УДК 004.023

В. Г. Кисіль
В. С. Озеранський

РОЗРОБКА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АНАЛІЗУ ВИКОРИСТАННЯ СОЦІАЛЬНИХ МЕРЕЖ СТУДЕНТАМИ

Вінницький національний технічний університет

Анотація

Розглянуто актуальність задачі аналізу використання соціальних мереж. Розглянуто існуючі методи розв'язання поставленої задачі, та запропоновано використання методу ієрархічної кластеризації для її вирішення.

Ключові слова: соціальні мережі, аналіз використання соціальних мереж, парсинг, таргетинг.

Abstract

The relevance of the problem of the analysis of the use of social networks is considered. The existing methods of solving the set task are considered, and the use of hierarchical clustering method for its solution is proposed.

Keywords: social networks, analysis of social networks usage, parsing, targeting.

Вступ

Соціальні мережі щодня все більше і більше охоплюють сфери нашого існування. Десятки тисяч багатонаправлених соціальних мереж дають можливість своїм користувачам стежити за світовими новинами, обмінюватися фотографіями, відео та музикою, спілкуватися. Гіганти, такі як Facebook, Instagram, Twitter, VKontakte, щоденно збирають сотні мільйонів користувачів на своїх платформах. І загальна аудиторія соціальних мереж оцінюється мільярдами користувачів. Соціальні мережі стали невід'ємною частиною життя користувачів - в середньому один користувач переглядає близько ста сторінок на день. [1].

Сьогодні аудиторію соціальних мереж можна порівняти з аудиторією телевізійних каналів. Лише 18% звичайних телевізійних рекламних кампаній приносять позитивну віддачу від інвестицій. Тим часом, навіть найпримітивніша рекламна кампанія соціальних мереж може примножити інвестиції. Аудиторія соціальних мереж є більш активною та уважною, ніж телевізійна аудиторія. Це пов'язано з тим, що в соціальних мережах взаємодія здійснюється безпосередньо з кожним користувачем, що враховує його індивідуальні характеристики, інтереси та потреби [2].

Метою роботи є підвищення точності підбору цільової аудиторії в соціальних мережах за рахунок використання методу ієрархічної кластеризації.

Огляд існуючих продуктів

На даний момент, в мережі інтернету вже є не мало програмних аналогів до програмного продукту який був розроблений мною, але кожний з цих сервісів або просто робить збір та парсинг аудиторії (при цьому з певними обмеженнями), чи просто надає отриману «суху» статистику не проводивши жодного аналізу щодо використання соціальних мереж. До того ж, всі сервіси які надають користувачам хоча б якусь інформацію щодо використання соціальних мереж певними аудиторіями – досить недешеві .

Як я вже наголошував, розроблений мною програмний модуль дозволить не лише збирати ID користувачів (в тому числі і студентів), і надавати «суху» статистику використання соціальної мережі та вподобань аудиторії, але й отримувати проаналізовану інформацію використовуючи яку можна буде

створювати більш точні рекламні кампанії та економити значні суми коштів для агенств та підприємств які рекламують свої послуги та товари в соціальних мережах.

Розробка моделі кластеризації даних

В даний час відомо безліч алгоритмів нечіткої кластеризації, таких як Fuzzy C-Means, FOPTICS і ін. Дані алгоритми формують кластери, межі яких розмиті, а об'єкт може належати більш ніж одному кластеру з різними ступенями приналежності. Однак слід зазначити, що більшість алгоритмів нечіткої кластеризації працюють з чіткими значеннями параметрів об'єктів, формуючи кластери, наприклад, на основі оцінки відстаней між об'єктами і центром кластера. Такий підхід не дозволяє ефективно здійснювати кластеризацію об'єктів з нечітко заданими значеннями параметрів. У зв'язку з цим актуальним завданням є розробка методів кластерного аналізу, здатних враховувати нечітку природу об'єктів, тобто працювати з параметрами, заданими в нечіткій формі у вигляді функцій приналежності.

Для вирішення багатьох практичних завдань в даний час використовується концептуальна кластеризація даних, яскравим представником якої є метод COBWEB.

В дипломному проекті використаємо математичну модель COBWEB.

Класичний варіант реалізації методу COBWEB не припускав роботу з параметрами, заданими в нечіткій формі, що актуалізує рішення поставленої вище завдання для даного методу.

Метод COBWEB будує дерево класифікації з ймовірними описами концептів. Вибір можливого способу кластеризації об'єктів заснований на значеннях функції корисності кластеризації. При побудові дерева класифікації використовуються наступні 4 операції:

– віднесення об'єкта до найкращого з існуючих кластерів;
– додавання нового кластера, що містить єдиний об'єкт;
– злиття двох існуючих кластерів в один новий з додаванням в неї цього об'єкта;
– розбиття існуючого кластера на два і віднесення об'єкта на краще з новостворених кластерів.

Покроковий опис методу концептуальної кластеризації в дипломному проекті:

1. Вводиться кореневої кластер C0, властивості якого збігаються з властивостями першого об'єкта O1 = [V11, ..., V1m]. Для кожного наступного об'єкта Oi = [Vi1, ..., Vim] виконується цикл, який реалізує кроки 2-6, в рамках яких виконуються 4 вище представлені операції.

2. Об'єкт Oi додається по черзі в кластери C1, C2, ..., Ck. Після кожного додавання обчислюється корисність кластеризації CU1, ..., CUk.

3. Для об'єкта Oi створюється новий кластер Ck + 1, об'єкт поміщається в кластер і обчислюється корисність кластеризації CUk + 1.

4. Об'єднуються два кластери з максимальними значенням корисності кластеризації з CU1, ..., CUk. Утворюється новий кластер, в нього додається об'єкт Oi. Обчислюється корисність кластеризації CUk + 2.

5. Об'єкт Oi додається в кластер з максимальним значенням корисності кластеризації з CU1, ..., CUk. Утворюється новий кластер з двома кластерами-нащадками. Обчислюється корисність кластеризації CUk + 3.

6. Вибирається максимальне значення корисності кластеризації серед корисностей CU1, ..., CUk, CUk + 1, CUk + 2, CUk + 3, відповідно до нього вибирається операція розбиття об'єктів по кластерам. [3].

Розроблений метод концептуальної кластеризації, на відміну від існуючих методів кластеризації, дозволяє працювати з об'єктами, що характеризуються нечіткими параметрами, і будувати модель концептуальної кластеризації для об'єктів нечіткої природи. Основу методу складає запропонована в роботі модифікована формула оцінки корисності концептуальної кластеризації для об'єктів, що характеризуються нечіткими параметрами. Експериментальним шляхом показано, що використання кусочно-лінійних функцій приналежності для завдання значень нечітких параметрів об'єктів дозволяє збільшити розділяє здатність кластерів в розробленому методі в порівнянні з П-образними функціями. Отримані в роботі теоретичні результати були використані для вирішення завдання автоматизації формування призначених для користувача ролей в корпоративної

інформаційної мережі, що включає в себе 22 користувача, кожен з яких описувався 18 параметрами. Отримані результати дозволили виділити користувачів, що характеризуються аномальним поведінкою в комп'ютерній мережі. На прикладі рішення задачі кластеризації розроблений метод показав 100% точність. На тих же даних точність відомих методів кластеризації EM і g-means склала відповідно 89% і 76,1%.

Створення інформаційної технології

Створення інформаційної технології закладається в тому, що потрібно спроектувати та продумати алгоритм парсингу соціальної мережі фейсбук.
Для парснигу було використано бібліотеку XPatch.

Xpath – це мова запитів до елементів xml або xhtml документа. Також як SQL, xpath є декларативною мовою запитів.

Не завжди вдається отримати доступ до цікавого вузлу за допомогою предиката або кроків адресації. Дуже часто на одному рівні ієрархії знаходиться наскільки вузлів однакового типу і необхідно вибрати «тільки перші» або «тільки другі» вузли. Для таких випадків передбачені колекції. Колекції xpath дозволяють отримати доступ до елементу по його індексу. Індекси відповідають тому порядку, в якому елементи були представлені в оригінальному документі. Порядковий номер у колекціях відраховується від одиниці.

Отже був написаний парсер який збирає дані з соціальної мережі по таким критеріям та робить графічну діагностику по таким параметрам. (Детальна інформація буде описана в розділі 3 з представленим програмним кодом):

- Пошук активної аудиторії. (Користувачі, які часто ставлять лайки, пишуть коментарі).
- Збір коментарів по запитам.
- Комбінування аудиторії
- Парсинг з параметрами.
- Збір контактів адміністраторів груп та власних сторінок (блогерів).
- Збір ботів, спамерів, не активних юзерів.
- Нові користувачі (збір нових користувачів за відведений час). [4].

Принцип технології заключається в тому що буде обработано максимальна кількість даних та зберігається в базі даних в власному кабінеті авторизованого користувача. Сервіс має велике навантаження на БД і працює з великими даними.

Висновки

Розглянуто існуючі методи аналізу використання соціальних мереж. Аналіз результатів роботи довів, що використання даної інформаційної технології надає можливість краще аналізувати цільову аудиторію за рахунок використання методу ієрархічної кластеризації. Тестування показало надійну роботу розробленої програми.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ших Клара. Эра Facebook. Как использовать возможности социальных сетей для развития вашего бизнеса/ М.: Манн, Иванов и Фербер,2010.- 295 с.. ISBN: 978-5-91657-103-5
2. Які соціальні мережі популярні в світі [електронний ресурс] // Режим доступу: https://marketer.ua/ua/top-social-media-2018/(дата звернення 10.12.2018) – Назва з екрану.
3. Описание методов API [електронний ресурс] // Режим доступу: https://fb.com/dev/methods/ (дата звернення 14.05.2017) – Назва з екрану.
4. Huisman M., Duijn M. Software for Social Network // Analysis Proceedings of the Sixth International Conf. on Logic and Methodology, August 17–20. – Amsterdam, The Netherlands, 2004. – pp. 578-600.

Кисіль Віталій Григорович — *студент групи 2КН-17м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, email: vitaliy.kusil@gmail.com*

Науковий керівник – ***Озеранський Володимир Сергійович*** *— ст.викл. кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.*

Kysil Vitalii Grigorovich *- student of group 2КН-17м, faculty of information technologies and computer engineering, Vinnytsia National Technical University, Vinnytsia, email: vitaliy.kusil@gmail.com*

Scientific supervisor **- Ozeranskyy S. Volodymyr** *- Senior lecturer. Department of Computer Science, Vinnitsa National Technical University, Vinnytsia.*