УДК 004.912:81’33

ЛІНГВІСТИЧНИЙ АНАЛІЗ ТЕКСТІВ НА ВИЯВЛЕННЯ ТЕРРОРИСТИЧНОЇ ЗАГРОЗИ

Бісікало Олег Володимирович – д.т.н, проф., декан факультету комп’ютерних систем і автоматики

Санаулла Фаяз Дава Хан – студент групи 2СІ-13б, факультет комп’ютерних систем та автоматики

Вінницький національний технічний університет.

У час стрімкого розвитку інформаційних технологій Інтернету результати їх розповсюдження все вагоміше впливають на всі сфери людського життя. За останні десятиліття Інтернет-технології знайшли застосування практично у всіх галузях суспільного життя – починаючи від комунікації та закінчуючи Інтернет-магазинами, розумними будинками, безпілотними автомобілями.

Окреслені тенденції і стали причиною збільшення обсягів вільно доступної інформації, що, у свою чергу, спричинило потребу в автоматизованій обробці і аналізі інформації. Цим займається напрям досліджень Data Mining, задачами якого є аналіз вхідної інформації і отримання на його основі нової корисної інформації. Останнім часом через високу нестабільність в світі, постійних зовнішньополітичних конфліктів і внутрішньої ворожнечі, все частіше з’являються терористичні загрози зі сторони людей, які бажають нанести шкоду мирному населенню та стабільності у цілому. З виникненням соціальних мереж та засобів комунікації типу Messenger та Skype з’явилась можливість виявляти загрози для суспільства на етапі їх формування або в процесі їх підготовки через аналіз інформаційного потоку у вигляді текстів. Зазвичай для цього використовується комбінація методів лексичного аналізу з аналізом даних та машинним навчанням.

Машинне навчання – це підгалузь інформатики, яка еволюціонувала з дослідження і розпізнавання образів, теорії обчислювального навчання та інших методів штучного інтелекту [1]. У 1959 році Артур Семюель визначив машинне навчання як «Галузь досліджень, яка дає комп'ютерам здатність навчатися без того, щоби їх явно програмували» [2]. Машинне навчання досліджує вивчення та побудову алгоритмів, які можуть навчатися з даних, а також виконувати передбачуваний аналіз на них [3]. Такі алгоритми діють шляхом побудови моделі зі зразкового *тренувального набору* вхідних спостережень з метою здійснювати керовані даними прогнози або ухвалювати рішення, що виражені як виходи – замість того, щоби суворо слідувати статичним програмним інструкціям.

Великі дані в інформаційних технологіях – це набори інформації настільки великих розмірів, що традиційні способи та підходи не можуть бути застосовані до них. Альтернативне визначення називає великими даними феноменальне прискорення нагромадження даних та їх ускладнення, що є дуєе

характерним явищем сучасного інформаційного суспільства. Важливо також відзначити те, що часто під цим поняттям у різних контекстах можуть розуміти як дані великого об'єму, так і набір відповідних інструментів та методів (наприклад, засоби масово-паралельної обробки даних системами категорії NoSQL, алгоритмами MapReduce чи програмними каркасами проекту Hadoop та Apache Spark).

Пропонується створити інформаційну лінгвістичну технологію аналізу текстів на виявлення терористичної загрози на основі лексичного аналізу текстових даних. Першим етапом є базове розбивання тексту на дискретні частини з подальшим парсингом на слова та частини мови. Далі формуються асоціативні пари зі слів, які залишились після очищення тексту від зайвого (множини стоп-слів), та окремо визначаються такі пари, що можуть вважатися загрозливими. Проводиться чисельна оцінка як сформованих пар, так і окремих слів. Наступним кроком є подання вхідних векторів на вхід логістичної регресії з метою створення прогнозу на основі попередньо навченої моделі. Особливістю запропонованого підходу є застосування платформи паралельних обчислень Apache Spark, на якій відбувається реалізація алгоритму, що дає можливість значно прискорити роботу і навчання моделі, а також забезпечує ефективну обробку великих обсягів даних.

Попереднє навчання відрізняється від звичайного запуску базового алгоритму лиш тим, що на вхід подаються тестові дані з попередньо визначеними результатами та на їх основі формується лінія розподілу логістичної регресії, яка і виконує прогнозування.

До основних переваг використання запропонованої технології можна віднести:

1. Отримання якісно нових знань за рахунок комплексного аналізу усієї інформації у єдиному аналітичному сховищі.
2. Розширення функціональності існуючих інформаційних систем підтримки бізнесу.
3. Збільшення ефективності використання апаратних ресурсів серверів.
4. Забезпечення мінімальної вартості використання всіх видів інформації за рахунок можливості використання програмних засобів з відкритим кодом і хмарних технологій.

Критика великих даних пов'язана з тим, що їх зберігання не завжди приводить до отримання вигоди, а швидкість оновлення даних і «актуальний» часовий інтервал не завжди розумно порівнянні.

Scala — мультипарадигмова мова програмування. Назва Scala утворена зі слів "scalable" та "language", для того щоб задекларувати, що мова може рости разом з вимогами користувачів. Програми мовою Scala виконуються на віртуальній машині Java за умови приєднання до дистрибутиву файлу scala-library.jar. Scala сумісна із існуючими програмами мовою Java, тобто код Scala може викликатися із Java-програм і навпаки. Починаючи з версії 2.11 Scala потребує принаймні Java 6 [1], а версія 2.12 потребуватиме Java 8 та матиме кращу інтеграцію із новими можливостями цієї версії Java [2].

Apache Spark — високопродуктивний рушій для оброблення даних, що зберігаються в кластері Hadoop. У порівнянні з наданим у Hadoop механізмом MapReduce, Spark забезпечує у 100 разів більшу продуктивність при обробленні даних в пам'яті й 10 разів при розміщенні даних на дисках [1]. Рушій може виконуватися на вузлах кластера Hadoop як за допомогою Hadoop YARN, так і у відокремленому режимі. Підтримується оброблення даних у сховищах HDFS, HBase, Cassandra, Hive та будь-якому форматі введення Hadoop.

У результаті дослідження було проведено аналіз існуючих методів, пов’язаних з обробкою великих обсягів текстової інформації. Запропоновано синтезувати інформаційну лінгвістичну технологію аналізу текстів на виявлення терористичної загрози шляхом поєднання методів опрацювання текстових даних з експертною моделлю, що відрізняється функціоналом подальшого масштабування на платформі для паралельних обчислювань.

Список використаної літератури

1. Spark latests documentation [Електронний ресурс]: - Режим доступу: http://spark.apache.org/docs/latest/

2. M. Odersky, L. Spoon, B. Venners. Programming in Scala. : Artima. – 2008. – P. 736

3. Барсегян А., Куприянов М., Холод И., Степаненко В. Методы и модели анализа данных: OLAP и Data Mining: Учеб. пособ. для вузов / Програмирование 2004 — 331с.