Парламентское собрание Союза Беларуси и России

Постоянный Комитет Союзного государства

Оперативно-аналитический центр
при Президенте Республики Беларусь

Государственное предприятие «НИИ ТЗИ»

Полоцкий государственный университет

КОМПЛЕКСНАЯ ЗАЩИТА ИНФОРМАЦИИ

Материалы XXII научно-практической конференции

(Полоцк, 16–19 мая 2017 г.)

Новополоцк
2017

УДК 004(470+476)(061.3)
ББК 32.81(4Беи+2)
К63

К63 **Комплексная защита информации :** материалы XXII науч.-практ. конф., Полоцк, 16–19 мая 2017 г. / Полоц. гос. ун-т ; отв. за вып. С. Н. Касанин. – Новополоцк : Полоц. гос. ун-т, 2017. – 282 с.
ISBN 978-985-531-564-4.

В сборнике представлены доклады ученых, специалистов, представителей ггосударственных органов и практических работников в области обеспечения информационной безопасности Союзного государства по широкому спектру научных напрвлений.

Адресуется исследователям, практическим работникам и широкому кругу читателей.

Тексты тезисов докладов, вошедших в настоящий сборник, представлены в авторской редакции.

УДК 004(470+476)(061.3)
ББК 32.81(4Беи+2)

ISBN 978-985-531-564-4

параметров модели. Получить оценки в явном виде не удалось, поскольку функция правдоподобия имеет сложный вид.

Для построения приближенных оценок $\hat{a}$ и $\hat{q}$ использовали следующий алгоритм.

1) Задаем начальное значение $\hat{q}^{(1)} = q_0$ и находим $\hat{a}^{(1)} = \arg\max l\left(a, \hat{q}^{(1)}; X\right)$ в предположении, что $q = \hat{q}^{(1)}$.

2) Полагая $a = \hat{a}^{(1)}$, находим $\hat{q}^{(2)} = \arg\max l\left(\hat{a}^{(1)}, q; X\right)$.

3) При $q = \hat{q}^{(2)}$ вычисляем $\hat{a}^{(2)} = \arg\max l\left(a, \hat{q}^{(2)}; X\right)$.

4) Шаги 2-3 повторяем до тех пор, пока на шаге *i* не выполнится $\hat{a}^{(i)} = \hat{a}^{(i+1)}$.

Тогда $\hat{a}^{(i)}$ и $\hat{q}^{(i)}$ и есть искомые приближенные оценки.

Чтобы найти оценки $\hat{a}^{(j)} = \arg\max l\left(a, \hat{q}^{(j)}; X\right)$, $j = \overline{1,i}$, при фиксированном q вычисляем значения функции правдоподобия при $a = 1, \ldots, n$ и выбираем то a, при котором значение функции максимально.

Оценки $\hat{q}^{(j)} = \arg\max l\left(\hat{a}^{(j-1)}, q; X\right)$, $j = \overline{2,i}$ находим сеточным методом на интервале $[0;1]$ с заданной точностью ε.

Список литературы

1. Криптология: учебник / Ю. С. Харин [и др.] - Минск: БГУ, 2013. - 511 с. - (Классическое университетское издание).

2. Мальцев, М.В. Моделирование и распознавание криптографических генераторов на основе цепей Маркова условного порядка : автореф. дис. … канд. физ.-мат. наук: 05.13.19 / М.В. Мальцев; Белорус. гос. ун-т. - Минск, 2015. – 24 с.

АНАЛИЗ ТЕКСТА ДЛЯ ПОСТРОЕНИЯ МОДЕЛИ ПОИСКА НЕЛИЦЕНЗИОННОГО КОНТЕНТА НА МЕДИАРЕСУРСАХ В СЕТИ ИНТЕРНЕТ

М.А. ДРАГУНКИН

Московский технологический университет

В работе представлен анализ текста для построения чёткой структуры медиаданных и выявление незаконного размещения аудио и видео контента. Используя название и автора композиции можно определить конкретный контент и с использованием вычислительных средств для анализа текста выявить правомерность его размещения. В работе использовались такие дисциплины, как распознавание, обработка, реферирование, аннотирование, категоризация и т. д. Все эти методики основаны на небольшом количестве базовых методик и алгоритмов, которые можно описать следующей последовательностью действий

Ключевые слова: анализ текста, анализ контента, нелицензионный контент.

Разделение текста

Основной задачей при обработке текста, является поиск и разделение текста на отдельные фрагменты или *токенизация*. Токенизация – это процесс выделения фрагментов текста, для многих текстов это слова. Выделение фрагментов текста происходит, чаще всего, по специальным *символам-разделителям* (Пример таблица №1).

Таблица 1 – Список пробельных символов

Представление	Символ	Обозначение	Расшифровка
\t	Табуляция	HT	Horizontal tabulation
\v	Вертикальная табуляция	VT	Vertical tabulation
\r	Возврат каретки	CR	Carriage return
\n	Перевод строки	LF	Line feed
\f	Конец страницы	FF	Form feed
\e	Escape-символ	ESC	Escape character
\b	Забой	BS	Backspace

Однако не всегда процесс токенизации, является простой задачей, сложностью этого процесса можно выделить несколько факторов:

Шумовые слова – люди могут совершать ошибки или даже специально использовать абсолютно бессмысленные слова с ошибками или случайные наборы символов, которые были бы крайне нежелательны в системе обработке текста, так как не несут никакого логического смысла и создают погрешности и ошибки.

Регистр символов – в некоторых случаях может играть важную роль в определении имен, мест, названий и т. д.

Результат процесса токенизации можно использовать для проверки правописания, определения типа сущности – но главное назначение – составление набора данных для дальнейшей обработки.

Определение границ частей названия

Следующий шаг в обработке текста является определение границ частей названия сущности (музыкальной композиции).

Общие методы поиска предполагают использование набора правил

Выделение отношений между элементами. Поиск именованных объектов

Используя подготовленные данные, можно, выполнить извлечение информации, например - найти в тексте именованные сущности(Name dentity recognition, NER)

Именованные сущности – это имена существительные, которые обозначают конкретные экземпляры объектов, имена, названия и т. д. Во многих ситуациях, также полезно кроме распознавания имён, мест и названий, узнать дату, числа указанное в тексте и другие сущности.

Идентификация имён людей, названий музыкальных проектов и групп, мест и других именованных сущностей позволяет определить характер сущности и предпринять соответствующие действия. Например, наличие этой информации, позволяет предложить дополнительные сведения о сущностях – рекомендовать сопутствующие материалы и, в конце концов, повысить интерес к приложению или веб-сайту.

Допустим, человек читает статью на новостном сайте, а сайт ему предлагает ссылки на соответствующие темы или людей, о которых ведется речь, человек, переходит по ссылке и дальше снова получает тематически ссылки на статью по этой тематике.

Одним из первых решений данного подхода были системы, основанные на большом количестве правил *регулярных выражений* (RegEx), суть ее заключалась в следующем: использую синтаксис регулярных выражений, создавались специальный шаблон, по которому находятся совпадения в тексте, например - шаблон для поиска email адреса в тексте:

Фактически, задача сводится к разработке анализатора названия музыкальной композиции.

Таким образом, актуальными являются исследования, связанные с разработкой информационных систем для анализа информации из сети Интернет.

Список литературы

1. Официальная документация проекта Apache OpenNLP – Режим доступа: https://opennlp.apache.org/documentation/1.5.3/manual/opennlp.html.

2. ГОСТ Р 7.0.8-2013. Делопроизводство и архивное дело. термины и определения.

3. Проект лингвистического корпуса Пенсильванского университета, [Электронный ресурс] режим доступа: http://www.cis.upenn.edu/~treebank/.

4. Обработка неструктурированных текстов - Грант Ингерсолл и др. ДМК-пресс, 2013.

5. О.В. Пескова, "Методы автоматической классификации текстовых электронных документов", ISSN 0548-0027, НТИ, Сер.2, Информ. процессы и системы, 2006, №3.

ПРОБЛЕМЫ СТАНДАРТИЗАЦИИ И БЕЗОПАСНОСТИ «УМНЫХ» СИСТЕМ

И.Д. КОТИЛЕВЕЦ

Московский технологический университет

В последние годы стала популярна концепция технологий, известных как «умный дом», «умный город», направленных на автоматизацию, ресурсосбережение, улучшение жизни. Но пока эти понятия практически никак не связаны, синергия направлена исключительно внутрь систем, из-за чего затруднено развитие инфраструктуры, жилищно-коммунального хозяйства и других служб на уровне самого города, муниципалитета и на уровне домов. Не происходит интеграции «умного дома» с «умным городом». Появляются лишние или избыточные службы или узлы, перерасход ресурсов, а также проблемы с безопасностью.

В связи с этим в настоящее время становится актуально создание стандарта проектирования систем «умных домов» для непосредственной интеграции в «умный город».

Умный дом. В основе как «умного города», так и «умного дома» лежит «интернет вещей» (IoT), объединяющий множество устройств самого разного назначения.

Поскольку система «умный дом» является разновидностью промышленной автоматизации, то универсальные настраиваемые средства могут быть основополагающими элементами для организации всего процесса управления с переходом от несвязанных подсистем к полной автоматизации и связям с высокоуровневыми приложениями вплоть до формирования отчетов о потребленных ресурсах [1, 2].