структура, состоящая из "текстовых фактов" [1]. Семантический анализ в рамках одного предложения называется локальным семантическим анализом. В общем случае семантическое представление является графом, семантической сетью, отражающим бинарные отношения между двумя узлами - смысловыми единицами текста. Глубина семантического анализа может быть разной, а в реальных системах чаще всего строится только лишь синтаксико-семантическое представление текста или отдельных предложений.

Тональность — это эмоциональное отношение автора высказывания к некоторому объекту (объекту реального мира, событию, процессу или их свойствам/атрибутам), выраженное в тексте. Эмоциональная составляющая, выраженная на уровне лексемы или коммуникативного фрагмента, называется лексической тональностью (или лексическим сентиментом). Тональность всего текста в целом можно определить как функцию (в простейшем случае сумму) лексических тональностей составляющих его единиц (предложений) и правил их сочетания[2].

Извлечение сущностей из текста (Entity Extraction).

Типичная задача извлечения информации: просканировать набор документов, написанных на естественном языке, и наполнить базу данных выделенной полезной информацией.

С данной задачей на данный момент доволе неплохо справляются комерческое решение Rosette и целая система библиотек для NLP – Spacy. Правда оба решения имеют неплохой процент качества работы именно с текстами на английском языке, хотя и поддерживают другие языки. Еще одна проблема – большой объем ресурсов, необходимый для работы Spacy. Один «поднятый» instance Spacy потребляет 2 Гб оперативной памяти. Так как в рамках приложения решается задача возможности заведения для каждого клиента своего словаря – появляется проблема с количеством доступной оперативной памяти.

Обучение классификатора. SVM.

В общем, задача машинного обучения сводится к получению набора выборок данных и, в последствии, к попыткам предсказать свойства неизвестных данных. Если каждый набор данных — это не одиночное число, а например, многомерная сущность (multi-dimensional entry или multivariate data), то он должен иметь несколько признаков.

Машинное обучение представляет собой обучение выделению некоторых свойств выборки данных и применение их к новым данным. Вот почему общепринятая практика оценки алгоритма в Машинном обучении — это разбиение данных вручную на два набора данных. Первый из них — это обучающая выборка, на ней изучаются свойства данных. Второй — контрольная выборка, на ней тестируются эти свойства.

SVM. Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Итогом проделанной работы стала рекомендательная система, ранжирующая новости в соответствии с требованиями пользователя-клиента данного сервиса. Разработаны бизнес-правила, разработана и реализована «гибкая» архитектура приложения, позволяющая без исправления кода настраиваться под требования нового клиента.

Список использованных источников:

1) Building Machine Learning Systems with Python. Luis Pedro Coelho, Willi Richert. Построение систем машинного обучения на языке Python. Луис Педро Коэльо, Вилли Ричарт.

2) Machine Learning with Spark: Create scalable machine learning applications to power a modern data-driven business using Spark. NickPentreath.

ЗАЩИТА ДАННЫХ В ОБЛАЧНОМ ХРАНИЛИЩЕ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Сыч Ю.В.

Прохорчик Р.В. – ассистент каф. ПОИТ, м.т.н.

Облачные хранилища данных – сервисы, предоставляющие возможность хранить свои файлы на удаленных серверах, а также получать к ним доступ из любой точки мира, где есть доступ в Интернет. В условиях роста необходимости хранения информации, данные сервисы стали популярны среди пользователей. В настоящее время множество компаний также хранит большую массу информации, необходимая для принятия решений. Она накапливается в различных источниках и хранилищах, превращаясь в опыт предприятия. Однако, отправляя свои файлы в облачное хранилище, пользователи не задумываются о защите информации в облаках от доступа третьих лиц. Стоит ли всецело доверяться держателям облачных сервисов, полагаясь на предпринимаемые ими меры защиты информации – личное дело каждого.

Опасность хранения файлов в облачном хранилище заключается в том, что пользователь передаёт файлы на хранение третьему лицу с неизвестными, по отношению к пользователю, намерениями. А опасность именно файлов, как объекта информации, в том и заключается, что с него можно сделать копию и

пользователь об этом факте никак не узнает. Данные компаний о заказчиках, счётах, денежных оборотах, хранящиеся в облаке, могут быть подвержены утечке, вследствие чего компании могут оказаться в крайне плохом положении.

В качестве защиты данных в облачном хранилище поставлена задача создать расширение, которое способно шифровать данные при их загрузке в сервис, а также обратная операция – расшифровка данных при скачивании из сервиса. Пользователь задаёт ключ и алгоритм для шифрования. Для создания расширения выбран барузер Chrome, а в качестве сервиса облачного хранилища используется Dropbox. Основным алгоритмом шифрования выбран aes (симметричный алгоритм блочного шифрования), обеспечивающий высокую скорость шифрования данных. В данный момент реализовано:

- Шифрование данных при загрузке на сервис перетягиванием (draganddrop) файлов в облако;
- Расшифровка данных при скачивании с главной страницы сайта;
- Выбор алгоритма шифрования у пользователя;
- Сохранение ключа для шифрования пользователя.

Принцип работы расширения для защиты облака:

- При попытке загрузки в сервис (скачивании из сервиса) данных, запрос на загрузку будет отменён;
- Полученные данные шифруются выбранным алгоритмом шифрования ключом пользователя;
- Формируется запрос отправки/скачивания зашифрованных данных;
- Отправка запроса (загрузка/скачивание данных).

Основные преимущества расширения для защиты облака:

- работа на всех популярных операционных системах, таких как: Windows, Mac, Linux;
- поддержка русского языка – управлять расширением просто, а все функции понятны;
- ключ и алгоритм для шифрования имеется только на стороне клиента, таким образом, третьи лица практически теряют возможность получить доступ к данным пользователя.

Основные недостатки расширения:

- данные шифруются непосредственно перед отправкой или загрузкой, а не на лету. Таким образом необходимо ожидать шифрования(расшифровки) данных, а затем отправку или скачивание, вследствие чего теряется скорость работы с облачным хранилищем;
- программа может работать только в браузере Chrome версии 4.0 и выше.

В дальнейшем необходимо реализовать шифрование данных при загрузке на сервис через форму облачного хранилища, расшифровку данных при скачивании из формы открытого файла, создание вспомогательного окна процесса шифрования.

В целом при совершенствовании расширения необходимо сформировать адаптивность для различных барузеров (IE, Firefox, Opera) и возможность шифрования для других сервисов облачных хранилищ, таких как Google Диск, OneDrive, Яндекс.Диск.

Список использованных источников:

1. Статья Способ удобного шифрования данных в облаке (собственными средствами) // Habrahabr [Электронный ресурс]. – Режим доступа: https://habrahabr.ru/post/241720/

2. Статья Шифрование данных в облаке, 2016: Исследование Gemalto и Ponemon [Электронный ресурс]. – Режим доступа: http://www.tadviser.ru/index.php/Статья:Шифрование_данных_в_облаке

СТАТИСТИЧЕСКИЙ ПОДХОД К УПРАВЛЕНИЮ РЕСУРСАМИ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Толкачёв А.В.

Куликов С.С. – к.т.н., доцент

В настоящее время облачная модель поставки и функционирования сетевых приложений и сервисов является доминирующей на рынке и продолжает демонстрировать стабильный рост. Использование методов статистического моделирования для управления ресурсами облачного вычислительного кластера может значительно повысить эффективность функционирования системы и минимизировать накладные расходы на балансировку нагрузки.

Облачные вычисления – подход, при котором ресурсы поставляются в виде услуги, могут быть сданы в аренду и предоставлены пользователям через сеть по запросу. Для максимально эффективной утилизации вычислительных ресурсов активно используются технологии виртуализации. Общая схема облачного вычислительного кластера приведена на рисунке 1.

В рамках одного физического сервера (хоста) при помощи специального программно-аппаратного комплекса (гипервизора) функционирует множество виртуальных машин, выступающих в качестве среды функционирования ПО. Гипервизоры связаны с системой мониторинга, ответственной за динамическое перераспределение нагрузки между физическими хостами. Трансфер виртуальной машины между хостами является достаточно ресурсоёмкой операцией, так как требует дополнительные сетевые ресурсы и вычислительные ресурсы для синхронизации [1]. Управление ресурсами, базирующееся на основе модели статистических предсказаний, позволяет минимизировать накладные расходы на миграцию виртуальных