УДК 004.735 : 004.912

А.О. Концевой

АНАЛІЗ ПРОФІЛІВ УЧАСНИКІВ СОЦІАЛЬНИХ МЕРЕЖ

Вінницький національний технічний університет

Анотація

В даній статті розглянуто перспективи, завдання, методи і додатки для аналізу мережевих (соціальні зв'язки між користувачами) і текстових (повідомлення і профілі користувачів) даних: визначення демографічних атрибутів користувачів, пошук описів подій в корпусах повідомлень, ідентифікація користувачів різних мереж, пошук спільнот користувачів і оцінка інформаційного впливу між користувачами.

Ключові слова: соціальні мережі; соціальні дані; призначені для користувача дані; мережевий аналіз; аналіз соціальних мереж; аналіз вмісту; веб-сервіси; мікроблоги; комп'ютерна лінгвістика; теорія графів; машинне навчання; розподілені алгоритми і системи

Annotation

This article considered such topics as problems, problems, methods and applications for analysis of network (social connections between users) and text (messages and user profiles) of data: was given definition of demographic attributes of users, search of event descriptions in message corps, identification of users of different networks , search for user communities and measure the informational impact among users.

Key words: social networks; social data; user data; network analysis; analysis of social networks; content analysis; web services; microblogging; computer linguistics; graph theory; machine learning; distributed algorithms and systems

Аналіз соціальних даних стрімко набирає популярність у всьому світі завдяки появі в 1990-х роках онлайнових сервісів соціальних мереж (SixDegrees, LiveJournal, Facebook, Twitter, YouTube та інші). З цим пов'язаний феномен соціалізації персональних даних: стали публічно доступними факти біографії, листування, щоденники, фото-, відео-, аудіоматеріали, замітки про подорожі тощо.

Отже, соціальні мережі є унікальним джерелом даних про особисте життя та інтереси реальних людей. Це відкриває безпрецедентні можливості для вирішення дослідних і бізнес-задач (багато з яких до цього неможливо було вирішувати ефективно через брак даних), а також створення допоміжних сервісів і додатків для користувачів соціальних мереж. Крім того, цим обумовлюється підвищений інтерес до збору і аналізу соціальних даних з боку компаній і дослідницьких центрів [1].

Фахівці з дослідницьких центрів і компаній по всьому світу використовують дані соціальних мереж для моделювання соціальних, економічних, політичних та інших процесів від персонального до державного рівня з метою розробки механізмів впливу на ці процеси, а також створення інноваційних аналітичних і бізнес-додатків та сервісів.

Разом з тим, при роботі з соціальними даними потрібно брати до уваги такі фактори, як нестабільність якості призначеного для користувача контенту (спам і неправдиві акаунти), проблеми із забезпеченням приватності особистих даних користувачів при зберіганні і обробці, а також часті поновлення моделі користувача і функціоналу. Все це вимагає постійного вдосконалення алгоритмів розв'язання різних аналітичних і бізнес-задач [2].

Обробка соціальних даних вимагає також розробки відповідних алгоритмічних і інфраструктурних рішень, що дозволяють враховувати їх розмірність. Наприклад, база даних соціальної мережі Facebook на сьогоднішній день містить більше 1 мільярда акаунтів користувачів і більше 100 мільярдів зв'язків між ними. Кожен день користувачі додають більше 200 мільйонів фотографій і залишають більше 2 мільярдів коментарів до різних об'єктів мережі. На сьогоднішній день більшість існуючих алгоритмів, що дозволяють ефективно вирішувати актуальні завдання, не здатні обробляти дані подібної розмірності за прийнятний час. У зв'язку з цим виникає потреба в нових рішеннях, що

дозволяють здійснювати розподілену обробку і зберігання даних без істотної втрати якості результатів.

Веб-інтерфейси соціальних мереж є джерелами даних реального часу і призначені для перегляду і взаємодії зі сторінками соціальної мережі у веб-браузері або для використання даних користувачів спеціалізованими додатками. Оскільки сценарії використання інтерфейсів соціальних мереж не передбачають автоматичного збору даних множини користувачів з метою побудови соціального графа, то виникає ряд проблем:

- приватність;
- слабка структурованість;
- обмеження доступу;
- розмірність даних обумовлює необхідність в паралельному методі збору даних, а також в методах отримання репрезентативної вибірки користувачів соціальної мережі (семплірування).

При заповненні свого профілю в соціальній мережі користувачі найчастіше помилково чи навмисно не заповнюють деякі поля або дають неправдиву інформацію про факти своєї біографії, інтереси та вподобання. Крім того, в тематичних мережах (Twitter, YouTube) призначений для користувача профіль часто обмежений набором базових атрибутів, недостатнім для вирішення багатьох задач, які передбачають персоналізацію результатів.

Таким чином, актуальні методи часткової ідентифікації авторів повідомлень за значеннями їх демографічних атрибутів. Зокрема, в системах інтернет-маркетингу і рекомендацій особливу важливість представляє визначення демографічних атрибутів користувача для цільового просування товарів і послуг в групах користувачів з однаковими значеннями атрибутів. Крім інтернет-сервісів, такі демографічні характеристики знаходять застосування в різних дисциплінах: соціологія, психологія, кримінологія, економіка, управління персоналом та ін.

Демографічні атрибути можна умовно розділити на категоріальні (стать, національність, раса, сімейний стан, рівень освіти, професія, працевлаштованість, релігійні і політичні погляди) і чисельні (вік, рівень доходів) [3].

Повідомлення користувачів соціальних мереж складають істотну частку текстового контенту сучасного Інтернету. Крім того, соціальні мережі часто виступають в ролі неформальних ЗМІ, де будь-який користувач може опублікувати новинне повідомлення про події, що відбуваються (інформаційні приводи).

Разом з тим, задача автоматичного завантажування набору повідомлень про невідомому заздалегідь подію є нетривіальною в силу наступних чинників:

- великий обсяг вхідних даних (наприклад, користувачі Twitter публікують кілька тисяч повідомлень щосекунди);
- велика кількість нерелевантних / неінформативних повідомлень;
- користувачі можуть по-різному описувати одну і ту ж подію;
- різні події можуть збігатися за часом;
- складність поділу події і його підподій (наприклад, Олімпійські ігри і конкретний футбольний матч в рамках цієї першості) [4].

Потенційною сферою застосування є пошук і складання короткої інфографіки реакції користувачів на невідомі або заздалегідь певні оффлайн- і онлайн-події. Прикладами таких подій можуть служити черговий випуск телевізійного шоу, спортивні події, стихійні лиха, політичні події, запуск нового сервісу для користувачів соціальної мережі тощо.

Природньою властивістю людського суспільства є тенденція до об'єднання в різні спільноти. Аналогічна картина спостерігається в соціальних мережах, де користувачі об'єднуються явно (використовуючи засоби мережі для створення груп і взаємодії всередині них) або неявно – встановлюючи зв'язки на основі загальної або схожою діяльності, ролі, соціального кола, інтересу або інших властивостей.

Пошук спільнот користувачів є важливим інструментом вивчення і аналізу соціальних мереж, що дозволяє досліджувати модульну організацію мережі і використовувати отриману інформацію для вирішення різних завдань. Наприклад, знання про структуру спільнот незамінні для передбачення зв'язків і атрибутів користувачів, розрахунку близькості користувачів в соціальному графі, оптимізації потоків даних в соціальної мережі, деяких аналітичних додатків тощо.

Інформація про спільноти (модульній структурі) соціальної мережі на глобальному рівні знаходить застосування в системах рекомендацій, фільтрації спаму і у багатьох інших додатках.

Автоматично певні спільноти найближчих контактів користувача в соціальній мережі можуть застосовуватися для оптимізації потоків вхідної та вихідної інформації (відправити повідомлення тільки спільноті "Колеги", прочитати новини тільки від спільноти "Близькі друзі") [5].

Отже, аналіз профілів учасників соціальних мереж є перспективним напрямом досліджень, що вимагає поєднання методів мережевого аналізу, комп'ютерної лінгвістики та машинного навчання. Отримані розв'язки актуальних задач цього напряму дозволять створити інноваційні та економічно обґрунтовані сервіси та додатки в індустрії ІТ.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Рудченко Д.В. Аналіз даних соціальних мереж для забезпечення безпеки / Д.В. Рудченко // Вісник національного технічного університету "ХНУ". – 2017. – 50 с.
2. Мазуренко В.В. Огляд моделей аналізу соціальних мереж / В.В. Мазуренко, С.Д. Штовба // Вінницький національний технічний університет "ВНТУ". – 2017. – 20 с.
3. Кондратенко Н. Р. Нечеткие модели принятия решений в задачах прогнозирования взаимоотношений в социальных группах / Н. Р. Кондратенко, С. В. Лужецкий // Наукові праці ВНТУ. — 2009 — № 2.
4. Куликова А. А. Подход к классификации пользователей социальных сетей / А. А. Куликова // Восточно-европейский журнал передовых технологий. — 2011. — Т. 3, № 2. — С. 14—18.
5. Губанов Д. А. Концептуальный подход к анализу онлайновых социальных сетей / А. А. Куликова // Восточно-европейский журнал передовых технологий. — 2011. — Т. 3, № 2. — С. 14—18.
6. Губанов Д. А. Социальные сети: модели информационного влияния, управления и противоборства / Д. А. Губанов, Д. А. Новиков. — М.: изд-во физико-математической литературы, 2010. — 228 с.

Концевой Антон Олександрович, студент групи 2СІ-14б, Факультет комп'ютерних систем та автоматики

Науковий керівник: Бісікало Олег Володимирович, д.т.н., професор, декан Факультету комп'ютерних систем та автоматики, Вінницький національний технічний університет, м. Вінниця

Kontsevoy Anton Olexandrovich, group 2CI-14b, Faculty of Computer Systems and Automation

Scientific supervisor: Oleg Bisikalo, Doctor of Technical Sciences, Professor, Dean of the Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsya