УДК 004.946

Сергеєв Єгор Ігорович
студент
Інституту прикладного системного аналізу
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»
Сергеев Егор Игоревич
студент
Института прикладного системного анализа
Национального технического университета Украины
«Киевский политехнический институт имени Игоря Сикорского»
Serheiev Yehor
Student of the
Institute for Applied System Analysis of the
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

ДОСЛІДЖЕННЯ ГЕТЕРОГЕННИХ ДЖЕРЕЛ ДАНИХ НА ПРИКЛАДІ СОЦІАЛЬНИХ МЕРЕЖ ТА РЕАЛІЗАЦІЯ ЇХ ОБРОБКИ І ПРОГНОЗУВАННЯ ДАНИХ

ИССЛЕДОВАНИЕ ГЕТЕРОГЕННЫХ ИСТОЧНИКОВ ДАННЫХ НА ПРИМЕРЕ СОЦИАЛЬНЫХ СЕТЕЙ И РЕАЛИЗАЦИЯ ИХ ОБРАБОТКИ И ПРОГНОЗИРОВАНИЯ ДАННЫХ

HETEROGENEOUS DATA SOURCES RESEARCH ON THE EXAMPLE OF SOCIAL NETWORKS AND IMPLEMENTATION OF THEIR DATA PROCESSING AND FORECASTING

Анотація. *Метою даної роботи є дослідження методів та технологій аналізу даних з гетерогенних джерел, з використанням різних засобів обробки великих даних.*

Ключові слова: *інтелектуальний аналіз даних, гетерогенні джерела даних, розподілені обчислення, великі дані.*

Аннотация. *Целью данной работы является исследование методов и технологий анализа данных с гетерогенных источников, с использованием различных способов обработки больших данных.*

Ключевые слова: *интеллектуальный анализ данных, гетерогенные источники данных, распределенные вычисления, большие данные.*

Summary. *The purpose of this work is to study the methods and technologies of data analysis from heterogeneous sources, using various methods for large data sources processing.*

Key words: *intellectual data analysis, heterogeneous data sources, distributed computing, big data.*

Постановка проблеми. На сьогоднішній день обсяг даних стає занадто великим для того, щоб була можлива їх обробка традиційними методами та алгоритмами. Таку проблему з великою кількістю даних ще називають великими даними (big data), а прикладом джерела такої кількості даних можуть бути гетерогенні джерела, такі як соціальні мережі.

Тому розробка програмного забезпечення, що може якісно і швидко завантажувати та обробляти дані з таких джерел є дуже актуальним напрямком дослідження, у час, коли кількість даних

в гетерогенних джерелах зростає з небувалою швидкістю, а єдиного підходу для інтелектуального аналізу не існує.

Формулювання цілей статті (постановка завдання). Огляд методології та існуючих методів для обробки великих масивів даних, а також прогнозування.

Виклад основного матеріалу. Наприкінці минулого століття принципово змінився стиль життя людини, яка отримала змогу налагоджувати ті контакти, які є важливими та цінними для неї, найбільшою мірою відповідають її інтересам. Тепер для того, щоб взаємодіяти з представниками інших культур, їй не обов'язково навіть виходити з дому. В Інтернет-середовищі формується власний специфічний дискурс, який змінює характер взаємодії між індивідами. Соціальні мережі практично не піддаються зовнішньому контролю, не мають єдиного центру, а тому кожен в праві діяти у них так, як вважає за потрібне. Інтернет як сучасний засіб масової комунікації формує з пасивного слухача активного співучасника та співтворця. Сучасна людина не просто прагне дізнаватись про події, вона прагне їх створювати. Дані, доступні в соціальних мережах, можуть дати уявлення про людину, громади та суспільство в цілому, що раніше було неможливим в таких глобальних масштабах. Такі цифрові медіа дані виходять за межі фізичного світу для вивчення людських відносин і допомагають визначати популярні соціальні та політичні настрої для регіональних груп без використання опитувань. Соціальні мережі регулярно фіксують основні маркетингові тренди, настрої та напрями та є ідеальними джерелами для вивчення та обробки. Однак, отримання необхідної інформації з соціальних мереж є дуже важким через конкретні виклики, що несуть такі дані. Методи інтелектуального аналізу даних можуть допомогти ефективно оброблювати інформацію з соціальних мереж та обійти три основні проблеми таких даних [1].

В першу чергу, набори даних соціальних мереж є величезними, наприклад, сотні мільйонів користувачів популярної соціальної мережі Facebook, що кожного дня генерують нові дані. Без автоматичного, програмного методу обробки та аналізу даних соціальних мереж, процедура обробки таких даних стає неосяжною та неможливою в виконанні в розумному проміжку часу.

По-друге, дані соціальних мереж містять в собі так званий «шум», а саме: спам, незмістовні дані, пусті повідомлення. Такі дані потрібно виокремлювати від існуючого набору чи ігнорувати, проте сам механізм виявлення шумів є досить складним і потребує використання необхідних алгоритмів розпізнання.

По-третє, такі дані є динамічними, часті зміни та оновлення впродовж короткого періоду часу є не лише загальним, але важливим аспектом, який слід враховувати при обробці таких даних. Інші набори даних можуть містити декілька із зазначених проблем, проте не всі зразу. Наприклад, набір традиційних веб-сторінок генерує дані, які є великими за об'ємом та містять шуми, проте, порівняно з даними із соціальними мережами, не є настільки динамічним, це пов'язано з тим, що соціальні мережі дозволяють глобальне створення різного типу контенту, інтерактивних даних різних форм та видів. Наприклад, стандартне повідомлення у блозі та коментарі до нього, при чому кожен коментар може містити різного типу дані, від звичайного тексту до зображень та відео.

Іншим важливим аспектом соціальних мереж є їх реляційний характер, який може ускладнювати аналіз обробки даних, проте вони це не є новою проблемою для пошуку даних, деякі методи обробки розроблені спеціально для ідентифікації шаблонів та правил, що базуються на реляційних атрибутах даних.

Методи інтелектуального аналізу викликані допомогти дослідникам та практикам обійти всі названі проблеми, використання таких технік та методів дозволяє покращити результати пошуків пошукових систем, допомогти в реалізації спеціалізованої маркетингової політики для бізнесу, надати нові знання про соціальні структури для соціологів, персоналізувати веб-системи для користувачів завдяки розвитку рекомендаційних систем і навіть допомогти розпізнавати спам та захищати від нього.

Оскільки соціальні мережі включають в себе неймовірну кількість даних, що оновлюється з кожним днем, то таке джерело даних є ключовим в дослідженнях соціальних наук. Пошук та обробка даних з соціальних мереж є складною задачею, через безліч факторів, в тому числі неоднозначність людської мови, багатозначність слів, псевдоніми для одного і того самого користувача, неправильне зображення даних та двозначність взаємовідносин між користувачами. В даному розділі розглянуто основні напрями та методи дослідження соціальних мереж.

Оскільки дані з соціальних мереж мають гетерогенний характер, а також оновлюються з надзвичайною швидкістю, то для обробки таких великих масивів даних потрібно використовувати спеціальні інструментарії перевірені часом, що зарекомендували себе як швидкі та надійні, одною з таких технологій є використання підходу MapReduce. Потрібно зазначити, що в даному контексті MapReduce розглядається не як конкретна реалізація, як наприклад, Hadoop MapReduce, а підхід для обробки великих масивів даних. Розглянемо основні принципи такого підходу.

На вхід функції Map можна подавати дані різного типу — масив, документ та інші, кожна така частина є сукупністю елементів і жоден з яких не зберігається в двох частинах даних. Усі виходи із завдання Map та входи завдання Reduce є парами ключів та значень, така форма входів є обумовле-

ною композицією декількох процесів MapReduce. Функція Map приймає на вхід деяке значення, наприклад, частину текстових даних та після обробки, що була задана, видає на вихід або пусту пару, або масив пар ключів та значень. Потрібно зазначити, що ключі повинні бути унікальними, однак самі пари можуть мати один і той же ключ проте з різними значеннями, які після обробки усіма вузлами Map згодом будуть згруповані та передані на відповідний обробник Reduce. При чому один з вузлів Map як і Reduce може оброблювати декілька потоків даних на вхід, тобто, після обробки одного пакету даних та генерації виходу, даний вузол може прийняти та обробити інші частини даних, якщо такі є і ще не оброблені.

Після обробки всієї послідовності вхідних даних вузлами Map, основний контролер системи MapReduce запускає процес групування даних по ключу, де отримані на вхід пари:

$$\ll K_1, V_1 >, \langle K_2, V_2 \rangle, \langle K_1, V_3 \rangle, \ldots, < K_N, V_N \gg .$$

Перетворюються в групові пари виду:

$$< K_1, < V_1, V_3, \ldots \gg, \ldots, < K_N, < V_N, \ldots \gg .$$

Після такого перетворення отримані дані можна передавати на подальшу обробку обробників операції Reduce. Виходом операції Reduce, як і в операції Map, є пари ключів та значень, при чому ключі та значення можуть мати інший тип, в залежності від поставленої задачі, але в основному вони матимуть один і той же тип. Отримані значення з усіх вузлів групуються в один файл, що і є результатом необхідної операції. Необхідно зазначити, що для виконання завдання зачасту може бути недостатньо виконання однієї операції MapReduce, тому вихідні значення попередньої операції можуть стати вхідними для наступної операції MapReduce.

Для виконання прогнозування даних можна використати нейронні мережі. При чому для використання нейронних мереж для прогнозування недостатньо просто створити мережу, мережа з входом і виходом є гарною для регресії, проте для прогнозування використовується так звана рекурсивна мережа, де на входи подається час, а на вхід наступної мережі подається вихід з попередньої [4], графічно це можна представити таким чином як на рисунку 1.

Хоча на рисунку зображено лише три нейронні мережі, що об'єднано в одну, для реальної задачі таких може бути набагато більше, при чому на рисунку 2.2 не показано скільки вузлів у одному шарі нейронної мережі, як і не показана кількість шарів. Тепер представимо, що і на вхід такої мережі на кожному кроці подається значення певного інтервалу, така мережа зможе прогнозувати дані, якщо її ваги на входах та між вузлами було правильно скореговано, використовуючи існуючи історичні дані. Можна переписати формулу 2.10 для такого випадку:

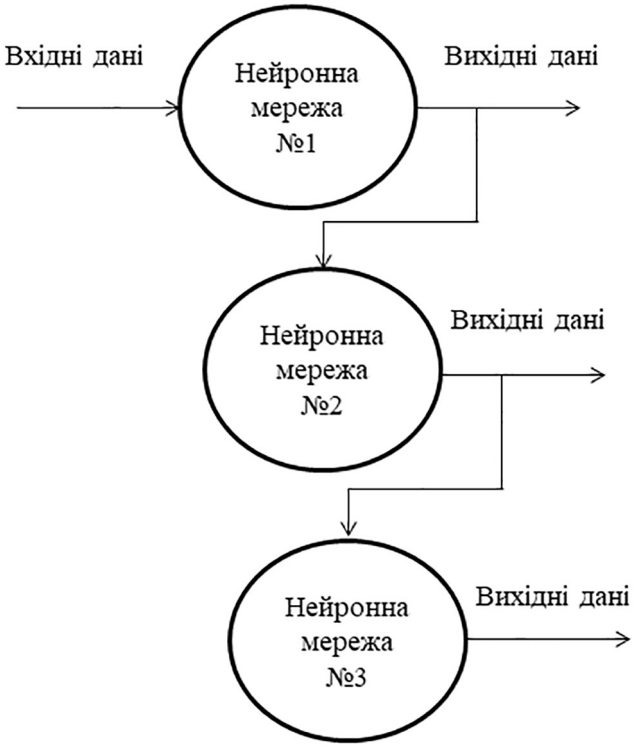

Рис. 1

$$\varphi_t\left(\omega^T a\right) = \tanh\left(\varphi_{t-1}\left(\omega^T a\right) + u^T x_t\right).$$

Де u — вага при вході на конкретну мережу в рекурсивній мережі, а коефіцієнт x — значення входу на конкретну мережу. Така проста мережа повинна досить непогано виконувати задачу прогнозування на короткий період часу.

Для того, аби виконувати прогнозування на великі терміни, використовується так звана LSTM (Long Short-Term) нейронна мережа, це покращена рекурсивна мережа, що складається з декількох мереж, кожна з яких має свою внутрішню змінну стану, що передається з одного вузла в інший і модифікується так званими затворами.

Затвор забуття. Його основна функціональність — змінювати внутрішній стан мережі. Затвор — функція, що приймає значення входу на конкретну мережу, вагу і т.д. і в результаті видає число, що потім перемножується зі значенням внутрішнього стану, при виході 0 із затвору, внутрішній стан стає нулем, саме тому його називають затвором забуття. Формула такого перетворення:

$$f_t = \sigma\left(\omega_f\left[h_{t-1}, x_t\right] + b_f\right).$$

Вхідний затвор використовує вихідне значення з попередньої мережі і вхідне значення, в результаті отримуючи число від 0 до 1, що згодом буде використано для перерахунку внутрішнього стану на даній ітерації:

$$i_t = \sigma\left(\omega_i\left[h_{t-1}, x_t\right] + b_i\right).$$

А значення внутрішнього стану мережі можна перерахувати за формулою:

$$C_t = f_t C_{t-1} + i_t C_t .$$

Вихідний затвор контролює значення на виході з коефіцієнтом внутрішнього стану, тобто яку частину внутрішнього стану мережі можна передати на вихід:

$$O_t = \sigma\left(\omega_o \left[h_{t-1}, x_t\right] + b_o\right),$$
$$h_t = O_t \tanh\left(C_t\right).$$

В результаті дану модель мережі можна використовувати для прогнозування на досить великі періоди часу, і на відміну від моделі лінійної регресії, можна використовувати будь-які дані, не лише ті, що можна змоделювати лінійно.

Одним з найпростіших методів, що дозволяє виконувати прогнозування для подальшої серії даних у часі є так званий Exponential Smoothing (експоненційного згладжування) метод. Даний метод дозволяє виконувати прогнозування на короткий період, як, наприклад, на наступний тиждень чи наступний день чи рік, проте ніхто не забороняє виконувати прогнозування на подальші інтервали, однак буде використано попередньо прогнозовані дані [2].

Приймемо, що дані за певний період часу t позначаються x_t, а прогноз для певного часу будемо позначати як s_t. В результаті основну формулу методу експоненційного згладжування можна записати так:

$$s_t = \alpha x_{t-1} + (1-\alpha) s_{t-1} .$$

З даної формули видно, що значення наступного прогнозу залежить від попереднього прогнозу та попереднього значення. При чому, прогноз для першого значення даних є рівним самому значенню, а якщо потрібно прогнозувати для більших інтервалів ніж на один, то, аналогічно отримане прогнозоване значення є рівним реальному, що у даному випадку є очікуваним. Параметр α обирається довільним, проте в основному береться рівним значенню 0,5 [3].

Даний алгоритм називається експоненційним, оскільки рекурсивне представлення можна переписати більш загальною формулою, що є експоненційною:

$$s_t = \alpha\left[x_t + (1-\alpha)x_{t-1} + (1-\alpha)^2 x_{t-2} + \ldots + (1-\alpha)^{t-1} x_1\right] +$$
$$+ (1-\alpha)^t x_0 .$$

Хоча даний алгоритм може дати досить точні результати при непогано вибраному параметру α однак існують покращені версії даного алгоритму інших порядків, а саме другого та третього.

Алгоритм другого порядку передбачає введення параметру b_t — найкращої оцінки для даного інтервалу часу, в результаті формулу 2.17 можна переписати так:

$$s_t = \alpha x_{t-1} + (1-\alpha)(s_{t-1} + b_{t-1}),$$

$$b_t = \beta(s_t - s_{t-1}) + (1-\beta) b_{t-1} .$$

Останнім, проте не менш важливим способом прогнозування значення даних в майбутньому, що розглядаються в контексті даної роботи, є використання ARIMA моделей. Такі моделі, теоретично, є найбільш загальними моделями для прогнозування часових рядів, які можна зробити стаціонарними шляхом диференціації (при необхідності), у поєднанні з нелінійними перетвореннями. Випадкова величина, що є часовим рядом є стаціонарною, якщо її статистичні властивості постійно змінюються, такі серії не мають певних тенденцій, а вони варіюються навколо свого середнього значення і мають постійну амплітуду, що коливаються послідовно. Модель ARIMA для прогнозування є лінійною, тобто схожою на регресивного типу. ARIMA (Auto-Regressive Integrated Moving Average) складається з основних трьох компонентів авторегресивного, інтегрованого та середнього переміщення [5].

Авторегресивний компонент посилається на використання попередніх значень в регресійному рівнянні для серій Y. В визначенні $ARIMA(p,d,q)$ параметр p впливає на кількість попередніх результатів у рівнянні, а сама модель авторегресії позначається як $AR(p)$. Для прикладу, модель $AR(2)$ або $ARIMA(2,0,0)$ буде виглядати так:

$$Y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + e_t .$$

Другим компонентом ARIMA моделі є інтегрований $I(d)$ компонент, де параметр d відповідає за ступінь різності в даній моделі. Різниця між серіями передбачає просту операцію віднімання значень теперішнього і попередніх значень d разів. Зачасту різність використовується для стабілізування серій значень коли стаціонарне допущення не виконується.

Третім компонентом ARIMA моделі є середнє переміщення, позначається як $MA(q)$ та позначає помилку моделі як комбінацію попередніх значень помилок e_t. Параметр q відповідає за число попередніх значень помилок у моделі. В результаті, модель середнього переміщення можна записати так:

$$Y_t = c + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} + e_t .$$

В результаті поєднання усіх трьох моделей дає нам ARIMA модель для розрахунку майбутніх значень, а саму модель модель можна записати як лінійне рівняння виду:

$$Y_t = c + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} + e_t .$$

Висновки. В даній роботі було розглянуто методологію обробки даних з можливих гетерогенних джерел даних на прикладі соціальних мереж. Було показано, що найкращим способом для такої обробки є використання технологій, що дозволяють використовувати операції MapReduce на паралельних кластерах. Було продемонстровано основні можливі методи для обробки та прогнозування отриманих даних, на основі збору попередніх, а саме ARIMA, Exponential Smoothing та рекурсивні нейронні мережі.

Література

1. Феномен соціальної мережі в інформаційному середовищі [Електронний ресурс] / А. Кромська // Науковий блог Національного університету «Острозька академія», 2015 — Режим доступу: https://naub.oa.edu.ua/2015/феномен-соціальної-мережі-в-інформац

2. Simple Exponential Smoothing [Electronic resource]. — OTexts, Режим доступу: https://www.otexts.org/fpp/7/1

3. Exponential Smoothing for Predicting Demand [Text] / Brown, Robert G. // Cambridge, Massachusetts — 1956, p. 15.

4. Neural Networks for Time Series Prediction [Electronic resource] / Touretzky D., Laskowski K. / Artigical Neural Networks, 2006 — Режим доступу: https://www.cs.cmu.edu/afs/cs/academic/class/15782-f06/slides/timeseries.pdf

5. Introduction to ARIMA: nonseasonal models [Electronic resource]. — DukePeople — Режим доступу: https://people.duke.edu/~rnau/411arim.htm