DOI 10.31891/2307-5732-2022-311-4-115-118
УДК 004.93, 004.8

КРИВЕНЧУК Ю. П.
Національний університет "Львівська політехніка"
https://orcid.org/0000-0002-2504-5833
e-mail: Yurii.P.Kryvenchuk@lpnu.ua
ХАНАС М.-Ю. Р.
Національний університет "Львівська політехніка"
e-mail: hanasura79@gmail.com

АЛГОРИТМ ВИДОБУВАННЯ ТА ОПРАЦЮВАННЯ СПОРІДНЕНИХ ДАНИХ В СОЦІАЛЬНИХ МЕРЕЖАХ

Основою даної роботи є один з алгоритмів, що використовуються в рекомендаційних системах - рекомендаційна система основана на фільтрації вмісту. Вона аналізує пости користувачів у Twitter та вираховує їх інтереси. Особливістю даної системи є те, що даний алгоритм використовує паралельні обчислення та частотний аналіз тексту. Це дає змогу об'єднати людей з однаковими інтересами.

Ключові слова: рекомендаційна система, дата майнінг, big data, nltk, tweepy.

Yurii KRYVENCHUK, Mykhailo-Yurii KHANAS
Lviv Polytechnic National University

ALGORITHM OF DATA MINING AND PROCESSING OF RELATED DATA IN SOCIAL NETWORKS

We live in a time of rapid growth of information technology, which is firmly entrenched in our daily lives. It is simply impossible to imagine a modern person without social networks, because they perform a communicative and informational function, namely: communication, information retrieval, news exchange, etc. Five hundred million tweets are posted daily, making Twitter a major social media platform from which topical information on events can be extracted. So, there is a lot of information available to the user, which is difficult to identify something specific and necessary in the usual way viewing. Accordingly, there is a need for technologies that can quickly process large amounts of data and highlight only the information that is useful to a particular user. This technology called recommender systems. It automatically suggest items to users that might be interesting for them. Due to the desire to unite people with common interests, it is relevant to develop a recommendation system based on social networks that help in personification of the user and compilation of his psychotype using his profile.

The paper has description and results of the creation of recommendation system. The basis of this work is one of the algorithms used in recommendation systems - the recommendation system is based on content filtering. It analyzes users' Twitter posts and calculates their interests. If we consider all the words, our model will not have good results and do not pay attention to what is important to use. Therefore, the most important step is always filtering data, so the number one task is to speed up the time of filtering text and retrieving data from the social network for further processing. The feature of this system is that this algorithm uses parallel calculations and frequency analysis of the text.

Keywords: recommendation system, date mining, big data, nltk, tweepy.

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Ми живемо в час швидкого росту інформаційних технологій, які міцно увійшли в наше щоденне життя. Сучасну людину уявити без соціальних мереж просто неможливо, бо вони виконують комунікативну-інформаційну функцію, а саме: спілкування, пошук інформації, обмін новинами і тд. Дана робота покликана об'єднати людей зі спільними інтересами.

Головною ідеєю дослідження є за максимально короткий час обробити, проаналізувати, а також відфільтрувати за ознаками та атрибутами великий потік дописів користувачів у соціальних мережах. Дані можуть мати велику кількість "зайвих слів", наприклад: and, or, the, firstly, does [1]. Якщо брати до уваги всі слова, то модель не буде давати хороших результатів і не звертатиме увагу на дійсно важливі вирази. Тому завжди найважливішим кроком є фільтрування даних[2], отже завдання номер один - пришвидшення часу фільтрації тексту та діставання даних з соціальної мережі для її подальшої обробки.

Аналіз останніх джерел

В роботі [3] було проведено всебічний огляд методів прогнозування дружби в LBSN та запропоновано таксономію для організації існуючих методів, було висунуто пропозицію про п'ять нових методів для усунення недоліків, намагаючись знайти баланс між оптимізацією обчислювальних ресурсів та покращенням передбачуваної потужності. Таким чином, вони представили загальну панораму завдання прогнозування дружби в області LBSN із збалансованою глибиною, щоб полегшити дослідження та розробку додатків у реальному світі щодо цього важливого питання. Недоліком роботи є погана точності при нестачі даних.

В роботі [4] покращується SLOMA, застосовуючи соціальну регуляризацію до факторізації підматриць, позначену як SLOMA++. Таким чином, запропонована модель може отримати користь як від соціальних рекомендацій, так і від локального припущення низького рангу. Експериментальні результати з двох реальних наборів даних, Yelp і Douban, демонструють перевагу запропонованих моделей над LLORMA і MF. Але недоліком даної роботи є те, що важко охопити складні відносини та взаємодії між користувачами та елементами.

В роботі [5] було запропоновано новий метод соціальних рекомендацій, а саме Probabilistic Relational Matrix Factorization (PRMF), яка спрямована дізнатися оптимальну соціальну залежність між користувачами, щоб підтвердити точність рекомендацій. У даній рекомендаційній системі є два основних завдання, а саме: рекомендація предметів та прогнозування рейтингу.

Формулювання цілей статті

Метою роботи є: швидко і надійно обробити Великі дані для отримання рекомендацій друзів, пов'язаних спільними інтересами.

Виклад основного матеріалу

Для створення рекомендаційної системи було використано декілька досить поширених бібліотек і методик. Бібліотеки: multiprocessing, nltk, tweepy. Методики: bag-of-words.

Multiprocessing:

Це пакет, який підтримує породження процесів з використанням API, подібно до модуля threading. Пакет multiprocessing пропонує як локальну, так і віддалену паралельність, ефективно обходячи Global Interpreter Lock, використовуючи підпроцеси замість потоків. Завдяки цьому модуль дозволяє програмісту повністю використовувати кілька процесорів на певній машині. Він працює як на Unix, так і на Windows.

Модуль multiprocessing також представляє API, які не мають аналогів у модулі threading.[2] Яскравим прикладом цього є об'єкт Pool, який пропонує зручний засіб паралельного виконання функції за кількома вхідними значеннями, розподіляючи вхідні дані між процесами (паралелізація по даних). Наступний приклад демонструє загальну практику визначення таких функцій в модулі, щоб дочірні процеси могли успішно імпортувати цей модуль.

NLTK:

Це провідна платформа для побудови програм Python, які обробляють людську мову. Він надає прості у використанні інтерфейси для більш ніж 50 мов, а також набір бібліотек обробки тексту для класифікації, токенізації, стемінгу, тегування, синтаксичного аналізу та семантичних міркувань, обгортки для бібліотек NLP індустріального рівня. Проект також має хороше ком'юніті[6].

Завдяки практичному посібнику, що пояснює основи програмування поряд з темами з обчислювальної лінгвістики, а також зрозумілу документацію щодо API, NLTK підходить для лінгвістів, інженерів, студентів, викладачів, дослідників та всіх кого цікавить обробка людської мови. NLTK доступний для Windows та Unix систем.[7]

Tweepy:

Це пакет Python з відкритим кодом, який дає дуже зручний спосіб отримати доступ до API Twitter. Tweepy включає набір класів і методів, що представляють моделі Twitter та ендпоінти API, він прозоро обробляє різні деталі реалізації, такі як:

- Кодування та декодування даних
- HTTP-запити
- Пагінація результатів
- OAuth
- Обмеження запитів
- Потоки

Якби ми не використовували Tweepy, то нам довелось б мати справу з деталями низького рівня, пов'язаними із запитами HTTP, серіалізацією даних, автентифікацією та обмеженнями швидкості. Це може зайняти багато часу та спричинити помилки. Натомість завдяки Tweepy ми зосередились на функціональних можливостях, які хотіли імплементувати. Майже всю функціональність, яку надає Twitter API, можна використовувати через Tweepy.

Bag-of-words:

Модель "торба слів" - це уявлення, яке перетворює довільний текст у вектори фіксованої довжини, підраховуючи, скільки разів з'являється кожне слово. Цей процес часто називають векторизацією, тобто проводиться відстеження кількості слів і ігнорування граматичних деталей та порядок слів.[8] Його називають «мішком» слів, оскільки будь-яка інформація про порядок чи структуру слів у документі відкидається.

Ідея роботи програми

Ми визначаємо певний набір користувачів Twitter. Після цього програма паралельно обробляє їх наступним чином. Для кожного із користувачів ми дістаємо усі його дописи (твіти). Пізніше об'єднуємо це все в один великий текст. Робимо певний препроцесінг, а саме видаляємо стоп слова і розділові знаки. Далі токенізуємо текст, тобто ділимо його на слова. Після цього здійснюємо стемінг - це процес скорочення слова до основи. Будуємо bag-of-words, і на основі результатів проводимо підбірку користувачів по найчастіше вживаних словах, припускаючи що це спільні інтереси.

Приклад роботи системи наведено на рис. 1 та на рис. 2.

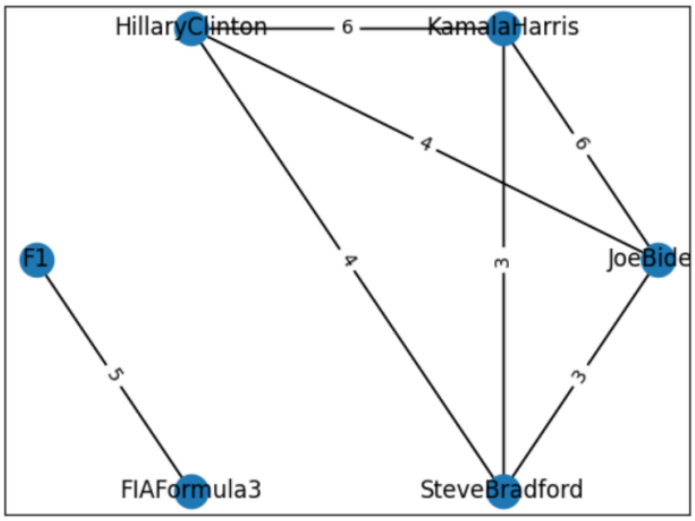

Рис. 1 Результат виконання програми

Для тестування програми було вибрано такі акаунти людей в Twitter: Joe Biden, Hillary Clinton, Kamala Harris, Steve Bradford, F1, FIAFormula3. Як бачимо програма нам видала 2 зважених графа. На ребрах вказано кількість інтересів між акаунтами. Дивлячись на результат програми, можна побачити, що вона досить добре впоралася, оскільки вона зрозуміла, шо є 2 різні групи акаунтів. Також можна зауважити, шо перший граф, містить всіх американських державних і політичних діячів, а другий граф містить акаунти пов'язані з F1. Можна зробити висновок, шо вони справді мають спільні інтереси.

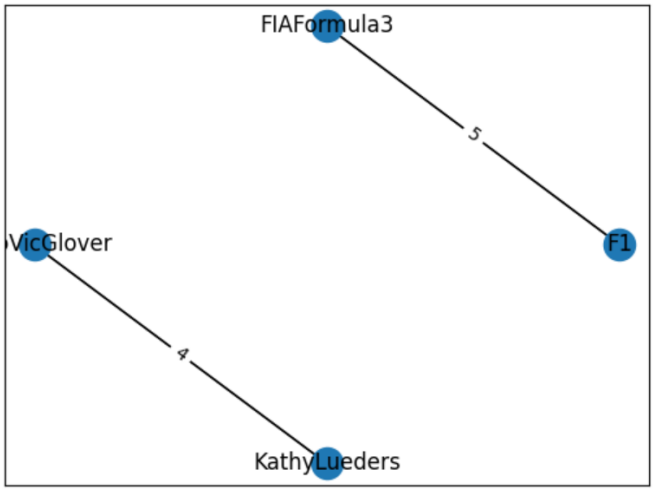

Рис. 2 Результат виконання програми

Для другого тесту взято групу людей, які повязані з космосом і F1. Було вибрано такі акаунти людей в Twitter: AstroVicGlover, KathyLueders, elonmusk, F1, FIAFormula3. Програма нам видала 2 зважених графа. Вона показала хороший результат, оскільки зрозуміла, шо є 2 різні групи акаунтів. Можна побачити, що графи не містять акаунта elonmusk, оскільки наша програма обробляє лише перших 200 публікацій і серед оброблених публікацій людина може почати цікавитися іншими речима ніж ми очікували.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

В результаті проведеної роботи створено рекомендаційну систему, яка швидко обробляє досить великі дані для отримання рекомендацій друзів, пов'язаних спільними інтересами. Ця ідея є дуже актуальною в наш час і дослідження цього напрямку є перспективним. Дана система, може широко застосовуватися у різних соціальних мережах та дасть змогу об'єднювати людей зі спільними інтересами.

Література

1. X. He, Z. He, X. Du, and T.-S. Chua, "Adversarial personalized ranking for recommendation," in The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, 2018

2. J. Tang, X. Hu, and H. Liu, “Social recommendation: a review,” Social Network Analysis and Mining, vol. 3, no. 4, 2013.

3. J. C. Valverde-Rebaza, M. Roche, P. Poncelet, and A. D. A. Lopes, The role of location and social strength for friendship prediction in location-based social networks, Information Processing and Management, vol. 54, no. 4, 2018, doi: 10.1016/j.ipm.2018.02.004.

4. H. Zhao, Q. Yao, J. T. Kwok, and D. L. Lee, Collaborative filtering with social local models, 2017, vol. 2017-November, pp. 645–654. doi: 10.1109/ICDM.2017.74.

5. Y. Liu, P. Zhao, X. Liu, M. Wu, and X.-L. Li. Learning Optimal Social Dependency for Recommendation. arXiv preprint arXiv:1603.04522, 2016.

6. Steven Bird, Ewan Klein, and Edward Loper, Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit, O'Reilly Media, 2009

7. J. Yu, M. Gao, J. Li, H. Yin, and H. Liu, “Adaptive implicit friends identification over heterogeneous network for social recommendation,” in Proceedings of the 27th ACM International Conference on Information and Knowledge Management. ACM, 2018

8. Davoudi and M. Chatterjee. Modeling trust for rating prediction in recommender systems. In SIAM Workshop on Machine Learning Methods for Recommender Systems, SIAM, 2016.

References

1. X. He, Z. He, X. Du, and T.-S. Chua, “Adversarial personalized ranking for recommendation,” in The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, 2018

2. J. Tang, X. Hu, and H. Liu, “Social recommendation: a review,” Social Network Analysis and Mining, vol. 3, no. 4, 2013.

3. J. C. Valverde-Rebaza, M. Roche, P. Poncelet, and A. D. A. Lopes, The role of location and social strength for friendship prediction in location-based social networks, Information Processing and Management, vol. 54, no. 4, 2018, doi: 10.1016/j.ipm.2018.02.004.

4. H. Zhao, Q. Yao, J. T. Kwok, and D. L. Lee, Collaborative filtering with social local models, 2017, vol. 2017-November, pp. 645–654. doi: 10.1109/ICDM.2017.74.

5. Y. Liu, P. Zhao, X. Liu, M. Wu, and X.-L. Li. Learning Optimal Social Dependency for Recommendation. arXiv preprint arXiv:1603.04522, 2016.

6. Steven Bird, Ewan Klein, and Edward Loper, Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit, O'Reilly Media, 2009

7. J. Yu, M. Gao, J. Li, H. Yin, and H. Liu, “Adaptive implicit friends identification over heterogeneous network for social recommendation,” in Proceedings of the 27th ACM International Conference on Information and Knowledge Management. ACM, 2018

8. Davoudi and M. Chatterjee. Modeling trust for rating prediction in recommender systems. In SIAM Workshop on Machine Learning Methods for Recommender Systems, SIAM, 2016.

Рецензія/Peer review : 13.06.2022 р. Надрукована/Printed :02.08.2022 р.