УДК 004.8

І.В. Стовпченко, М.М. Колотило

МЕТОДИ WEB MINING ДЛЯ ДОСЛІДЖЕННЯ СОЦІАЛЬНИХ ГРУП

Анотація. В роботі проводиться аналіз методів Web Mining, пропонується структура веб додатку для аналізу та візуалізації соціальних груп в мережі Twitter.

Ключові слова: web mining, соціальні групи.

Вступ

Необхідність автоматичного аналізу інформації з Інтернету викликана її високою доступністю великої кількості, яка постійно поповнюється інформацією, а також зростаючою популярністю вебпослуг серед усіх категорій користувачів. Розвиток мережі Інтернет в глобальну інформаційну структуру дозволило звичайним користувачам бути не тільки споживачами інформації, але й ще її творцями та розповсюджувачами. Тому для більш ефективного вирішення завдань пошуку, аналізу та структурування інформації в мережі, яка в основному є хаотично організованою, створений новий напрямок в методології аналізу даних - Web Mining.

Цей напрямок розвивається на перетині таких дисциплін як виявлення знань в базах даних, ефективний пошук інформації, штучний інтелект, машинне навчання та обробка природних мов.

Завдання Web Mining

Пошук інформації. Для знаходження необхідної інформації користувачі зазвичай користуються пошуковими ресурсами. При цьому вони часто використовують прості запити за ключовими словами. Результатом виконання запиту є список сторінок, відсортованого за індексом релевантності, що описує ступінь збігу результату із запитом, але існуючі пошукові механізми мають недоліки. Основним з них є низька точність результату, викликана недостатнім урахуванням семантичних зв'язків і контексту знайдених в тексті виразів. Індексація цікавих сегментів мережі з використанням інтелектуального аналізу даних, що застосовує алгоритми

математичної лінгвістики та обробки природних мов, є перспективним напрямком в області Web Mining.

Аналіз структури сегмента мережі. Цей метод полягає в аналізі структури посилань між різними веб-сторінками, внутрішніми та зовнішніми сайтами в виділеному мережевому сегменті.

Виявлення знань з веб-ресурсів. Це завдання перетинається з уже описаною проблемою пошуку інформації. Тільки тут у дослідника вже є набір веб-сторінок, отриманих в результаті запиту. Далі треба провести їх обробку з точки зору автоматичної класифікації, складання змістів, виявлення ключових слів і загальних тем. Виявлені знання можуть представлятися у вигляді дерев або графів, що описують структури документів або у вигляді логічних і семантичних виразів. Рішення частини цих проблем пропонує Text Mining - технологія автоматичного вилучення знань у великих обсягах текстового матеріалу, що заснована на поєднанні лінгвістичних, семантичних, статистичних та машинних методик.

Пошук шаблонів в поведінці користувачів. Метою є пошук закономірностей в шаблонах взаємодії користувача з веб-ресурсом з метою прогнозування його наступних дій. Аналізовані дії можуть включати не тільки переходи по посиланнях, але і відправку форм, прокрутку сторінок, додавання в обрані сторінки тощо. Знайдені шаблони використовуються в подальшому для оптимізації структури сайту, вивчення цільової аудиторії та для прямого маркетингу.

Дослідження соціальних груп

Social Mining – застосування методів і алгоритмів Data Mining для пошуку і виявлення залежностей і знань в соціальних мережах. Найбільш часто використовуваний засіб для аналізу і візуалізації в даній області - це граф, де вузлами є люди або групи, а дуги показують взаємини (зв'язки) або потоки інформації між вузлами.

Соціальну мережу можна уявити як «велику систему», яка має свої властивості. Як єдине ціле вона здатна взаємодіяти з навколишнім середовищем і реагувати на зовнішні процеси, що відбуваються (рисунок 1). З іншого ж боку, мережа складається з окремих елементів, їх зв'язків, властивостей і взаємин, що функціонують відповідно до певних закономірностей.

Рисунок 1 - Взаємодія соціальної мережі с навколишньою середою

Усередині соціальної мережі утворюються різні групи і спільноти за інтересами, наприклад, любителів музики, автомобілів, навчання, роботи. Зв'язки між учасниками таких об'єднань досить сильні, що дозволяє їх легко ідентифікувати. Розглянемо рисунок 2, де зображений приклад описаної ситуації. Усередині групи між учасниками зв'язків більше і вони сильніше, ніж з іншими членами соціальної мережі. У складі спільнот можуть з'являтися підгрупи, таким чином утворюючи ієрархію (на рисунку 2 таке можна спостерігати в групі 1).

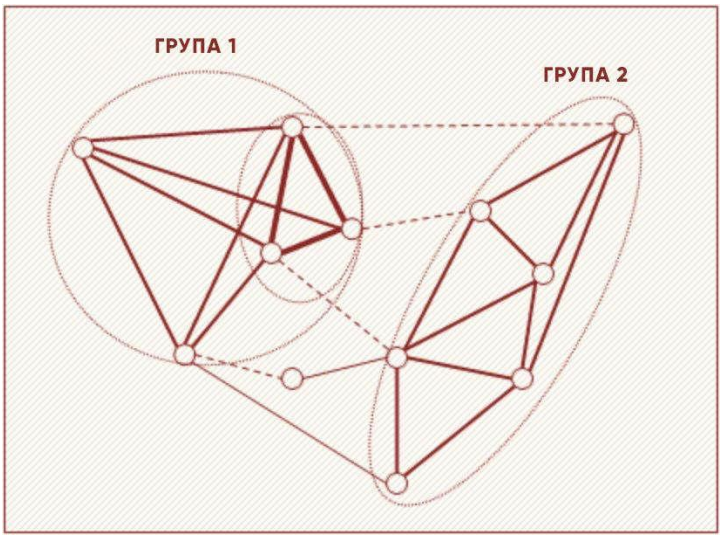

Рисунок 2 - Групи всередині соціальної мережі

Аналіз соціальної мережі Twitter

В рамках роботи створюється система для аналізу та візуалізації даних соціальних груп, яка застосовується до соціальної мережі Twitter.

Twitter - досить популярна соціальна мережа, деякий мікроблог, де люди пишуть свої думки з приводу подій у зовнішньому світі та коментують один одного. Застосувавши методи

Web Mining для цього мікроблогу, можна, наприклад, дослідити реакції жителів конкретних регіонів на конкретні події в світі, кластеризувати їх, побудувати граф взаємовідносин між ними тощо. На рисунку 3 продемонстровано загальну структуру та процес роботи додатку для обробки даних соціальної мережі Twitter. Можна побачити що дані із ВЕБ отримуються модулем отримання даних (у нашому випадку це - скрипт, що звертається до публічного API Twitter'у кожні 15 хвилин тому, що Twitter лімітує можливу кількість запитів до свого API за певний період часу, та записує дані у базу даних), обробляються модулем обробки даних (скрипт аналізує зв'язкі між зібраними користувачами, будує лінії "дружби" та записує у файл в спеціальному форматі) та на даному етапі візуалізуються за допомогою додатку Gephi.

Рисунок 3 - Схема роботи додатку

На рисунку 4 зображений соціальний граф після первинної візуалізації.

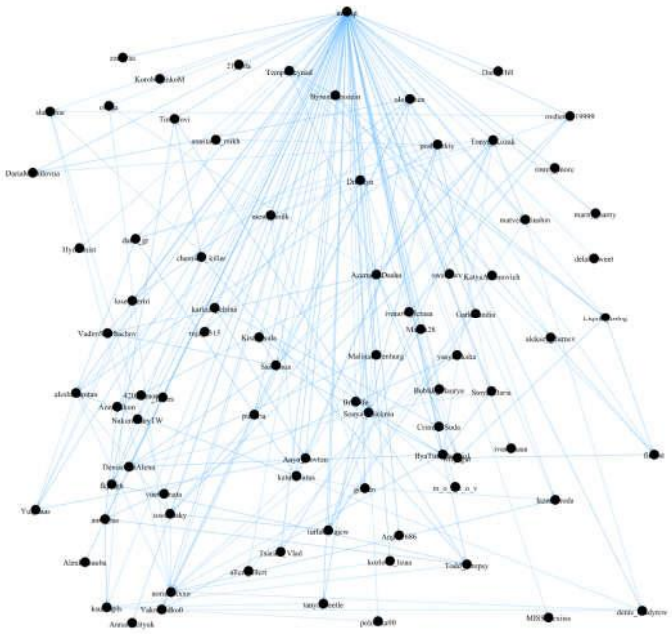

Рисунок 4 - Мій соціальний граф мережі Twitter

 ISSN 1562-9945

Таким чином ми отримуємо інструмент для аналізу та дослідження соціальних груп у режимі онлайн.

Висновки

Web Mining є новим перспективним напрямком аналізу інтернет-ресурсів для оптимізації структури веб-сайтів, отримання знань про відвідувачів сайту, опису соціальних мереж і спільнот, а також для автоматичного пошуку і структуризації інформації з інтернету.

В подальшому на основі поточної роботи планується створити веб-ресурс що дає змогу проводити аналіз соціальних груп мережі Twitter.

ЛІТЕРАТУРА

1. Градосельская Г. В. Сетевые измерения в социологии: Учебное пособие. М.: Изд. Дом «Новый учебник», 2004

2. Давыдов А. А. Системная социология: Social Networks Mining. М.: ИС РАН, 2009.

3. Айвазян С. А., Бухштабер В. М., Юнюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. - М.: Финансы и статистика, 1989.

4. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery" - Tandem Computers Inc., 1996.

5. Дюк В.А. Обработка данных на ПК в примерах. - СПб: Питер, 1997.