УДК 004.822:004.912

Д. В. Ланде, Б. О. Березін

Аналіз і розробка засобів статистичної обробки інформації з Інтернету для технології OSINT

1. Постановка проблеми

Під OSINT (Open Source Intelligence, розвідка на основі відкритих джерел) розуміють технології збору даних із загальнодоступних джерел для використання у розвідувальних цілях. У складі OSINT виділяють етапи формулювання вимог, збору даних, обробки, аналізу і добування необхідної інформації, розповсюдження готової продукції. Отримані результати значною мірою залежать від використання методів статистичної обробки на етапі аналізу.

2. Мета роботи

Розвинути та адаптувати основні методи та засоби статистичної обробки інформації з Інтернет для технологій OSINT. Розробити метод побудови та використання семантичних моделей (СМ) для аналізу суспільної думки з результатами лінгвостатистичного аналізу текстів із мережі Інтернет.

3. Аналіз методів і засобів статистичної обробки

Серед методів, що використовуються для статистичної обробки інформації у технологіях OSINT в [1] відзначають методи виявлення подій (event-detection). В роботі розглянуто два методи виявлення подій на основі потоку повідомлень з мережі Twitter: цілеспрямований на конкретний тип подій, та не цілеспрямований. Виявлення подій без попереднього знання типу, характеру, масштабу є більш складною проблемою; відповідний метод використовує кластеризацію. Обидва ці методи застосовуються для виявлення подій; швидких змін у поведінці груп; виявлення змін суспільних настроїв; подій, орієнтованих на певний регіон тощо.

У роботі [2] запропоновано методи виявлення подій з потоку новинних повідомлень, а також реалізації двох відповідних програмних засобів NEXUS та PULS і наведено їхню оцінку. В [2] також запропоновано кластерну модель класифікації для підозрілих повідомлень електронної пошти та для категоризації новин. У моделі використано простий ID3-алгоритм для класифікації та K-means алгоритм для кластеризації.

Значна кількість методів обробки інформації в OSINT пов’язана з аналізом соціальних мереж (social network analysis — SNA) та використовується для аналізу терористичних мереж. У [2] запропоновано показник для оцінки організації мережі та виявлення найбільш впливових осіб у мережі. Тестування, проведене на основі даних з реальних терористичних мереж, показало ефективність використання запропонованого методу виявлення впливових осіб у мережі. Також запропоновано та показано ефективність показника для аналізу важливості зв’язків у терористичній мережі. Крім того, в [2] запропоновано метод виявлення прихованих зв’язків між вузлами мережі на основі аналізу доступної інформації. Експерименти, проведені на основі даних з реальних терористичних мереж, показали можливість використання методу.

Серед програмних засобів, що використовують для статистичної обробки інформації в OSINT відзначають доступні для вільного використання пакет для аналізу мереж і візуалізації Gephi (http://gephi.org), а також пакет Maltego (https://www.paterva.com), який крім збору інформації, забезпечує аналіз зв’язків між різними видами об’єктів. До засобів статистичної обробки інформації в OSINT також відносять мови програмування Python, R та інші, що забезпечують використання бібліотек для обробки текстів, аналізу мереж, кластеризації тощо. На Python розроблена відкрита нейромережева бібліотека Keras, що дозволяє використовувати для аналізу методи глибинного навчання.

4. Підходи до розробки засобів статистичної обробки

У роботі запропоновано метод побудови та використання СМ (моделі предметної області, що має вигляд орієнтованого графа, вершини якого відповідають концептам предметної області, а дуги задають відносини між ними) для задач моніторингу суспільної думки в мережі Інтернет. Метод передбачає три етапи [3]: побудову та кластеризацію СМ; відбір документів і визначення тональності тематик; візуалізацію результатів.

На першому етапі проводиться: вибірка масиву документів для побудови СМ; знаходьження концептів; визначення зв'язків СМ шляхом побудови компактифікованого графа горизонтальної видимості (КГГВ) [4]; кластеризація графа; формування запитів, відповідних кластерам (на основі знайдених кластерів експертами виділяються тематики і формулюються запити для відбору відповідних документів). Для визначення зв'язків між концептами та побудови семантичної моделі використовується алгоритм КГГВ [4], особливість використання якого в даній роботі в тому, що його перші два кроки виконуються окремо для кожного речення тексту, яке аналізується. Після цього отримана мережа компактифікується. Для кластеризації графів СМ у даній роботі розглядалося застосування різних відомих алгоритмів [5], кращі результати були отримані при використанні алгоритмів Louvain, Leading Eigenvector, а також Walktrap.

На другому етапі проводиться: відбір документів відповідних тематик (підтем) із загального інформаційного потоку за допомогою запитів; визначається їхня частка в загальному потоці документів; визначається тональність документів відповідних тематик. На третьому етапі тематики з тональностями: візуалізуються на карті; записуються в базу даних (БД) системи моніторингу для подальшого отримання динаміки зміни результатів у часі.

Для реалізації запропонованого методу використані засоби програмного пакету Gephi, а також програмні засоби, розроблені на R-мові програмування для статистичних розрахунків.

Можливості застосування запропонованого методу побудови та використання СМ для моніторингу суспільної думки проаналізовано на основі результатів моніторингу Інтернет-ресурсів по декількох темах [3]: One Belt, One Road; Nord Stream та іншим.

5. Висновки

Розглянуто основні методи та засоби статистичної обробки інформації, що використовуються в OSINT: виявлення подій, кластеризація, класифікація, методи аналізу мереж, програмні пакети, мови програмування тощо.

Запропоновано метод побудови та використання СМ для моніторингу суспільної думки, що включає три етапи. Показано побудову СМ за допомогою алгоритму КГГВ, застосування методів кластерного аналізу для визначення актуальних тематик, оцінювання частки та тональності окремих підтем у складі загального тематичного потоку інформації.

Отримані результати підтверджують можливість використання запропонованого методу моніторингу суспільної думки в різних предметних областях OSINT.

1. Hobbs C., Moran M., & Salisbury D. (Eds.). Open source intelligence in the twenty-first century: new approaches and opportunities. Springer, 2014. 191 p.

2. Wiil U. K. Counterterrorism and open source intelligence. Springer, 2011. 458 p.

3. Aleksandr Dodonov, Dmitry Lande, Boris Berezin Semantic Models at Task Monitoring Public Opinions. «Информационные технологии и безопасность». Материалы XVIII Международной научно-практической конференции ИТБ-2018. Киев: ООО «Инжиниринг», 2018. С. 338–346. CEUR Workshop Proceedings (ceur-ws.org). Vol. 2318 urn:nbn:de:0074-2318-4. Selected Papers of the XVIIIШ International Scientific and Practical Conference on Information Technologies and Security (ITS 2018). URL: http://ceur-ws.org/Vol-2318/paper1.pdf]

4. Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V. The use of horizontal visibility graphs to identify the words that define the informational structure of a text. 12th Mexican International Conference on Artificial Intelligence (MICAI), 2013. P. 209–215. DOI: 10.1109/MICAI.2013.33

5. Harenberg S., Bello G., Gjeltema L., Ranshous S., Harlalka J., Seay R., Samatova N. Community detection in large-scale networks: a survey and empirical evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2014. Issue 6(6). P. 426–439.