OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.912

ИСПОЛЬЗОВАНИЕ СОЦИАЛЬНЫХ СЕТЕЙ ДЛЯ ПОИСКА КОМПЕТЕНТНЫХ ГРУПП ЛЮДЕЙ

Стратнев П.Ю.

Югорский государственный университет, г. Ханты-Мансийск, Россия
Pavel.Stratnev@gmail.com

В статье рассмотрен метод поиска определённых групп пользователей, имеющих компетентность в некоторой области деятельности.
Ключевые слова: социальные сети, классификация, компетентность, алгоритм.

Введение

За последние несколько лет уровень информатизации общества резко возрос. Одним из аспектов этого роста является развитие такого информационного ресурса как социальные сети. Подобные ресурсы представляют собой автоматизированную социальную среду для обеспечения коммуникации как отдельных, так и групп пользователей, объединенными общими интересами. Таким образом, данные Web-сайты формируются за счет пользовательского контента, и могут по праву считаться удобным средством обмена информацией. Имея в своём составе богатую информационную базу, а также, учитывая тот факт, что количество пользователей в этих сетях увеличивается с огромной скоростью [Boyd et al., 2007], такие сервисы могут использоваться в качестве источника актуальных и уникальных данных.

На сегодняшний день, в условиях сложившейся демографической ситуации обусловленной кризисом 90-х годов, имеет место проблема обеспечения организаций достаточным количеством наиболее подготовленного контингента. Профориентационная работа в школах направлена на решение данной проблемы, но в масштабах населения России требуются дополнительные инструменты и возможности разрешения этой задачи.

Данная статья посвящена поиску компетентных групп людей на основе данных, полученных из социальной сети.

1. Система анализа данных

1.1. Сбор данных из социальной сети

Интерфейс социальной сети является источником данных реального времени и предназначен для просмотра и взаимодействия со страницами социальной сети в веб-браузере. Для использования данных пользователей в специализированных приложениях ресурс предоставляет определённый набор функции (API) для получения данных из своей информационной базы. Поскольку сценарии использования интерфейсов социальных сетей не предполагают автоматического сбора данных множества пользователей, то возникает ряд проблем:

- *приватность данных* – чаще всего доступ к данным пользователей разрешён только для зарегистрированных и авторизованных участников сети, что требует поддержки эмуляции пользовательской сессии с помощью специальных учетных записей (*аккаунтов*);
- *слабая структурированность данных* – во многих случаях API социальных сетей имеют ограниченный функционал, что требует поддержки получения с помощью пользовательского веб-интерфейса статистических копий HTML-страниц, корректной обработки их динамической части (включая исполнение асинхронных запросов к серверу социальной сети), извлечения нужных данных с помощью алгоритма и/или шаблона и построения их структурированного представления, удобного для дальнейшей автоматической обработки;
- *ограничение доступа и блокировки* – с целью предотвращения несанкционированного автоматического сбора данных и ограничения нагрузки на инфраструктуру сервиса социальной

сети администраторы сервисов зачастую вводят явные или скрытые ограничения на допустимое количество запросов от одного пользовательского аккаунта и/или IP-адреса в единицу времени, что требует учёта количества посылаемых запросов, а также поддержки динамической ротации используемых для сбора данных пользовательских аккаунтов и IP-адресов;

- *размерность данных* обуславливает необходимость в параллельном методе сбора данных, а также в методах получения репрезентативной выборки пользователей социальной сети.

В связи с необходимостью постоянного получения больших наборов данных из социальной сети, была разработана программа реализации этой функции.

Созданный инструмент поддерживает скачивание данных из социальной сети «Вконтакте», пользующеюся огромной популярностью в русском сегменте сети Интернет. Загрузка данных происходит постранично, с последующей обработкой элементов внутренней разметки страницы. Та часть элементов, получение значений которых, реализовано через API сервиса, автоматически структурируется и вносится в локальную базу данных приложения. Кроме функций скачивания, приложение реализует механизм автоматического выбора учетной записи социальной сети для каждого запроса, а также поддержка прокси-соединений. Это обеспечивает устойчивость к блокировкам по IP-адресам и учетным записям.

Для оценки производительности инструмента были проведены эксперименты, в которых скачивались и записывались в базу данных профили пользователей социальной сети «Вконтакте». В среднем производительность достигала 250 пользователей в час.

1.2. Предварительная обработка данных

Так как вся основная информация размещённая на веб-ресурсе представлена в виде HTML-страницы, то логично предположить что большая часть данных является обычным текстом. Для дальнейшего использования информационной базы социальной сети в рамках данного проекта, требуется предварительная обработка собранных данных, которая в первую очередь включает в себя непосредственное выделение информации, признаков и параметров со страницы учетной записи. Это позволит внести некоторую структурированность в систему данных.

Модель «Множество слов» (bag-of-words) – упрощающее предположение, использующееся в обработке естественного языка и поиске информации (information retrieval). В этой модели, текст (предложение или документ) представляется как неупорядоченный набор слов, без учета грамматики и порядка слов в частности. Для этого необходимо убрать HTML-теги, знаки препинания, стоп-слова, все слова привести к нижнему регистру [Lee et al., 2010].

Для определения базиса пространства требуется получить все основы слов в используемом наборе текста. Процесс нахождения основы слова для заданного исходного слова называется *стемминг*.

Существует много алгоритмов решающих поставленную задачу. В результате анализа был выбран стеммер Портера. Главным плюсом этого метода заключается в том, что он не использует никаких словарей и выделение основы осуществляется путём преобразования слова согласно определённым правилам [Porter, 1980]. К тому же, алгоритм поддерживает работу с русским языком.

1.3. Классификация профилей пользователей

Задача классификации представляет собой задачу отнесения образца к одному из, как минимум, двух попарно не пересекающихся множеств.

Пусть X – множество описаний объектов, Y – множество наименований (номеров) классов.

$$y : X \to Y . \quad (1)$$

Существует неизвестная целевая зависимость – отображение, значение которой известны только на объектах конечной обучающей выборки X^m, по которой производится настройка модели.

$$X^m = \{(x_1, y_1), ..., (x_m, y_m)\}. \quad (2)$$

Требуется построить алгоритм a, способный классифицировать произвольный объект $x \in X$.

Пусть дан набор пользователей n пользователей P.

$$P = \{p_1, p_2, ..., p_n\}. \quad (3)$$

Каждый пользователь p_i имеет профиль u_i и связанный с ним контент. Профиль представляет собой страницу, которая создаётся и контролируется пользователем.

Задача заключается в определении компетентности пользователя p_i в математике, русском языке, иностранном языке, литературе и т.п. (область знаний зависит от обучающей выборки) с помощью классификатора c, используя u_i.

$$c : u_i \to \{1,0\} . \quad (4)$$

Классификатор c даёт прогноз, является ли пользователь p_i компетентным кадром. Для

построения классификатора необходимо определить, какие параметры влияют на принятие решения о том, принадлежит ли объект классу. Для этого необходимо иметь достаточно размерность пространства признаков (количество параметров входного вектора).

Точность классификатора можно оценить с помощью тестовой выборки с заранее известными данными о принадлежности элементов к классу. Оценку качества классификации, сделанную по тестовой выборке, можно применить для выбора наилучшей модели или дальнейшей модификации.

1.4. Структура приложения

Эксперимент требовал создания дополнительного программного обеспечения реализующего алгоритм классификации, и поэтому была разработана архитектура системы, представленная на рис. 1, состоящая из 3 основных узлов, каждый из которых выполняет соответствующую функцию.

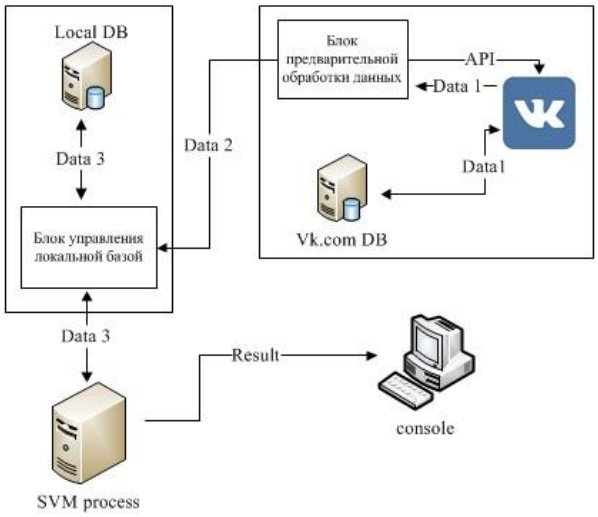

Рисунок 1 – Общая структура приложения

Любая социальная сети (в данном случае - «Вконтакте») имеет в своём составе закрытую информационную базу, хранящую все профили своих пользователях. Так как обращение к базе данных напрямую не доступно, то параметры необходимые для непосредственной классификации, или информация необходимая для вычисления определённых атрибутов выделяет из многопользовательской системы с помощью API, которое предоставляет сервис vk.com.

Блок предварительной обработки данных включает в себя модуль сбора информации и реализует предварительную обработку данных. В качестве входных параметров используется поток данных data1, который содержит данные в формате целевого сервиса. Обработка параметров включает вычисление некоторых атрибутов, преобразование и передачи данных в блок управления локальной базой (data2).

Блок управления локальной базой берёт на себя функции работы с главной базой данных приложения, процедуру преобразования данных к векторной форме, необходимой для работы алгоритма классификации и непосредственно передачи преобразованных данных в узел SVM process (data3). Самостоятельность модуля необходима для обеспечения расширяемости системы в целом.

Важной частью системы является блок SVM process, в котором представлен алгоритм машинного обучения, который может работать с любыми данными из локальной базы данных. Модуль не зависит от набора данных data1, что делает его универсальным, предоставляя возможности по обработке данных из других многопользовательский сетей. Необходимым и достаточным условием станет лишь корректировка или добавление блока предварительной обработки данных для соответствующей системы.

Результаты работы приложения выводят в консоль.

1.5. Метод исследования

Анализ показал, что среди методов машинного обучения высокие перспективы имеет метод опорных векторов [Воронцов, 2007]. Этот метод позволяет добиться высокого качества в области классификации.

Метод опорных векторов относится к бинарным классификаторам, которые определяют только принадлежность (или не принадлежность) объекта к классу, но не сам класс. Иными словами полностью соответствуют выражению (4). Суть самого метода заключается в поиске гиперплоскости разделяющей два класса точек в n-мерном пространстве. Обучающая выборка задаёт целевую зависимость, значения которой заранее известны. Требуется построить алгоритм, аппроксимирующий целевую зависимость на всём пространстве.

Для построения классификатора требуется определение пространства признаков, по которым будет проходить классификация пользователей. Профиль пользователя будет представлен в качестве вектора большой размерности. Элементы вектора формируются путём предварительной обработки данных, полученных со страницы пользователя социальной сети.

Исследования проводились с использованием компьютера на базе Intel Core i5 с 4 Гб оперативной памяти. В качестве основного инструментария для проведения классификации использовался язык программирования python и библиотека libsvm реализующая работу метода. Благодаря этим программным средствам поставленная задача свелась к формированию обучающей выборки (настройка классификатора) и непосредственно классификация контрольной выборки.

1.5.1. *Настройка классификатора*

Обучение классификатора будет строиться на выборке, в которую войдёт часть профилей пользователей, которые обучаются (или обучались) в Институте (НОЦ) систем управления и информационных технологий Югорского государственного университета. В данном случае принадлежность к классу будет определять статус «студент института». Таким образом, мы получаем выборку вида:

$$V^i = \{(v_1,1),...,(v_i,1)\} \quad (5)$$

где v_i - вектор, элементы которого являются предварительно обработанные данные, полученные из профиля i - го пользователя студента. Предполагается, что 100 профилей будет достаточно для построения модели.

1.5.2. *Тестирование точности*

После получения настроенного классификатора, оценим его точность на новой выборке людей, по характеру отбора схожей с обучающей выборкой. Отличительной особенностью будет лишь то, что элементы этих множеств не будут пересекаться. Таким образом, мы имеем массив профилей, с заранее известной характеристикой принадлежности к классу.

Из 100 опытов, классификатор спрогнозировал 71 верный случай принадлежности объекта к классу. Точность 71% является допустимой в рамках эксперимента, однако стоит обратить внимание на один немало важный факт. Так как профили в социальной сети контролируются людьми и вся информация полученная из таких профилей является собственноручно заполненной, есть вероятность получить погрешности при построения модели, ввиду возможным нежеланием пользователей указывать достоверную информацию. Это, пожалуй, самый важный недостаток в использование информационной базы многопользовательской сети.

2. Перспективы развития системы

Первоначальной идеей создания системы являлась поддержка принятия решения в приёмных кампаниях высших учебных заведений. Ввиду того что эта область применения довольна требовательна к точности прогноза, в настоящее время использовать систему не представляется возможным. Достигнув точности классификации в 90%, возможно, этот метод станет востребованным.

В целом, системы классификации решают огромное число задач статистического характера. В частности, адаптация предложенного метода классификации может решать не только задачу определения компетентности пользователей социальной сети, но так же может помочь получить результаты в анализе тональности сообщений пользователей и/или сообществ и выделение необходимых сегментов пользователей. Кроме того, оперируя уникальными данными, подобные проекты найдут применение и в сфере разработок искусственного интеллекта.

3. Заключение

После своего появления социальные сети развивались в основном количественным путём, охватывая всё больше пользователей, сегодня же социальные сети переходят в стадию качественного развития, изобретая всё новые инструменты взаимодействия с пользователями. Поэтому технологическое развитие станет необходимым условием для выживания в конкурентной среде. Развитие технологий будет многогранным, но даже сейчас можно с уверенностью заявить развитие инструментов работы с контентом станет одной из ключевых тенденций [Семенов, 2011].

Представленный подход поддерживает расширяемость в плане количества источников информации. Добавление дополнительных блоков обработки данных положительно скажется на точности классификации пользователей. Кроме того, точность системы может быть увеличена путём преобразования данных к единому виду для корректного использования с другими методах машинного обучения.

В данной статье предложен метод обработки данных из социальной сети с целью получения компетентных групп пользователей. Рассмотрена архитектура системы исследования и классификации пользователей. Изложенная идея представляет собой законченный эксперимент, результаты которого были описаны в п.1.4.2.

Библиографический список

[Boyd et al., 2007] Boyd. D. Social Network Sites: Definition, History, and Scholarship / D. Boyd [et al.] // Journal of Computer-Mediated Communication. – 2007 – Vol.13 №1 – P.210-230.

[Lee et al., 2010] Uncovering Social Spammers: Social Honeypots+Machine Learning / K. Lee , J. Caverlee, S. Webb // Proceedings of the 33rd Annual ACM SIGIR Conference (SIGIR 2010). – July 2010, Geneva, Switzerland. – Geneva, 2010

[Porter, 1980] Porter M.F. An Algorithm for Suffix Stripping / Porter M.F. // Program: electronic library and information systems 14 (3). – P.130-137

[Воронцов, 2007] Воронцов К.В.. Лекции по методу опорных векторов [Электронный ресурс] / К. В. Воронцов // Режим доступа: http://www.ccas.ru/voron/download/SVM.pdf

[Семенов, 2011] Семенов Н. Все о социальных сетях. Перспективы развития [Электронный ресурс] / Н. Семенов // Режим доступа: http://secl.com.ua/article-vse-o-socialnyh-setjah-perspectivy-razvitija.html

USING SOCIAL NETWORKS FOR FINDING QUALIFIED GROUPS OF PEOPLE

Stratnev P.U.

Ugra State University, Khanty-Mansiysk, Russia

Pavel.stratnev@gmail.com

The article describes the method of searching for specific user groups with expertise in a certain field of activity.