СЕРВИС ДЛЯ ИССЛЕДОВАНИЙ ПРОФЕССИОНАЛЬНЫХ ИНТЕРЕСОВ СТУДЕНТОВ В ИНТЕРНЕТ

Ю. П. Парфенов, Н. В. Рогачева
Институт радиоэлектроники и информационных технологий
Уральский федеральный университет имени первого Президента России Б. Н. Ельцина
Екатеринбург, Россия
u.p.parfenov@urfu.ru

Аннотация. Предлагаемый сервис предназначен для исследований интересов любой группы пользователей путем анализа баз данных, формируемых Интернет-браузерами. Задача решается путем автоматической классификации запросов и заголовков посещаемых сайтов предварительно обученной моделью. Универсальность инструмента обеспечивается настройкой иерархического классификатора интересов в соответствии с целями исследования.

Ключевые слова: *Интернет; классификация интересов пользователей; обучаемые модели*

Доступные интернет-сообществу продукты для мониторинга активностей пользователей в основном ориентированы на задачи SEO-оптимизации и анализа сайтов конкурентов. Вместе с тем, Интернет дает большие возможности [1] для анализа и понимания преобладающих интересов разных групп пользователей, в частности социальных и профессиональных интересов и потребностей студентов.

Предлагаемый сервис предназначен для проведения исследований интересов любой группы пользователей путем анализа баз данных, формируемых Интернет-браузерами. Использование браузера вместо традиционных методов социологического опроса и анкетирования обеспечивает объективность исходных данных и не создает дополнительную нагрузку на исследуемую группу.

Хранящаяся в базе браузеров история посещения страниц веб-сайтов содержит детальные сведения об активности респондента в Интернет, а заголовки посещенных сайтов и тексты запросов к поисковым системам могут служить источником данных для анализа интересов. Целенаправленный отбор и анализ получаемой из Интернет информации обеспечивается использованием в сервисе настраиваемого для конкретного исследования иерархического классификатора интересов.

Задача изучения интересов пользователей Интернет решается путем автоматической классификации истории запросов и посещаемых веб-сайтов предварительно обученной моделью. Перед началом исследования создается классификатор, содержащий иерархию групп изучаемых интересов. Затем на предварительной выборке обучается модель автоматической классификации. Создание произвольного классификатора обусловливает возможность выяснения определенных интересов для любых групп пользователей.

Исследования выполняются на условиях анонимности и добровольности путем предоставления доступа к данным выполненных Интернет-запросов и обращений к веб-сайтам из браузера респондента. При этом респонденту предоставляется возможность удаления из анализа (фильтрации) нежелательных, по его мнению, сведений.

Общая схема подготовки и проведения исследования представлена на рис. 1.

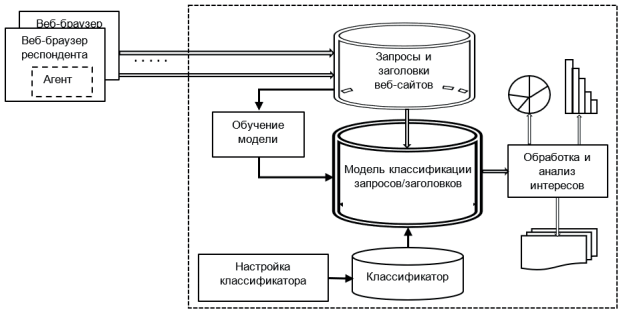

Рис. 1. Схема подготовки и проведения исследования

Респонденты в дополнение к веб-браузеру устанавливают «тонкое» клиентское приложение-агент, которое на интервале исследования систематически передает историю браузера в базу сервиса. Таким образом, для анализа доступны времена обращения, веб-адреса и заголовки посещаемых сайтов, а также тексты исполняемых запросов. При этом весь процесс сбора данных, их систематизация и анализ выполняется «прозрачно» для респондентов.

С целью более детального анализа и систематизации интересов по обобщающим группам в классификаторе они задаются трехуровневыми древовидными графами. Листовые вершины графа

задают изучаемые интересы респондентов, по которым собирается информация об их активности в Интернет, а вершины верхних уровней предназначены для обобщений результатов и представления данных с разной степенью детализации. Названия вершин и структура графа определяется целями проводимого исследования. В сервисе создан интерактивный графический редактор, позволяющий строить, изменять и сохранять классификаторы в своей базе данных. Пример части возможного классификатора интересов студентов представлен на рис. 2.

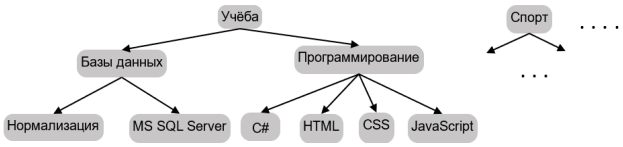

Рис. 2. Пример классификатора исследуемых интересов

Для автоматической классификации интересов по текстам запросов и заголовков веб-сайтов в сервисе используются модели, обучаемые учителем, проводящим исследование. Обучение проводится в диалоговой форме путем сопоставления конкретного запроса или сайта из обучающей выборки с определенным интересом, внесенным в классификатор.

Ввиду обычно малого объема обучающей выборки применение нейросетей в настраиваемой модели классификации не рассматривалось. В сравнении для использования в сервисе участвовали только простые методы классификации тексов, ориентированные на короткие выборки и реализованные моделями языка Python. Использовалось четыре претендента: k ближайших соседей, случайный лес, полиномиальный наивный Байес и логистическая регрессия.

Так как реальные выборки по классифицируемым интересам скорее всего не сбалансированы, то не учитывающая этот фактор мера Accuracy для оценки качества предсказания не использовалась. Качество моделей оценивалось следующими типовыми метриками:

— precision (точность), представляющая долю правильно определённого интереса среди всех отнесенных к данному интересу,

— recall (полнота) показывает, долю правильно определенного интереса среди фактических запросов и сайтов, относящихся к этому интересу,

— F1 score (интегрированный показатель) представляет собой гармоническое среднее между точностью (precision) и полнотой(recall).

Для эксперимента выбора наиболее точного метода был построен классификатор интересов, содержащий 23 класса в трех разных уровнях иерархии и создан набор данных из 1059 запросов и обращений к сайтам, соответствующим заданным интересам. Количество экспериментальных запросов по всем 23 классам интересов показано на рис. 3.

В ходе эксперимента для сравнения моделей классификации весь набор данных был случайным образом разделен на наборы для обучения — основная выборка (75 %) и тестирования (25 %). На основной выборке были обучены все четыре модели.

Затем с помощью этих моделей была проведена классификация запросов и сайтов для тестовой выборки. Далее, сравнивая предсказанные моделями классы интересов и советующие им фактические классы, рассчитаны метрики качества проверяемых моделей.

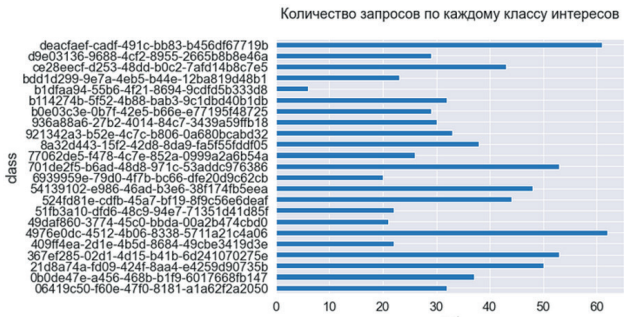

Рис. 3. Распределение количества запросов по классам интересов

Для каждого интереса, заданного в классификаторе, вычислялись значения для precision, recall, и F1 score-мер [2], а также количество запросов и сайтов в каждом классе.

Для сравнения качества классификации использованы средние оценки F1 score, для которой получены следующие результаты:

— k ближайших соседей — 0.63,

— случайный лес (Random forest) — 0.77,

— полиномиальный наивный Байес — 0.72,

— логистическая регрессия — 0.79.

Наилучшие результаты показали модели, использующие методы случайного леса и логистической регрессии со значениями оценки F1 score 0,77–0,79. Поэтому в сервисе используются обе модели. В случае несовпадения результата классификации выбирается класс, у которого точность (precision) предсказания в целом выше.

Сервис реализуется в виде WPF-приложения на языке Python с использованием библиотек машинного обучения. Классифицируемые запросы

и заголовки сайтов хранятся в базе под управлением SQLite.

Для окончательного анализа обработанных моделью и сохраненных в базе результатов исследователю предоставлен интерактивный конструктор запросов, позволяющий систематизировать данные по разным группам интересов (узлам классификатора) и отображать их в виде деловой графики.

Библиографический список

1. Д. О. Стребков Социологические опросы в Интернете: возможности и ограничения. Материалы Интернет-конференции Социология и Интернет: перспективные направления исследования. https://iq.hse.ru/more/sociology/sociologicheskie-oprosi-v-internet (дата обращения 12.10.2021)

2. Пратик Джоши. Искусственный интеллект с примерами на Python. Создание приложений искусственного интеллекта / П. Джоши. — М.: Вильямс, 2019. — 448 с.