УДК 004.9:519.2 DOI: 10.30748/zhups.2018.56.16

О.Ю. Чередніченко[1], В.В. Ткаченко[2], М.А. Вовк[1], О.О. Масихнович[1]

[1] *Національний технічний університет «Харківський політехнічний інститут», Харків*
[2] *НДЦ ЗС України «Державний океанаріум», Одеса*

ДОСЛІДЖЕННЯ ПРОФІЛІВ КОРИСТУВАЧІВ СОЦІАЛЬНИХ МЕРЕЖ

Стаття присвячена питанням дослідження профілів користувачів соціальних мереж. Проведено огляд існуючих проблем збору та обробки даних з соціальних мереж. Визначено основні проблеми при автоматичному зборі даних з профілів Інтернет користувачів: конфіденційність даних, слабка структурованість даних, обмеження доступу і блокування, розмірність даних. Проведено огляд методів кластеризації даних, отриманих з соціальних мереж. Проаналізовано можливості використання даних моніторингу соціальних мереж для прийняття рішень. Проведено аналіз існуючих програмних рішень та виявлено,що більшість програмних продуктів дозволяє лише побудувати граф даних з соціальної мережі. Програмне забезпечення, яке б дозволило проводити аналіз профілів користувачів соціальних мереж, відсутнє на ринку. У статті розглянуто проектування такої програмної системи та розроблено вимоги до неї.

Ключові слова: *профіль користувача, соціальні мережі, збір даних, методи кластеризації, програмні продукти.*

Вступ

Постановка проблеми. Активне зростання аудиторії соціальних медіа в мережі Інтернет призвело до становлення цих ресурсів як нового джерела даних для прийняття рішень. Специфіка роботи з такими даними несе в собі цілий ряд переваг і недоліків. До переваг належить швидкість доступу до інформації, охоплення аудиторії. Однією з головних переваг, як і серйозною перешкодою, є обсяг цих даних. Так, згідно зі статистикою [1], щомісяця в соціальних мережах близько 30 мільйонів унікальних авторів публікують майже 580 мільярдів повідомлень.

Аналіз соціальних даних стрімко набирає популярність у всьому світі завдяки появі в 1990-х роках онлайнових сервісів соціальних мереж [1]. З цим пов'язаний феномен соціалізації персональних даних: стали публічно доступними факти біографії, листування, щоденники, фото, відео, аудіоматеріали, замітки про подорожі тощо. Таким чином, соціальні мережі є унікальним джерелом даних про особисте життя і інтереси реальних людей [2]. Це відкриває безпрецедентні можливості для вирішення дослідницьких і бізнес-задач, а також створення допоміжних сервісів і додатків для користувачів соціальних мереж.

Актуальність даної роботи обумовлена необхідністю розвитку методологічного апарату, який дозволив би використовувати великі обсяги даних користувачів соціальних мереж для вирішення комплексу задач по прийняттю рішень. 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 бізнес-додатків і сервісів. Разом з тим, при роботі з соціальними даними потрібно брати до уваги такі фактори, як нестабільність якості користувацького контенту (спам і неправдиві акаунти), проблеми із забезпеченням приватності особистих даних користувачів при зберіганні і обробці, а також часті оновлення користувальницької моделі і функціоналу. Все це вимагає постійного вдосконалення алгоритмів розв'язання різних аналітичних і бізнес-задач.

Аналіз останніх досліджень і публікацій. Дослідницька компанія Statista [3] опублікувала результати звіту, які показують, що кількість унікальних користувачів соціальних мереж буде продовжувати стрімко зростати (рис. 1).

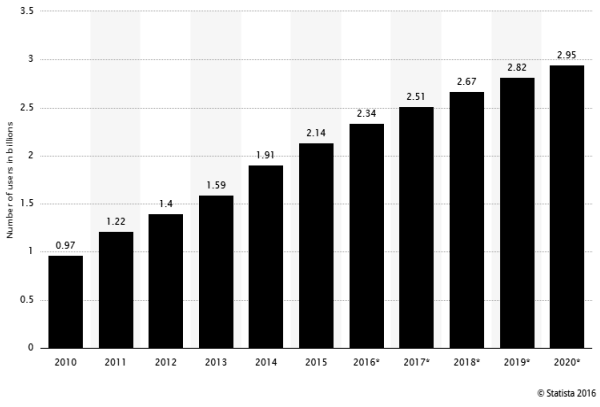

Рис. 1. Кількість користувачів соціальних мереж в 2010–2020 р.р. [3]

Згідно з даними прогнозу [4] кількість користувачів соціальних мереж в 2018 році має досягти більш ніж 2,5 мільярди. У ряді країн (рис. 2) кількість користувачів соціальних мереж або досягає

половину від загального числа користувачів мережі Інтернет, або перевищує це значення.

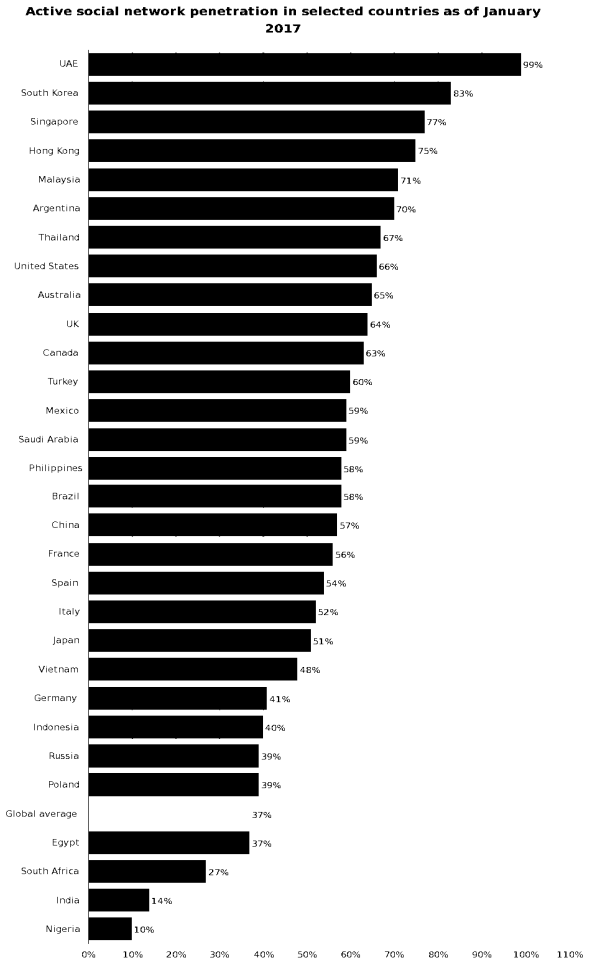

Рис. 2. Поширення соціальних мереж,
% від загального числа Інтернет користувачів [4]

Дослідницька компанія Alexa представила звіт про поширення соціальних мереж по країнам [4]. За даними, що наведено у звіті соціальна мережа Facebook поширена в більшості країн. Хоча кількість користувачів Facebook на території країн Східної Європи зростає, найбільшу частку популярності тут займають соціальні мережі «Odnoklassniki» і «VKontakte».

Аналітичне агентство Gartner у 2012 році опублікувало звіт [5]. Згідно зі звітом, технологіям "Соціальна аналітика" (Social analytics) і "Великі дані" (Big data) в даний час приділяється велика увага. Зокрема, дослідженнями соціальних даних активно займаються такі університети, як Стенфорд та Оксфорд, а також компанії Facebook, Google, Yahoo!, LinkedIn і багато інших. Компанії-власники сервісів соціальних мереж активно інвестують в розробку вдосконалених інфраструктурних і алгоритмічних рішень для обробки великих масивів даних.

Обробка соціальних даних вимагає також розробки відповідних алгоритмічних і інфраструктурних рішень, що дозволяють враховувати їх розмірність. 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 в веб-браузері або для використання даних користувачів спеціалізованими додатками. Оскільки сценарії використання інтерфейсів соціальних мереж не передбачають автоматичного збору даних для аналітичних цілей, то виникає ряд проблем [5].

1. Конфіденційність даних – часто доступ до даних користувачів дозволений тільки для зареєстрованих і авторизованих учасників мережі, що вимагає підтримки емуляції користувальницької сесії за допомогою спеціальних облікових записів.

2. Слабка структурованість даних – у багатьох випадках програмні інтерфейси (API) соціальних мереж мають обмежений функціонал, що вимагає отримання статичних копій HTML-сторінок, коректної обробки їх динамічної частини вилучення потрібних даних за допомогою алгоритму або шаблону і побудови їх структурованого уявлення.

3. Обмеження доступу і блокування – з метою запобігання несанкціонованому автоматичного збору даних і зайвого навантаження власники сервісів часто вводять обмеження на допустиму кількість запитів від одного користувача в одиницю часу, що вимагає обліку кількості запитів, а також підтримки динамічної ротації використовуваних IP-адрес.

4. Розмірність даних – обумовлює необхідність в паралельному методі збору даних, а також в методах отримання репрезентативної вибірки користувачів соціальної мережі.

Однією з фундаментальних проблем при використанні соціальної інформації про користувача є її фрагментованість серед безлічі різних соціальних мереж. Незважаючи на те, що існують спроби щодо забезпечення єдиного способу взаємодії між різними соціальними платформами, вони не набули широкого застосування, а нові соціальні сервіси продовжують з'являтися.

Таким чином, завдання ідентифікації користувачів в різних соціальних мережах має на увазі зіставлення акаунтів користувачів в рамках списків контактів деякого центрального користувача в різних соціальних мережах. Таке завдання часто виникає при роботі з контактами користувачів в соціальних мета-сервісах, які, зокрема, можуть служити для об'єднання новинних потоків або надання єдиної системи обміну повідомленнями. Подібна задача виникає також при використанні функції автомати-

чного об'єднання контактів з різних джерел (телефонна книга, соціальні мережі, месенджери).

Метою роботи є аналіз проблем використання даних профілів користувачів соціальних мереж, існуючих методів дослідження профілів користувачів соціальних мереж та програмного забезпечення для аналізу даних соціальних мереж.

Виклад основного матеріалу

Щороку з'являється велика кількість як універсальних, так і спеціальних соціальних сервісів, і для активних користувачів Інтернет типово мати кілька профілів в різних соціальних мережах. Оскільки пошук акаунтів користувача в різних мережах в загальному випадку вимагає наявності актуальних даних про всіх користувачів даних мереж, доцільно обмежити простір пошуку найближчими сусідами будь-якого користувача, акаунти якого в досліджуваних мережах відомі.

Така задача вирішується методами кластеризації [6–10]. У багатьох прикладних задачах вимірювати ступінь подібності об'єктів істотно простіше, ніж формувати ознакові описи. Загальний алгоритм кластеризації виглядає так [9–10].

На першому етапі відбувається підготовка даних до кластеризації. Дані для кластеризації найчастіше представляють у вигляді таблиць, де кожен стовпець – це один з атрибутів, рядок – об'єкт даних.

На другому етапі вибирають, як охарактеризувати схожість об'єктів.

Для цього використовуються різні міри близькості [9]. Міра близькості підбирається індивідуально для конкретних типів даних. Іноді адекватної міри близькості підібрати не вдається.

На третьому етапі вибирають алгоритм, за яким будують модель даних, тобто алгоритм групування об'єктів.

На четвертому етапі реалізується алгоритм, результатом є угруповання об'єктів а кластерами.

На п'ятому етапі отримане групування намагаються представити в найбільш зручному для інтерпретації вигляді. Представлення результатів кластеризації призване допомогти найбільш точно інтерпретувати результати виконання алгоритму.

На останньому етапі кластеризації результати виконання алгоритму інтерпретуються, з них виходять знання, тобто корисні правила, які можна використовувати в подальшому для віднесення нових об'єктів до тієї чи іншої групи – кластеру.

Необхідним інструментарієм для дослідження виступають додаткові програмні продукти, що дозволяють маніпулювати потоками даних і організовувати їх максимально зручним для подальшого аналізу. Програмне забезпечення для аналізу даних соціальних мереж зазвичай використовує теорію графів для дослідження соціальних структур, як аналітично, так і візуально. В табл. 1 наведено програмні продукти, які можна безкоштовно використовувати задля некомерційних цілей.

Таблиця 1

Програмно-інформаційний інструментарій для дослідження комунікацій в соціальних мережах

Найменування	Функціональні можливості інструментарію
AlchemyAPI	Забезпечує технологією семантичного аналізу та отримання мета-даних з текстів, веб-документів, дозволяє визначати у великих масивах тексту ключові слова, які з великою часткою релевантності характеризують текст; реалізує екстрагування іменованих об'єктів з тексту, включаючи визначення людей, компаній, географічних об'єктів, організацій, об'єктів інфраструктури. Технологія в цілому дозволяє проводити якісний аналіз і фільтрацію текстів [11].
NVivo	Забезпечує обробку як текстових джерел, так і зображень для організації та аналізу нечислових і неструктурованих потоків даних [12].
Cytoscape	Програмне забезпечення з відкритим вихідним кодом для візуалізації складних даних і їх інтеграції з будь-яким типом атрибута даних [13].
YahooPipes	Веб-додаток з графічним інтерфейсом для створення додатків, які об'єднують стрічки новин, веб-сторінки та інші сервіси. Дозволяє користувачам отримувати потоки інформації з різних джерел і створювати правила з управління отриманим контентом, наприклад, використовуючи фільтри, сортування і переклад з інших мов об'єднання RSS потоків з різних джерел [14].
DrupalCMF	Дозволяє організувати бібліотеку структурованої інформації, отриманої з різних джерел і її подальшу обробку, включаючи пошукові запити різного ступеня складності і відображення результатів в табличному або іншому зручному вигляді [15].
Cuttlefish	Додаток, який візуалізує дані мереж з використанням деяких найбільш відомих алгоритмів компонування [16].
Gephi	Підтримує інтерактивну візуалізацію для всіх видів мереж і складних систем, динамічних і ієрархічних графів; працює на Windows, Linux і Mac OS; підтримує всі типи мереж [17].

Закінчення табл. 1

Найменування	Функціональні можливості інструментарію
libSNA	Бібліотека з відкритим вихідним кодом для соціального аналізу мереж, під ліцензією LGPL [18].
NodeXL	Дозволяє візуалізувати граф соціальної мережі, а також проводити статистичну обробку отриманих даних; отримують структуровані дані в одному зі спеціальних форматів і забезпечують візуальне, а також чисельне відображення таких ключових параметрів мережі як кластери користувачів, ранги вузлів мережі, щільність графа тощо [19].
GraphChi	Може виконувати дуже великі обчислення графу на одній машині, використовуючи новий алгоритм для обробки графу з диску (SSD або жорсткий диск); також підтримує потокові поновлення графа моделювання та видалення ребер з графа [20].
JUNG	Бібліотека програмного забезпечення, яка представляє загальну та розширювану мови для моделювання, аналізу та візуалізації даних, які можуть бути представлені у вигляді графіка або мережі [21].
Netlytic	Хмаро-базований текстовий аналізатор та візуалізатор даних соціальних мереж, який може автоматично підсумовувати великі обсяги тексту від розмов на сайтах до соціальних медіа, коментарів в блогах, форумах і чатах [22].

Після проведення аналізу існуючого програмного забезпечення можна зробити висновок, що:

1) більшість програмних продуктів дозволяє лише побудувати граф даних з соціальної мережі;

2) на ринку програмного забезпечення відсутні продукти, які дозволяють проводити аналіз профілів користувачів соціальних мереж для бізнес-цілей.

На підставі проведеного аналізу можна сказати, що проблема дослідження профілів користувачів соціальних мереж існує. На ринку є тільки продукти, що дозволяють проводити базове дослідження на основі даних соціальних мереж, в багатьох випадках тільки для побудови графіку або діаграми даних. Автоматизація в досліджуваній предметній області дає багато переваг.

Аналіз вимог до інформаційної системи дозволив виділити акторів інформаційної системи та основні варіанти використання системи. Інформація про основні варіанти використання приведена у табл. 2.

Таблиця 2

Варіанти використання системи

Назва	Опис
Логін	Авторизація користувача у системі
Реєстрація	Реєстрація нового користувача системи
Модель для аналізу соціальних даних	Відобразити параметри моделі з аналізу соціальних даних
Редагування моделі	Редагування параметрів моделі з аналізу соціальних даних
Список проектів	Відображення та редагування списку проектів користувача
Додати проект	Додати новий проект у систему
Аналіз вхідних соціальних даних	Проведення аналізу вхідних соціальних даних та формування маркетингового звіту
Звіт	Відображення звіту з виконаного аналізу соціальних даних
Довідкова інструкція	Відображення довідкової інформації по використанню системи
Список користувачів	Відображення та редагування даних користувача з переліку користувачів системи
Управління проектами системи	Перегляд та видалення існуючих проектів з системи

Візуальне моделювання з використанням нотації UML можна уявити як процес порівневого спуску від найбільш загальної і абстрактної концептуальної моделі вихідної бізнес-системи до логічної, а потім і до фізичної моделі відповідної програмної системи. На рис. 3–5 наведено основні варіанти використання системи, що розробляється для різних типів користувачів (акторів).

Для створення конкретної фізичної системи необхідно реалізувати всі елементи логічного представлення в конкретні матеріальні сутності. Для опису таких реальних сутностей призначений інший аспект модельного уявлення, а саме – фізичне представлення моделі. У мові UML для фізичного представлення моделей систем використовуються так звані діаграми реалізації, які включають в себе дві

окремі канонічні діаграми: діаграму компонентів і діаграму розгортання.

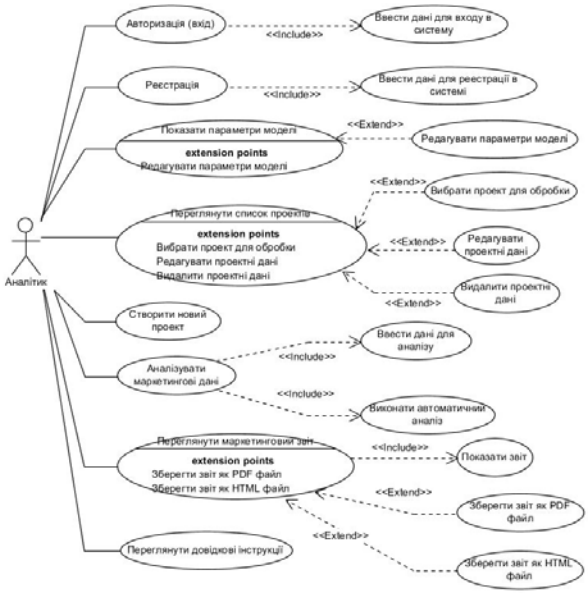

Рис. 3. UML-діаграма основних варіантів використання для аналітика системи

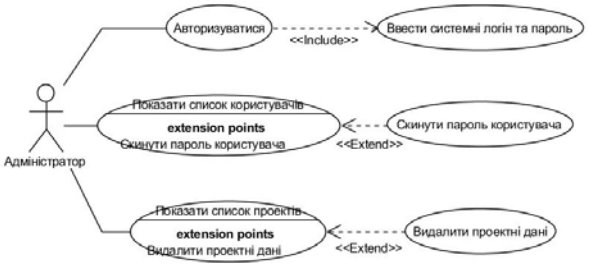

Рис. 4. UML-діаграма основних варіантів використання для адміністратора

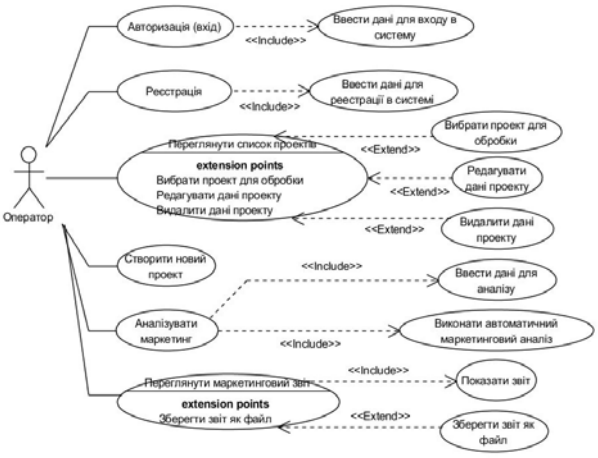

Рис. 5. UML-діаграма основних варіантів використання для оператора

Діаграма компонентів забезпечує узгоджений перехід від логічного представлення до конкретної реалізації проекту у формі програмного коду. Одні компоненти можуть існувати тільки на етапі компіляції програмного коду, інші – на етапі його виконання. Діаграма компонентів відображає загальні залежності між компонентами (рис. 6). Діаграма розміщення компонентів системи наведена на рис. 7.

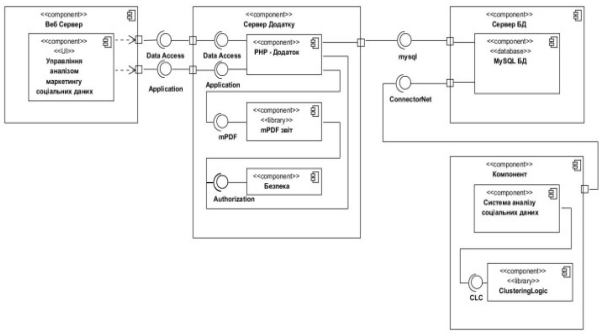

Рис. 6. UML-діаграма основних програмних компонентів системи

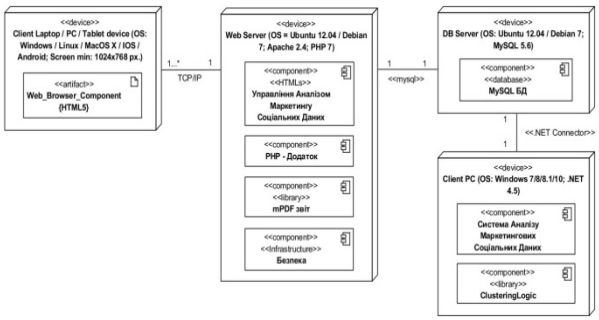

Рис. 7. UML-діаграма розміщення

Висновки

Ідентифікація користувача в різних соціальних мережах дає змогу отримати більш повну картину про соціальну поведінку даного користувача в мережі Інтернет. Виявлення акаунтів, що належать одній людині, в декількох соціальних мережах, дозволяє отримати більш повний соціальний граф, що може бути корисно в багатьох задачах, таких, як інформаційний пошук, інтернет-реклама, рекомендаційні системи і т.п.

Визначено основні проблеми при автоматичному зборі даних з профілів інтернет користувачів: конфіденційність даних, слабка структурованість даних, обмеження доступу і блокування, розмірність даних. Аналіз існуючих програмних рішень довів відсутність програмних продуктів, здатних проводити аналіз профілів користувачів соціальних мереж. Проектування такої програмної системи планується та розроблено вимоги до неї. Проведений аналіз проблем використання даних профілів користувачів соціальних мереж та існуючого програмного забезпечення дозволив сформулювати вимоги та побудувати архітектуру програмної системи, реалізація якої дозволить здійснювати збір та обробку даних в соціальних мережах для подальшого використання в системах підтримки прийняття рішень.

Список літератури

1. Najork M. Breadth-first crawling yields high-quality pages / M. Najork, J.L. Wiener // Proceedings of the 10th international conference on World Wide Web. – ACM, 2011. – P. 114-118.
2. Leskovec J. Sampling from large graphs / J. Leskovec, C. Faloutsos // Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – P. 631-636.
3. Number of social media users worldwide from 2010 to 2020 //https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.
4. World map of social networks research by Alexa and SimilarWeb //http://vincos.it/world-map-of-social-networks/.
5. Key Trends to Watch in Gartner 2012 Emerging Technologies Hype Cycle // http://www.forbes.com/sites/gartnergroup/2012/09/18/key-trends-to-watch-in-gartner-2012-emerging-technologies-hype-cycle-2/
6. Будильский Д.В. Дослідження застосовності моделей глибокого навчання для завдання аналізу текстових повідомлень / Д.В. Будильский, А.Г. Подвесовский. – Брянськ, 2015. – С. 117-126.
7. Мандель І.Д. Кластерний аналіз / І.Д. Мандель. – М.: Фінанси і Статистика, 2008.
8. Уілліамс У.Т. Методи ієрархічної класифікації / У.Т. Уілліамс, Д.Н. Ланс; під ред. М.Б. Малюта. – М.: Наука, 2006. – С. 269-301.
9. Jain A. Data clustering: A review / A. Jain, M. Murty, P. Flynn // ACM Computing Surveys. – 2003. – Vol. 31, No. 3. – P. 264-323.
10. Lance G.N. A general theory of classification sorting strategies. Hierarchical systems / G.N. Lance, W.T. Willams // Comp. J. – No. 9. – Pp. 373-380.
11. Semantic Text Analysis APIs Using Natural Language Processing // http://www.alchemyapi.com/, 15.11.2016.
12. QSR International: NVivo qualitative data analysis software // http://www.qsrinternational.com/nvivo-product/, 16.11.2016.
13. Cytoscape: An Open Source Platform for Complex Network Analysis // http://www.cytoscape.org/, 15.11.2016.
14. Yahoo Pipes // http://pipes.yahoo.com/, 16.11.2016.
15. Content Management Filter // https://www.drupal.org/project/cmf, 16.11.2016.
16. Cuttlefish software tool // http://cuttlefish.sourceforge.net/, 16.11.2016.
17. Gephi software tool // https://gephi.github.io/, 17.11.2016.
18. libSNA. Open-source library for Social Network Analysis //http://www.libsna.org/, 17.11.2016.
19. NodeXL: Network Overview, Discovery and Exploration for Excel – Home // https://nodexl.codeplex.com/, 17.11.2016.
20. GraphChi software tool // https://github.com/GraphChi/graphchi-cpp, 18.11.2016.
21. Java Universal Network / Graph Framework software tool // http://jung.sourceforge.net/, 18.11.2016.
22. Netlyticsoftware tool. Official website //https://netlytic.org/home/, 19.11.2016.

References

1. Najork, M. and Wiener, J.L. (2011), Breadth-first crawling yields high-quality pages, *Proceedings of the 10th international conference on World Wide Web*, ACM, pp. 114-118.
2. Leskovec, J. and Faloutsos, C. (2006), Sampling from large graphs, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 631-636.
3. Number of social media users worldwide from 2010 to 2020 //https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.
4. World map of social networks research by Alexa and SimilarWeb //http://vincos.it/world-map-of-social-networks/.
5. Key Trends to Watch in Gartner 2012 Emerging Technologies Hype Cycle // http://www.forbes.com/sites/gartnergroup/2012/09/18/key-trends-to-watch-in-gartner-2012-emerging-technologies-hype-cycle-2/
6. Budilskiy, D.V. and Podvesovskiy, A.G. (2015), "*DoslIdzhennya zastosovnostI modeley glibokogo navchannya dlya zavdannya analizu tekstovih povidomlen*" [*The study of the applicability of the models of deep learning for the purpose of analyzing text messages*], Bryansk, pp. 117-126.
7. Mandel, I.D. (2008), "*Klasternyi analiz*" [*Cluster analysis*], Finance and Statistics, Moscow, 215 p.
8. Uilliams, U.T. and Lans, D.N. (2006), "*Metody Iierarkhichnoi Klasyfikatsii*" [*Methods of hierarchical classification*], Moscow, 301 p.
9. Jain, A., Murty, M. and Flynn, P. (2003), Data clustering: A review, *ACM Computing Surveys*, Vol. 31, No. 3, pp. 264-323.
10. Lance, G.N. and Willams, W.T. A general theory of classification sorting strategies. Hierarchical systems, *Comp. J*, No. 9, pp. 373-380.
11. Semantic Text Analysis APIs Using Natural Language Processing // http://www.alchemyapi.com/, 15.11.2016.
12. QSR International: NVivo qualitative data analysis software // http://www.qsrinternational.com/nvivo-product/, 16.11.2016.
13. Cytoscape: An Open Source Platform for Complex Network Analysis // http://www.cytoscape.org/, 15.11.2016.
14. Yahoo Pipes // http://pipes.yahoo.com/, 16.11.2016.
15. Content Management Filter // https://www.drupal.org/project/cmf, 16.11.2016.
16. Cuttlefish software tool // http://cuttlefish.sourceforge.net/, 16.11.2016.
17. Gephi software tool // https://gephi.github.io/, 17.11.2016.
18. libSNA. Open-source library for Social Network Analysis //http://www.libsna.org/, 17.11.2016.
19. NodeXL: Network Overview, Discovery and Exploration for Excel – Home // https://nodexl.codeplex.com/, 17.11.2016.
20. GraphChi software tool // https://github.com/GraphChi/graphchi-cpp, 18.11.2016.
21. Java Universal Network / Graph Framework software tool // http://jung.sourceforge.net/, 18.11.2016.
22. Netlyticsoftware tool. Official website //https://netlytic.org/home/, 19.11.2016.

Надійшла до редколегії 2.04.2018
Схвалена до друку 15.05.2018

Відомості про авторів:

Чередніченко Ольга Юріївна
кандидат технічних наук доцент
доцент кафедри «Національного технічного університету «Харківський політехнічний інститут»,
Харків, Україна
https://orcid.org/0000-0002-9391-5220

Ткаченко Василь Володимирович
Начальник науково-дослідного відділу
НДЦ ЗС України «Державний океанаріум»
Одеса, Україна
https://orcid.org/0000-0003-2944-8987

Вовк Марина Анатоліївна
кандидат економічних наук
доцент кафедри програмної інженерії та інформаційних технологій управління Національного технічного університету «Харківський політехнічний інститут»,
Харків, Україна
https://orcid.org/0000-0003-4119-5441

Масихнович Олександр Олександрович
студент кафедри програмної інженерії та інформаційних технологій управління Національного технічного університету «Харківський політехнічний інститут»
Харків, Україна

Information about the authors:

Olga Cherednichenko
Candidate of Sciences Associate Professor
Senior Lecturer of Department of National Technical University "Kharkiv Polytechnic Institute",
Kharkiv, Ukraine
https://orcid.org/0000-0002-9391-5220

Vasyl Tkachenko
Head of Research Department in Research Center of the Armed Forces of Ukraine "State Oceanarium",
Odesa, Ukraine
https://orcid.org/0000-0003-2944-8987

Maryna Vovk
Candidate of Sciences
Senior Lecturer of Software Engineering and Information Technologies Management Department of National Technical University "Kharkiv Polytechnic Institute",
Kharkiv, Ukraine
https://orcid.org/0000-0003-4119-5441

Oleksandr Masyhnovych
Student of Software Engineering and Information Technologies Management Department of National Technical University "Kharkiv Polytechnic Institute",
Kharkiv, Ukraine

ИССЛЕДОВАНИЕ ПРОФИЛЕЙ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ

О.Ю. Чередниченко, В.В. Ткаченко, М.А. Вовк, О.О. Масихнович

Статья посвящена вопросам исследования профилей пользователей социальных сетей. Проведен обзор существующих проблем сбора и обработки данных из социальных сетей. Определены основные проблемы при автоматическом сборе данных из профилей интернет пользователей: конфиденциальность данных, слабая структурированность данных, ограничения доступа и блокировки, размерность данных. Проведен обзор методов кластеризации данных, полученных из социальных сетей. Проанализированы возможности использования данных мониторинга социальных сетей для принятия решений. Проведен анализ существующих программных решений и выявлено, что большинство программных продуктов позволяет только построить граф данных из социальной сети. Программное обеспечение, которое бы позволило проводить анализ профилей пользователей социальных сетей, отсутствует на рынке. В статье рассмотрено проектирования такой программной системы и разработаны требования к ней.

Ключевые слова: *профиль, социальные сети, сбор данных, методы кластеризации, программные продукты.*

STUDY OF SOCIAL NETWORKS USER'S PROFILES

O. Cherednichenko, V. Tkachenko, M. Vovk, O. Masihnovich

The article is devoted to studying of social networks user's profiles. An overview of the current problems in collecting and processing data from social networks is completed. The main problems in the automatic collection of data from profiles of Internet users are identified. They are confidentiality of data, poor data structuring, access restriction and blocking, the dimension of data. Often access to user data is only allowed for registered and authorized network members, which requires support for user session emulation using special accounts. In many cases, the social interface APIs have a limited functionality that requires support for receiving static copies of HTML pages with the user interface. In order to prevent unauthorized automatic data gathering and limiting the load on the infrastructure of a social network service, service owners often introduce explicit or concealed restrictions on the permissible number of requests from one user account and/or IP address per unit time counting the number of queries referenced. The dimensionality of the data necessitates a parallel method of data collection, as well as methods for obtaining a representative sample of users of the social network (sampling). An overview of the methods of data clustering obtained from social networks has been reviewed. The possibilities of using social network monitoring data for decision making have been analyzed. An analysis of existing software solutions was made and it was found that most software products only allow constructing of a graph of data from a social network. Software that allows analyzing of user profiles of social networks is not available on the market. The article deals with the design of such a software system and the requirements for it were developed.

Keywords: *profile, social networks, data collection, clustering methods, software products.*