Огляд Видів Аналізу Соціальних Мереж для Забезпечення Інформаційної Безпеки

Тамара Радівілова
кафедра інфокомунікаційної інженерії
Харківський національний університет радіоелектроніки
Харків, Україна
tamara.radivilova@gmail.com

Людмила Кіріченко
кафедра прикладної математики
Харківський національний університет радіоелектроніки
Харків, Україна
lyudmyla.kirichenko@nure.ua

Даніїл Рудченко
кафедра інфокомунікаційної інженерії
Харківський національний університет радіоелектроніки
Харків, Україна
daniil.rudchenko@nure.ua

Overview Types of Analysis of Social Networks for Information Security

Tamara Radivilova
Department of Infocommunication Engineering
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
tamara.radivilova@gmail.com

Lyudmyla Kirichenko
Department of Applied Mathematics
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
lyudmyla.kirichenko@nure.ua

Daniil Rudchenko
Department of Infocommunication Engineering
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
daniil.rudchenko@nure.ua

***Анотація*—У даній статті представлений огляд основних видів аналізу даних в соціальних мережах для забезпечення інформаційної безпеки. Розглянуто основні математичні методи аналізу даних, які використовуються в кожному з вказаних видів аналізу. Для кожного з видів аналізу вказано дані, які можуть бути проаналізовані, дані, які будуть отримані після аналізу, та їх область та доцільність використання. Представлено типові завдання аналізу соціальних мереж, такі як кластерізація, прогнозування, глибинний аналіз текстів, моніторинг інформаційного простору, визначення взаємозв'язків користувачів та спільнот, класифікація даних, виявлення спільнот та експертів в мережі, виявлення лідерів в спільнотах, аналіз потоків подій, аналіз та контент аналіз та інші.**

***Abstract*—The paper provides an overview of the main types of data analysis in social networks for information security. The basic mathematical methods for analyzing the data used in each of these types of analysis are considered. The main types of social network analysis are hypothesis testing, data operations, sorting, comparison of elements, analysis of photographic and video images, deep analysis of texts, monitoring of the information space, evaluation of user involvement, defining relationships, evaluation of coverage, demographic and geographical classification, analysis tonality, the evaluation of transactions on**

Internet, event stream analysis, prediction. For each type of analysis, the data that can be analyzed, as well as the methods that are most often used to analyze this data, are indicated. Also for each type of analysis, the data to be obtained as a result of the analysis, and their scope and usefulness are indicated. It is shown that main important methods of analysis are graph theory methods and comtlex networks methods, text mining and data mining methods. Typical tasks of analyzing social networks, such as clustering, prediction, determining users relationships and communities, classifying data, identifying communities and experts on the network, identifying leaders in communities, event stream analysis, statistical analysis and content analysis and others are presented.

Ключові слова—аналіз соціальних мереж, методи аналізу даних, передбачення, моніторинг інформаційного простору.

Keywords—social network analysis, data analysis methods, prediction methods, prediction, monitoring of information space.

I. Вступ

У сучасному суспільстві Інтернет і соціальні мережі відіграють дуже важливу роль. Вони дозволяють людям всієї планети спілкуватися з друзями, шукати роботу, займатися просуванням власного бізнесу або організацією різних заходів. Але соціальні медіа - Facebook, Twitter або відеохостінговий сайт YouTube - все в більшій мірі стають інструментами впливу звичайних людей, їх користувачів, на міжнародну політику, соціальні процеси.

Виділяють наступні основні типи кінцевих користувачів, зацікавлених в аналізі, прогнозуванні та управлінні соціальними мережами: 1) органи державної влади та місцевого самоврядування; 2) підприємства державного та приватного сектора економіки, в тому числі комерційні організації; дослідні організації; засоби масової інформації; силові структури; 3) суспільство, в тому числі політичні партії, науково-освітні організації; окремі фізичні особи. Аналіз соціальних мереж, зокрема, використовується в розвідувальних, контррозвідувальних та правоохоронних заходах. Таким чином аналіз інформації та поведінки людей в комп'ютерних соціальних мережах викликає великий інтерес у сучасних дослідників та інших зацікавлених організацій та осіб [1-3].

II. Аналіз інформації соціальних мереж

Тема аналізу інформації дуже велика, і в даній роботі представлена частина видів можливого аналізу інформації соціальних мереж. Мета аналізу - виявлення патернів подій: взаємозв'язків між людьми, закономірностей, за якими відбуваються ці події та передбачення ходу аналогічних подій в подальшому. В ході аналізу соціальних мереж використовують безліч математичних методів. При цьому можуть використовуватися методи багатовимірного статистичного аналізу і штучного інтелекту (в тому числі методи data mining, text mining, image / video mining), а також методи аналізу мережевих структур.

Нижче наведені найпростіші види аналізу даних.

1. Тестування гіпотез. Зіставляючи факти, непрямі ознаки, логічні висновки з фактів можна визначити найбільш ймовірну гіпотезу. Побудова гіпотез (як і будь-якої версії) - це розумовий процес переходу від неповних ймовірних знань до повних і достовірних знань. Висування гіпотези включає всебічне вивчення явищ, що спостерігаються, аналіз і відбір фактів, які перебувають в причинно-наслідковом часовом зв'язку з зазначеними обставинами, аналіз окремих фактів і відносин між ними. Висування гіпотези (версії) складається з наступних кроків: аналіз окремих фактів і відносин між ними; синтез фактів, їх узагальнення; формулювання припущення. Перевірка гіпотези (версії) - цілеспрямований збір доказів, які підтверджують або спростовують ці припущення. Існує кілька способів підтвердження істинності гіпотези (версії):

а) виведення гіпотез, що випливають з наслідків та їх перевірка - встановлення відповідності фактичних даних;

б) безпосереднє виявлення об'єкта, думка про існування якого була основним змістом гіпотези;

в) дедуктивне виведення гіпотези з іншого, але достовірного, знання;

г) підтвердження підстави гіпотези, якщо при побудові гіпотези вона не була достовірною;

д) розширення підстави гіпотези до меж, достатніх для достовірного знання, і т. п.

Спростування гіпотези здійснюється шляхом встановлення фальсифікації наслідків, що випливають з неї, тобто шляхом встановлення їх невідповідності фактичним даним.

2. Операції з даними: сортування, зіставлення елементів і т.п. Можна визначити спільноти, структури, до яких може мати відношення об'єкт, що цікавить. Часто використовують методи кластеризації, основними з яких є наступні [4]:

а) Алгоритми, засновані на поділі даних (Partitioning algorithms), в т.ч. ітеративні: поділ об'єктів на k кластерів; ітеративний перерозподіл об'єктів для поліпшення кластеризації.
б) Ієрархічні алгоритми (Hierarchy algorithms): агломерація: кожен об'єкт спочатку є кластером, кластери, з'єднуючись один з одним, формують більший кластер і т.д.
в) Методи, засновані на концентрації об'єктів (Density-based methods): засновані на можливості з'єднання об'єктів, ігнорують шуми, знаходження кластерів довільної форми.
г) Грід-методи (Grid-based methods): квантування об'єктів в грід-структури.
д) Модельні методи (Model-based): використання моделі для знаходження кластерів, найбільш відповідних даним.
є) Графічні методи (відображення графової структури взаємодії / знайомств об'єкта, груп і т.д).

3. Аналіз фото- і відео-зображень. Використовується для виконання таких завдань: аналіз даних, що випадково потрапили в кадр, як вид з вікна, частина назви географічного пункту або відображення в дзеркалі на задньому плані; виявлення «підозрілих» предметів; ідентифікація осіб, які перебувають «в розшуку»; авторизація доступу за відбитками пальців; записи

відеоспостережень; відбитки пальців, знімки сітківки ока; зображення осіб; невербальні сигнали, які дозволять судити про людину (складки на обличчі, що відкривають переважну емоцію людини; улюблені жести; характер взаємовідносин з іншими людьми і т.п.).

Найбільш використовуваними є наступні методи і алгоритми [5]: методи визначення біометричних характеристик, метод головних компонент (зменшує розмірність даних), інтелектуальні методи розпізнавання (віднесення вихідних даних до певного класу за допомогою виділення істотних ознак, що характеризують ці дані, із загальної маси несуттєвих даних), метод еластичних графів (особа представляється у вигляді графа, вершини якого розташовані на ключових точках особи - вузлах, таких як контури голови, губ, носи та їх крайніх точках, кожна грань позначена відстанями між її вершинами), методи нейронних мереж (система з'єднаних і взаємодіючих між собою простих процесорів), генетичні алгоритми (евристичний алгоритм пошуку, який використовується для вирішення завдань оптимізації та моделювання шляхом випадкового підбору, комбінування і варіації шуканих параметрів з використанням механізмів, аналогічних природному відбору в природі), фрактальний аналіз (обчислення фрактальних характеристик для всього зображення, для окремих фрагментів або об'єктів в полі сканування) та інші.

В даний час існує чотири основні методи розпізнавання особи: eigenfaces (використовує двовимірні зображення в градаціях сірого, які представляють відмінні характеристики зображення особи, часто використовується в якості основи для інших методів розпізнавання особи), аналіз «відмінних рис» (подібна методиці «Eigenface», але більшою мірою адаптована до зміни зовнішності або міміки людини), аналіз на основі нейронних мереж (використовують алгоритм, який встановлює відповідність унікальних властивостей людини, що перевіряється, і параметрів шаблону, що знаходиться в базі даних, при цьому застосовується максимально можливе число параметрів), метод «автоматичної обробки зображення обличчя» (використовує відстані та ставлення відстаней між легко визначними точками особи, такими як очі, кінець носа, куточки рота).

4. Глибинний аналіз текстів. Автоматичне вилучення понять і фактів і формалізованих масивів інформації (баз даних, таблиць) і неструктурованих текстів, представлених в веб-просторі, виявлення глибинних зв'язків між поняттями. Для вирішення цих завдань використовуються технології Knowledge discovery, концепції глибинного аналізу даних і текстів (Data Mining, Text Mining) [1, 3-4]. Завдання Text Mining: витяг з тексту його характерних елементів або властивостей, які можуть використовуватися в якості метаданих документа, ключових слів, анотацій; віднесення документа до деяких категорій з заздалегідь заданої схеми класифікації; семантичний пошук документів. До основних методів Text Mining відносять класифікацію (Classification, завдання розпізнавання, де за деякою контрольною вибіркою система відносить новий об'єкт до якоїсь категорії), кластеризації (Clustering, процес виділення компактних підгруп об'єктів з близькими властивостями), побудову семантичних мереж (аналіз зв'язків між поняттями, що екстрагуються з документів), витяг фактів, понять (Feature Extraction, отримання деяких фактів з тексту з метою поліпшення класифікації, пошуку, кластеризації та побудови семантичних мереж), реферування (Summarization, складання коротких викладів матеріалів), відповіді на запити (Question Answering, формування пошукових образів документів), тематичне індексування (Thematic Indexing), пошук за ключовими словами (Keyword Searching), засоби створення і підтримки таксономії (Taxonomies), тезаурусів (Thesauri) і онтологій (Ontology).

5. Моніторинг інформаційного простору - адаптація концепції глибинного аналізу текстів (Text Mining) і класичних методів контент аналізу до умов формування і розвитку динамічних інформаційних масивів даних. Завдання вирішуються системами моніторингу: створення сучасної ефективної інформаційної технології та заснованого на ній он-лайнового сервісу, що акумулюють різні математичні методи та алгоритми, орієнтовані на моніторинг і багатофакторний аналіз медіа-простору. При цьому повинні виконуватися наступні умови: безперервний автоматизований моніторинг медіа-простору, що включає в себе тисячі найбільш рейтингових веб-ресурсів, основні державні і регіональні інтернет-джерела, ЗМІ та телевізійні канали (відеомоніторинг); акумулювання різноманітних розрізнених баз даних в єдину інформаційну систему; накопичення в структурованому вигляді результатів моніторингу для подальшого аналізу; використання сучасних методів кількісного і якісного аналізу інформації. В цілому існуючі на ринку системи можуть надавати такі можливості: SiteSputnik - програмний комплекс, аналоги якого відсутні, що дозволяє вести пошук і обробку його результатів у видимому і невидимому Інтернеті, використовуючи всі необхідні користувачу пошуковики; WebSite-Watcher – дозволяє проводити моніторинг веб-сторінок, включаючи захищені паролем, моніторинг форумів, RSS каналів, груп новин, локальних файлів. Володіє потужною системою фільтрів. Моніторинг ведеться автоматично і поставляється в зручному для користувача вигляді; http://web-data-extractor.net/ – універсальне рішення для отримання будь-яких даних, доступних в інтернеті; CaptureSaver – професійний інструмент дослідження інтернету. Програма, що дозволяє захоплювати, зберігати і експортувати будь-яку інтернет інформацію, включаючи web сторінки, блоги, RSS новини, електронну пошту, зображення і багато іншого; http://www.orbiscope.net/en/software.html – система веб моніторингу; Maltego – принципово нове програмне забезпечення, що дозволяє встановлювати взаємозв'язок суб'єктів, подій і об'єктів в реалі і в інтернетіта т.п.

6. Оцінка залученості користувачів (аналіз передруків, ретвітів, лайків, коментарів) [4-5]. В результаті аналізу збирається статистика того, хто або що збирає максимум лайків від об'єкта, виявляється оцінка впливовості авторів і кількості ботів серед них. Розглядаючи окремо кількість лайків, репостів, коментарів, як і кількість передплатників, можна

зіткнутися з таким, що вводить в оману, параметром успіху, в якому зростання окремих метрик (від числа лайків до кількості передплатників і реєстрацій) розглядається як безсумнівний показник ефективності в соціальних мережах. Насправді ці показники, відірвані від інших даних, не впливають ні на що. У свою чергу, загальна кількість цих метрик (обсяг) може бути використана для розрахунку важливіших показників: коефіцієнта залучення і (в деяких випадках) вартості залучення. Коефіцієнт залучення відповідає на два головних питання: наскільки актуальний та цікавий контент; наскільки аудиторія хоче отримати інформацію.

Для отримання оцінки використовуються семантичний аналіз (компонентна оцінка кількості слів або фраз, що визначають основний зміст тексту – семантичне ядро, і статистичних показників) [6], дослідження каскадів поширення інформації (необхідність того, щоб інформація, яку потрібно донести до аудиторії, мала властивості інформаційного каскаду. Це показник віральності – цікавості контенту) [7]. Метод океану (методи створення безконкурентних ринків, які дозволяють миттєво отримувати додаткову вигоду компанії, її покупцям, її працівникам за допомогою пошуку нового попиту і роблячи конкуренцію, як таку, непотрібною). Визначення ботів і стеження за їх активністю. Робот пошукової машини (бот від англ. Bot), який ходить по сайтах, індексуючи їх контент. Це автоматичний скрипт, який працює за визначеним розкладом. Пошукові боти мають свої юзер-агенти, через які їх легко визначити і мають певну функцію. Наприклад, існують боти, які індексують текстовий контент, відео та аудіо контент, зображення, боти для блогів, для новин та інші роботи. Методи визначення: за допомогою різних сервісів (FakeOFF, VkFak, Twitter Audit і ін.), за IP-адресою і ін. Для виявлення подібності принципів взаємовідносин учасників соціальних мереж використовується такий напрямок аналізу як рольові алгебри, які фокусуються на виявленні логіки взаємодій учасників мережі в блокових моделях [4, 8]. Аналіз діад і тріад дає важливий показник сили зв'язків між учасниками, яка визначається як лінійна комбінація тривалості, емоційної насиченості, інтимності або конфіденційності та значущості взаємних послуг, які характеризують дані взаємодії.

7. Визначення взаємозв'язків. Виявлення неочевидних закономірностей і зв'язків з текстів веб-сторінок і виявлення їх взаємозв'язків, побудова матриць і графів взаємозв'язків. База даних може розглядатися у вигляді графа, вершинами якого виступають об'єкти - терми, поняття та ін., а ребрами - їх зв'язки. Основа пошуку - це пошук вершин, тобто пошук об'єктів. При проектуванні баз даних зв'язків використовуються перспективні рішення в області створення інформаційно-аналітичних систем – Text Mining, Information Extraction, Big Data (технології баз даних надвеликих обсягів), Complex Network (концепція складних мереж). В рамках теорії складних мереж вивчаються характеристики, пов'язані з топологією мережі, статистичні феномени, розподіл ваг окремих вершин і ребер, ефекти провідності і протікання в мережах.

Пов'язані підгрупи (спільноти) в мережі характеризуються наявністю великої кількості зв'язків між учасниками, що входять до них, та істотно меншим числом зв'язків з іншими учасниками. Аналіз спільнот дозволяє вивчати стійкість соціальних структур [1,3,9]. Найпростіший випадок пов'язаної групи – це спільнота, де кожен учасник пов'язаний з кожним, і до цієї групи не можуть бути включені інші учасники мережі, оскільки вони не мають зв'язків з усіма членами спільноти (кліки). Таким чином, кліка - це максимальний повний підграф даного графа. Якщо аналізувати процеси поширення інформації в графах, то можна дати інше визначення спільноти, як множини учасників, де шлях між двома будь-якими учасниками не містить більше однієї проміжної вершини. В результаті інформація від одного учасника до іншого в пов'язаній групі передається з мінімальними спотвореннями. Пов'язані групи також можуть бути виділені за допомогою багатовимірного шкалювання (аналіз і візуалізація даних за допомогою розташування точок, відповідних досліджуваним – шкаліруемим об'єктам, в просторі меншої розмірності ніж простір ознак об'єктів) або факторного аналізу матриці зв'язків графа (зниження розмірності матриці зв'язків графа, тобто виділення у всій сукупності ознак тих, які дійсно впливають на зміну залежною змінною) [7].

8. Оцінка охоплення. Одна справа – кількість згадок, і зовсім інша справа – охоплення аудиторії, одне без іншого не має практично ніякого сенсу. Охоплення (reach) – це кількість представників цільової аудиторії, які в рамках кампанії мали контакт з поширюваною інформацією (новиною, рекламою, дезінформацією і т.д.) певну кількість разів [3,9]. Чим більше кількість потенційних людей, до яких реклама була донесена, тим більше охоплення. Ідеальне охоплення – отримання повідомлення ста відсотками цільової аудиторії. Однак на практиці повне охоплення залишається, як правило, недосяжною метою. Охоплення аудиторії може бути представлене як характеристика аудиторії, яка бачила (чула) інформацію певну кількість разів, і як характеристика аудиторії, яка бачила (чула) інформацію не менше певної кількості разів. Оцінка охоплення здійснюється по кожному носію інформації. Найбільш важливою є оцінка того, чи вдалося забезпечити досягнення цілей охоплення і частоти. Охоплення може бути виражене як кількісно, тобто в одиницях, тисяча, мільйони людей, так і у відсотках від цільової аудиторії. Для збільшення охоплення використовують різні методи, включаючи використання негативної інформації. Справа в тому, що негативна інформація поширюється набагато швидше і охочіше, а також часто має набагато більшу вагу. Тобто картина співвідношення позитиву і негативу в абсолютній кількості знайдених згадок сильно відрізняється від картини співвідношення при оцінці за охопленням.

9. Демографічна і географічна класифікація. Стать, вік, трудові ресурси, географічний розподіл, расовий, етнічний, релігійний склад населення і інше. Основні методи класифікації: за допомогою дерев рішень (метод автоматичного аналізу даних та подання правил в ієрархічній, послідовній структурі, де кожному об'єкту

відповідає єдиний вузол, що дає рішення) [4]; байєсівська (наївна) класифікація (простий імовірнісний класифікатор, заснований на застосуванні теореми Байєса зі строгими (наївними) припущеннями про незалежність); за допомогою штучних нейронних мереж (система з'єднаних і взаємодіючих між собою простих процесорів (штучних нейронів) [10]; метод опорних векторів (набір схожих алгоритмів навчання з учителем, що використовуються для задач класифікації і регресійного аналізу) [8]; статистичні методи, зокрема, лінійна регресія (регресійна модель залежності однієї змінної від іншої або кількох інших змінних з лінійною функцією залежності) [6]; за допомогою методу найближчого сусіда (метричний алгоритм для автоматичної класифікації об'єктів. Основним принципом методу найближчих сусідів є те, що об'єкт присвоюється тому класу, який є найбільш поширеним серед сусідів даного елемента) [4]; CBR-метод (це метод отримання рішення шляхом пошуку подібних проблемних ситуацій в пам'яті, що зберігає минулий досвід вирішення завдань, і адаптації знайдених рішень до нових умов) [3,4]; за допомогою генетичних алгоритмів (евристичні алгоритми пошуку, які використовуються для вирішення завдань оптимізації та моделювання шляхом випадкового підбору, комбінування і варіації параметрів, які шукаються, з використанням механізмів, аналогічних природному відбору в природі) [10].

10. Аналіз тональності висловлювань по відношенню до інформаційних об'єктів. Проводиться кластеризація повідомлень за сюжетами з можливістю знайти кожне окреме повідомлення і його характеристики (в тому числі тональність); ранжування сюжетів за обговорюванням. Також використовується для попередження про можливі ризики – полягає в спостереженні за кількісною та якісною динамікою обговорення інформаційного об'єкта і прогнозі подальшої динаміки. Отримання інформації з коротких неформальних повідомлень, таких, як твіти та смс, відстеження подій, вилучення проблем, виявлення сарказму та ін. Традиційний підхід до класифікації тональності ґрунтується на присутності негативних і позитивних слів або піктограм, що зображують емоцію (емотикони), які служать індикаторами позитивного або негативного забарвлення. Застосовуються також змішані підходи, де лексичні ресурси (словники тональності) комбінують з методами машинного навчання. Аналіз проводиться з використанням рішення бінарної класифікації [11], завдання класифікації послідовностей (sequential classification) [12], завдання тематичного моделювання або традиційної задачі кластеризації [13]. Мета класифікації – визначити, чи є терміни, іменники та словосполучення шуканим об'єктом (аспектом, щодо якого висловлюється деяка думка). В [14] пропонується підхід, заснований на правилах, в якому використовуються дерева залежностей для пропозицій. Також використовуються методи умовних випадкових полів (Conditional Random Fields, CRF) для класифікації послідовностей в завданні вилучення аспектних термінів.

11. Оцінювання транзакцій в інтернеті (фінансові транзакції через кредитні карти, покупки і використання авіаквитків, інформація про інших людей, з якими об'єкт веде справи, а також про те, як це все вписується в загальну історію транзакцій і т.д.) як частина «загального патерну активності» користувача. Використовуються методи та інструменти оцінки ризиків на базі моніторингу транзакцій, методи машинного навчання, статистичні методи.

12. Аналіз потоків подій (можливість «пов'язувати» і оцінювати вплив зовнішніх подій на ключові події / заходи). Застосовується для своєчасного виявлення шахрайства в сфері фінансових послуг; для оперативного аналізу активностей на біржових торгових майданчиках; для оперативного аналізу відтоку клієнтів і проведення маркетингових кампаній, для оцінки впливу політики на соціальну та економічну області та ін.

Обробка потоків подій (англ. Event Stream Processing, ESP) - набір технологій, призначених для побудови інформаційних систем обробки подій. ESP технологія включає в себе візуалізацію подій, їх зберігання, кероване по подіям сполучне програмне забезпечення та мови програмування обробки подій. Основним завданням для ESP є обробка потоків подій (даних) з метою виявлення в них значущих шаблонів, використовуючи такі методи як пошук взаємозв'язків між подіями, кореляція подій, ієрархії подій, та інші аспекти, такі, як причинність, аналіз складових подій і часових рядів.

Для аналізу використовується концептуальна модель обробки подій. Вона розглядається в термінах мережі обробки подій, яка лежить в її основі (Event Processing Network), та асоційованої з нею концептуальної архітектури для обробки подій (Conceptual Architecture for Event Processing). Також використовується теорія ризиків (ризик - це комбінація ймовірності події та її наслідків), кореляційний аналіз (метод, що дозволяє виявити залежність між декількома випадковими величинами).

13. Прогнозування. Методи прогностичного моделювання, які дозволяють за минулим передбачати майбутнє, застосовуються при вирішенні безлічі завдань, таких як рекомендуючі системи, виявлення шахрайства та зловживань або профілактика захворювань і нещасних випадків.

Прогноз успішності процесу передачі дезінформації, пошук закономірностей в соцмережах для генерації даних, необхідних для прогнозування районів, в яких з найбільшою ймовірністю можуть відбуватися злочини і терористичні акти, рекомендація товарів і послуг, ґрунтуючись на звичках об'єктів, допомога лікарям приймати рятівні профілактичні заходи з урахуванням сприйнятливості людини до конкретних захворювань, для передбачення епідемій, а деякі навіть пророкують стихійні лиха і їх наслідки, аналізуючи контент соціальних медіа [8]. Прогностичні моделі використовуються і для реалізації функції класифікації, коли результатом є клас або категорія. Найбільш популярними є методи: нейронних мереж (NN) [10], кластеризації [4], метод опорних векторів (SVM) , асоціативні правила, дерева рішень, лінійна і логістична регресія і оціночні таблиці. Для передбачення появи нових зв'язків в мережі за проміжок часу в [2] застосовується автоматичне моделювання процесу

розвитку соціальної мережі з залученням деяких характеристик мережі, таких як кількість спільних сусідів, найкоротший шлях, впливовість вершини, момент першого попадання в соціальну мережу. Також застосовуються прогностичні моделі засновані на машинному навчанні, що використовують особисту інформацію про користувачів мережі для підвищення точності передбачення [15]. Іноді застосовують ієрархічні, імовірнісні (марковські) і реляційні моделі для виявлення зв'язків між користувачами [16].

В інших моделях [9, 16] за основу пропонується брати самі властивості користувачів, і, наприклад, наявність великої кількості зв'язків (в блогосфері), що може бути пояснено шляхом зіставлення демографічних груп, спільних інтересів або географічною близькістю.

III. Висновки

У даній статті розглянуто основні види аналізу даних в соціальних мережах. Для кожного виду аналізу наведено перелік необхідних даних, короткий огляд основних математичних методів, які використовуються для проведення аналізу. Також описані показники, які можна отримати після аналізу, їх інформативність та використання в подальшому аналізі або представлені результатів аналізу. Описано взаємозв'зки показників даних та аналітичних розрахунків для кожного виду аналізу соціальних мереж.

Література REFERENCES

[1] Т.В. Батура Модели и методы анализа компьютерных социальных сетей. // Батура Т.В. / Программные продукты и системы, № 3. – 2013. - С. 130-137.

[2] Liben-Nowell D., Kleinberg J. The Link Prediction Problem for Social Networks // Proceedings of the 12th International Conference on Information and Knowledge Management. N. Y .: ACM Press, 2003. P.556-559.

[3] Д.В.Ланде, А.А.Снарский , И.В.Безсуднов Интернетика: навигация в сложных сетях: модели и алгоритмы. М. книжный дом «Либроком», 2009. – 264с.

[4] F.Sebastiani Machine Learning in Automated Text Categorization. URL: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf.3.

[5] Н.Г.Федотов Теория признаков распознавания образов на основе стохастической геометрии и функционального анализа. – М.:ФИЗМАТЛИТ, 2010. – 304с.

[6] G.Sidorov, S.Miranda-Jimenez, F.Viveros-Jimenez, A.Gelbukh, N.CastroSanchez, F.Velasquez, J.Gordon Empirical study of machine learning based approach for opinion mining in tweets // Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2013. P. 1-14.

[7] Как использовать социальный граф для распространения контента URL: http://www.pvsm.ru/sotsial-ny-e-seti/64097)

[8] CJC Burges. A Tutorial on Support Vector Machines for Pattern Recognition. URL: http://www.music.mcgill.ca/_rfergu/adamTex/references/Burges98.pdf

[9] R.Kumar, J.Novak, P.Raghavan, A.Tomkins Structure and Evolution of Blogspace // Commun. ACM. 2004. Vol. 47. No. 12. P. 35-39.

[10] Leonid Lyubchyk, Eugenie Bodyansky, Arkady Rivtis. Adaptive harmonic components detection and forecasting in wave non-periodic time series using neural networks. URL: https://pdfs.semanticscholar.org/c879/8f1ac64fa076c9df437a860645fd5e463040.pdf.

[11] A.M. Popescu, O.Etzioni Extracting product features and opinions from reviews // Natural language processing and text mining. ACL, 2007. P. 9-28.

[12] N.Jakob, I.Gurevych Extracting opinion targets in a single-and crossdomain setting with conditional random fields // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. ACL, 2010. P. 1035-1045.

[13] Y.Zhao, B.Qin, T.Liu Clustering product aspects using two effective aspect relations for opinion mining // Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer International Publishing, 2014. P. 120-130.

[14] S.Poria, E.Cambria, W.Ku, C.Gui, A.Gelbukh A rule-based approach to aspect extraction from product reviews // Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP). 2014. P. 28-37.

[15] B.Taskar, M.Wong, P.Abbeel, D.Koller Label and Link Prediction in Relational Data. URL: http://kdl.cs.umass.edu/srl2003_upload/files/taskar-paper.pdf.

[16] Saoussen Aouay, Salma Jamoussi, Faiez Gargouri, Ajith Abraham. Modeling Dynamics of Social Networks: A Survey. Sixth International Conference on Computational Aspects of Social Networks. – 2014. P.49-53.