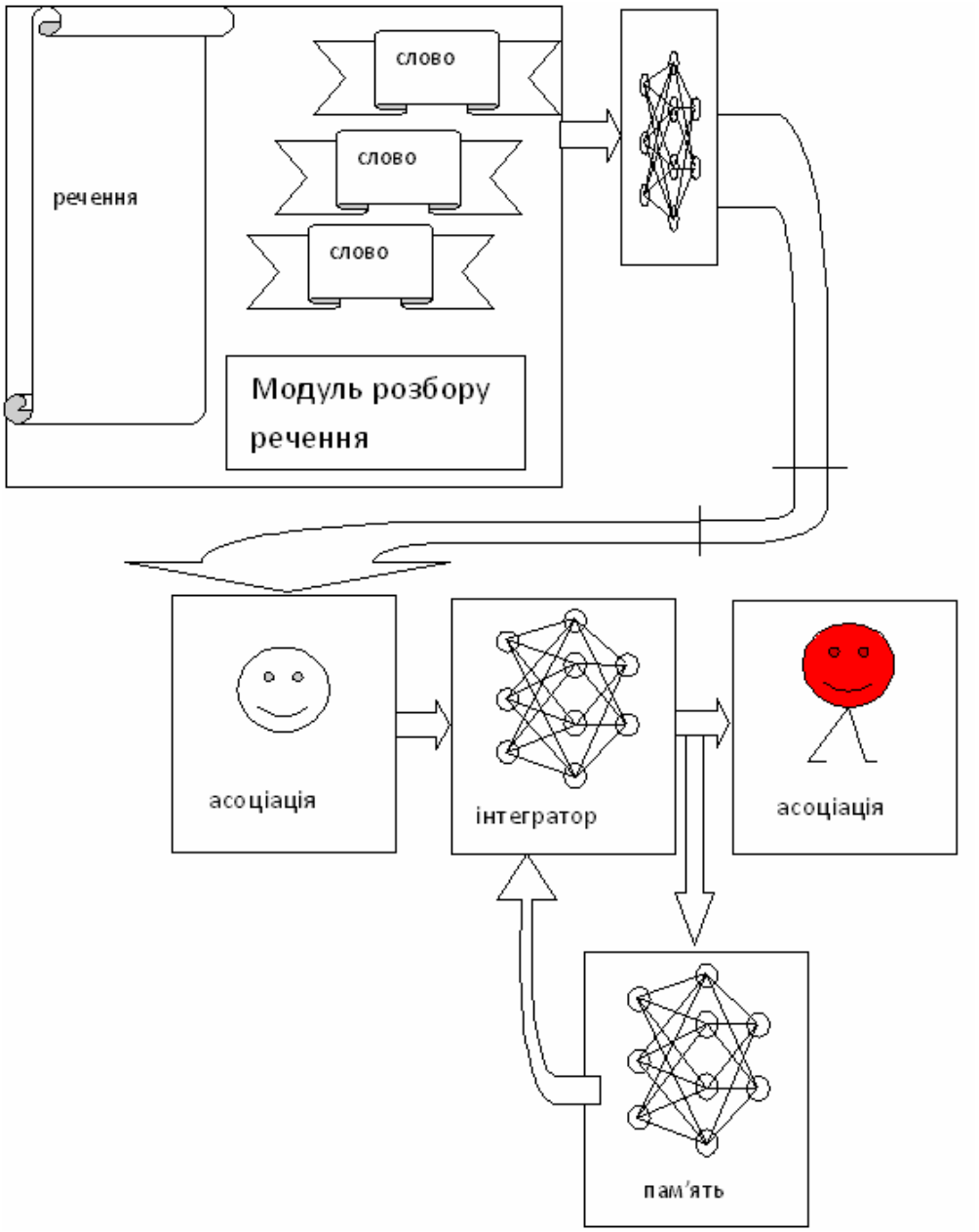

Рис. 7. Схема перетворення речення у інтегровану асоціацію

Висновки

Використання асоціацій у класифікації текстової інформації це наступний крок у розумінні принципів побудуви кластеризованих нейронних мереж у природі. Вершиною яких можна вважати людський розум. Всі компоненти які було використано при створенні функціональної схеми сьогодні існують. Залишилося їх тільки об'єднати та підібрати потрібні алгоритми їх функціонування. Наприклад, передатні функції нейронів у мережах та алгоритми навчання. Також потрібно створити базу асоціацій.

Є ще одна проблема яку я спробую вирішити у наступних статтях – це сприйняття тексту у цілому, а не тільки окремих речень.

Список літератури

1. *Савельев М.С.* Внутренний СПАМ // Источник: PCWeek/RE, №32, 2003. – С. 13 – 20.

2. *Саймон Хайкин* Нейронные сети: полный курс. – М.: Издательский дом «Вильямс». – 2006. – 1105 с.

3. *Чугреев В.Л., Яковлев С.А.* Анализ структуры текста и прогнозирование нечисловых величин // ВУЗОВСКАЯ НАУКА – РЕГИОНУ: Материалы 1-й Общероссийской нучн.-техн. конф. – Вологда: ВоГТУ, 2003. – С. 202–204.