Секція: Технічні науки

Василенко Сергій Сергійович
студент факультету комп'ютерних наук і програмної інженерії
Харківського політехнічного інституту
м. Харків, Україна

ОГЛЯД ОСНОВНИХ ПРОБЛЕМ ТА МЕТОДІВ КЛАСТЕРИЗАЦІЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ З МЕТОЮ ВИЯВЛЕННЯ КЛЮЧОВИХ СЛІВ ЕКСТРЕМІСТСЬКОГО НАПРЯМУ В МЕРЕЖІ ІНТЕРНЕТ

За останнє десятиліття терористичні і екстремістські організації значно збільшили свою присутність в мережі Інтернет та соціальних мережах, активно використовуючи ці кошти для вербування нових членів і їх навчання, підготовки і організації терористичних атак, пропаганди насильства, поширення екстремістської літератури і т.п. Заходи, спрямовані на виявлення терористів і пов'язаних з ними осіб, припинення поширення екстремістських матеріалів, запобігання готуючихся терактів вимагають аналізу всієї інформації, що надходить від представників екстремістських угруповань. В силу величезного обсягу поширюваної через Інтернет інформації, її мовного різноманіття та вимоги її моніторингу в режимі реального часу необхідно використовувати автоматичні процедури текстового аналізу з метою виявлення потенційно небезпечних користувачів, своєчасного видалення екстремістських матеріалів, аналізу інформації про терористів і теракти. Основними завданнями при створенні автоматичних засобів аналізу інформації терористичної спрямованості є вибір відповідних даних для тестування алгоритмів і розробка алгоритмів, придатних для вирішення завдання виявлення терористичної активності.

Таким чином, розробка автоматичних засобів тематичного аналізу дозволить істотно підвищити ефективність вирішення завдань пошуку в Інтернеті документів і окремих повідомленні терористичної та екстремістської спрямованості, що, в свою чергу, призведе до можливості запобігання готуються терактів, зменшення впливу екстремістських груп і підвищення рівня національної безпеки.

У задачі автоматичного анотування як текстів використовуються окремі фрагменти документа, наприклад, пропозиції. Таким чином, колекція фрагментів документа представляється у вигляді числової матриці A e RmX", рядки якої відповідають ознакам, а стовпці – фрагментами. В якості ознак в моделі "мішок слів" використовуються терми – лексеми, що входять в текст. Мета попередньої обробки тексту – залишити тільки ті ознаки, які найбільш інформативні, тобто найбільш сильно характеризують текст. До того ж скорочення числа аналізованих ознак призводить до зменшення обсягу використовуваних обчислювальних ресурсів.

Для складання анотації вибирається певна кількість пропозиції з найбільшими значеннями отриманої релевантності. Таким чином, ідея запропонованого методу автоматичного анотування полягає у виділенні основних тематик в тексті документа і знаходженні фрагментів тексту, які найкращим чином описують виділені тематики, шляхом розрахунку їх релевантності.

Для побудови результуючого документа, що не містить інформаційного шуму, вибираються його фрагменти з максимально релевантної, сума яких не перевищує заданий відсоток інформації, як правило, дорівнює не більше 30% [1].

Таким чином, розробка автоматичних засобів тематичного аналізу дозволить істотно підвищити ефективність вирішення завдань пошуку в Інтернеті документів і окремих повідомлень терористичного і

екстремістской спрямованості, що, в свою чергу, призведе до можливості запобігання готуються терактів, зменшення впливу екстремістських груп і підвищення рівня національної безпеки.

В даний час існує велика кількість систем автоматичного вилучення ключових фраз з тексту на природній мові. Ключовими факторами при відборі аналогів в даній статті були рекомендації експертів, кілька дослідницьких робіт, присвячених аналогам, а також популярність відповідних систем в сучасному IT-співтоваристві.

OpenCalais – Web-сервіс, призначений для автоматичного вилучення семантичних метаданих з текстів природною мовою. Починаючи з 2007 року, розвитком і підтримкою сервісу займається корпорація Thomson Reuters.

Семантичні метадані представлені у вигляді іменованих сутностей (англ. Named entity), а також пов'язаних з ними фактів і подій. Іменовані суті, в свою чергу, можуть розглядатися як ключові слова і фрази вихідного тексту.

Функціонування системи OpenCalais засноване на методах обробки природної мови, машинного навчання та інших алгоритмах. Для вилучення семантичних метаданих застосовуються попередньо підготовлені онтології різних предметних областей в форматі RDF. Оригінальний текст піддається попередній обробці (графематіческой і морфологічної розмітки), потім розмічені словосполучення проходять ідентифікацію за допомогою навченої моделі розпізнавання іменованих сутностей, між якими ведеться пошук семантичних відносин. Отриманий граф сутностей і відношень між ними перетворюється в набір RDF-трійок [2].

Extractor – система автоматичного вилучення термінів, що функціонує з 2002 року і використовувана багатьма організаціями у власних рішеннях по обробці природної мови.

Робота системи Extractor заснована на застосуванні генетичних алгоритмів в поєднанні з методами машинного навчання та статистичними методами обробки природної мови. Початкове навчання системи ведеться на основі розміченого корпусу текстів.

TerMine – Web-сервіс вилучення термінів, розроблений в британському Національному центрі аналізу тексту (англ. The National Centre for Text Mining).

Сервіс TerMine працює на основі методу C-value і застосовує аналізатор TreeTagger для попередньої морфологічної розмітки тексту.

Демонстраційний Web-інтерфейс TerMine дозволяє обробляти тексти виключно англійською мовою.

Maui – система тематичної класифікації текстових документів, що працює на основі методів обробки природної мови та машинного навчання.

Схема функціонування системи складається з двох етапів роботи: етапи початкового побудови і навчання моделі і етапу застосування навченої моделі до вирішення завдання тематичної класифікації тексту.

Результати тематичної класифікації, отримані за допомогою Maui, можуть розглядатися в якості тегів (міток) вихідного тексту. Без використання навченої моделі Maui функціонує як система автоматичного вилучення ключових фраз [3].

Кластеризація – це завдання розбиття множини об'єктів на групи, які називаються кластерами [4]. Усередині кожної групи повинні з'явитись "схожі" об'єкти, а об'єкти різних груп повинні бути якомога більш відмінні.

У загальному випадку задача кластеризації тексту розпадається на дві:

– технічна задача перетворення в деяку матричну, векторну або будь-яку одному модель;

– математична задача кластеризації.

Спочатку необхідно виконати попередню обробку документів. Вона включає в себе наступні етапи:

– фільтрація – видалення спецсимволів і пунктуації;
– токенізація – розбивання тексту на терміни – слова або словосполучення;
– стемінг – приведення слова до основи;
– видалення стоп-слів;
– скорочення – видалення низькочастотних слів (є необов'язковим параметром);
– створення виваженої матриці терм-документ – перехід до векторів документа.

Метод TD-IDF – використовується для оцінки важливості слова в контексті документа, що є частиною колекції документів або корпусу. Вага деякого слова пропорційний кількості вживання цього слова в документі, і обернено пропорційна частоті вживання слова в інших документах колекції.

TF (term frequency – частота слова) – відношення числа входження деякого слова до загальної кількості слів документа.

$$\mathrm{tf}(t,d) = \frac{n_i}{\Sigma_k k_k} \qquad (1)$$

де – n_i число входжень слова в документ, а в знаменнику – загальна кількість слів у документі.

IDF (inverse document frequency – зворотна частота документа) – інверсія частоти, з якою деяке слово зустрічається в документах колекції. Облік IDF зменшує вагу широковживаних слів. Для кожного унікального слова в межах конкретної колекції документів існує тільки одне значення IDF.

$$\mathrm{i\,df}\,(t,D) = \log \frac{|D|}{|(d_i \supset t_i)|}, \qquad (2)$$

де – |D| кількість документів в корпусі; $|(d_i \supset t_i)|$ – кількість документів, в яких зустрічається t_i (коли $n_i \neq 0$).

Міра TF-IDF є твором цих двох характеристик:

$$tf\,i\,df(t, d, D) = tf(t, d) * i\,df\,(t, D) \quad (3)$$

В Основі методу K-means (К-середніх) лежить ітеративний процес стабілізування Центроїд кластерів. Основною характеристикою кластера є його ценроід і вся робота алгоритму спрямована на стабілізований або, в кращому випадку, повне припинення зміни центроїда кластера [4].

Переваги методу:

– низька обчислювальна складність – O (knT), де n – число документів, k – число кластерів, T – кількість ітерацій;
– метод не потребує навчанні та при необхідності може накопичувати відомості для подальшого збільшення точності роботи – використання байєсівського оцінок параметрів кластеризації.

Недоліки методу:

– потрібно завдання кількості кластерів, як мінімум на початкових етапах – до використання апріорної інформації;
– порівняно низька точність.

В основі методу жадібного пошуку лежить кластеризація наборів електронних документів виконувалася з використанням так званого жадібного алгоритму, який визнаний методом, що дає досить хороші результати при кластеризації корпусу наукових статей близької тематики, хоча і володіє порівняно великою обчислювальною складністю.

Процес можна описати кроками, циклічно повторюваними до тих пір, поки не буде "вільних" документів, які не включені ні в один з результуючих кластерів.

Метод LDA використовується для рішення завдання тематичного аналізу, яке ускладнюється низкою факторів. Інформація, поширювана терористичними групами, різнорідна, повідомлення в соціальних мережах досить короткі, містять сленг і закодовані слова, що робить безглуздим семантичний аналіз. Найбільш часто в такій ситуації використовується метод прихованого розподілу Діріхле LDA [5].

Текстовий зміст аналізується за допомогою LDA. Результати аналізу подаються в різні моделі часових рядів для прогнозування активності вербування. Кількісним аналіз показує, що використання заснованих на LDA тематик в якості предикторів в моделях часових рядів зменшує помилку прогнозування в порівнянні з випадковим блуканням, авторегресії проінтегрувати змінного середнього і експоненціальним згладжуванням.

Література

1. Машечкин И. В., Петровский М. И., Царёв Д. В. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования // Вычислительные методы и программирование. – 2013. – Т. 14. – №. 1. – С. 91-102.
2. Пескова О. В. Автоматическое формирование рубрикатора полнотекстовых документов // Тр. X Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2008). Дубна, 7–11 октября 2008 г. – С. 139–148.
3. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. Пособие. М.: МИЭМ, 2011. – 272 с.
4. Барахнин В. Б., Нехаева В. А., Федотов А. М. О задании меры сходства для кластеризации текстовых документов // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. – 2008. – Т. 6, вып. 1. – С. 3–9.

5. Федотов А. М., Барахнин В. Б. К вопросу о поиске документов «по аналогии» // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. – 2009. – Т. 7, вып. 4. – С. 3–14.