УДК 004

Исследование методов и реализация алгоритма автоматического поиска официальных сайтов организаций по косвенным данным

Терентьев Александр Владимирович
Волжский политехнический институт (филиал) «Волгоградский государственный технический университет
студент

Короткова Неля Николаевна
Волжский политехнический институт (филиал) «Волгоградский государственный технический университет
к.т.н., доцент

Рыбанов Александр Александрович
Волжский политехнический институт (филиал) «Волгоградский государственный технический университет
к.т.н., доцент, заведующий кафедрой информатики и технологии программирования

Аннотация
В данной статье приведён обзор методов и реализаций алгоритмов автоматического поиска официальных сайтов организаций по косвенным данным. Указаны аналогичные исследования в этой области. Перечислены актуальные проблемы, связанные с алгоритмом поиска по косвенным данным. Также представлен план необходимых действий по реализации алгоритма автоматического поиска. Описана система сбора данных об организациях из всемирной сети.
Ключевые слова: автоматический поиск, косвенные данные, машинное обучение

Research of the methods and implementation of the algorithm of automatic search for the official sites of organizations by indirect data

Terentyev Alexander Vladimirovich
Volzhsky Polytechnic Institute (branch) of Volgograd state technical University
student

Korotkova Nelya Nikolaevna
Volzhsky Polytechnic Institute (branch) of Volgograd state technical University
Ph. D., associate Professor

Rybanov Alexander Alexandrovich
Volzhsky Polytechnic Institute (branch) of Volgograd state technical University
Ph. D., associate Professor, head of the Department of Informatics and programming technology

Abstract
This article provides an overview of the methods and implementations of the automatic search algorithms for the official websites of organizations using indirect data. Similar studies are indicated in this area. The current problems associated with the indirect search algorithm are listed. A plan of the necessary actions to implement the automatic search algorithm is also presented. A system for collecting data about organizations from the world wide web is described.
Keywords: automatic search, indirect data, machine learning

Алгоритмы поиска необходимой информации в большом объеме данных являются наиболее востребованными в наше время. Работа с большими данными относиться к бурно растущему направлению информационных технологий. Известные компании, такие как Яндекс, Google, Facebook, Twitter и прочие, ежедневно обрабатывают огромные массивы информации для решения своих бизнес-задач. В работу с большими объемами данных входят все принципы построения информационных систем: анализ и хранение данных, координация сервисов, планирование задач, эффективность использования [1].

Объектом нашего исследования является поиск официальных сайтов организаций, а значит система должна по косвенным данным обнаруживать, желательно без ошибок, необходимую информацию. Помимо того, что работа будет проводиться с большими данными, так еще и с использованием машинного обучения [2]. Наша система будет обучаться на основе полученных ею данных. В этой области есть наработки, которые следует упомянуть.

Реализация алгоритма поиска очень важна с практической точки зрения. Существует множество проблем, которые приходится решать для получения достоверных результатов поиска. Одной из таких проблем является выбор нужного элемента данных по косвенным данным. Такой элемент, в самом общем виде должен обладать следующими свойствами:

1. Точность полученных данных;
2. Высокое качество полученных данных;
3. Доступность данных.

Ниже представлены некоторые результаты сравнения методов и алгоритмов машинного обучения на задаче поиска информации по косвенным признакам.

В статье «How Search Engines Use Machine Learning: 9 Things We Know for Sure» [6] сравниваются некоторые методы и сигналы, помогающие найти достоверную информацию в интернете. Методы Байса и Роше косвенно указаны, но какой из них приоритетней использовать зависит от постановки задачи.

При прочтении статьи «How Machine Learning in Search Works: Everything You Need to Know» [5] стало понятно, что такой поисковой гигант, как Google пользуется своими наработками в области машинного обучения. Программный продукт данной компании RankBrain, позволяет ранжировать

результаты поиска, устанавливая на самом верху самые релевантные результаты запроса.

Результатом статьи «The 10 Algorithms Machine Learning Engineers Need to Know» [7] стал обзор всевозможных методов и алгоритмов машинного обучения. Неясно точно какой подойдет для нашей конкретной задачи, но сравнение помогло расставить приоритеты в выборе математической модели.

Методы алгоритмов поиска данных, основанных на обучении, впервые введены в рассмотрение в 1960-е годы. В наше время разработано множество методов машинного обучения, которые применяются для решения широкого спектра задач. Многие из них применяются для поиска необходимых данных в большом объеме информации. Ниже перечислены основные методы машинного обучения, которые подходят для нашей задачи.

Метод Байеса, основанный на теории вероятности, в самом общем виде гласит, что вероятность какого-либо события можно определить, если произошло другое статистически взаимосвязанное с ним событие. Этот метод основан на анализе совместных распределений признаков поиска схожих объектов. Если подвести к нашей задаче, то если мы находим верную информацию при определенном наборе косвенных данных, тогда последующие найденные данные будут также верны. Метод Байеса обладает высокой скоростью работы и простотой математической модели. Этот метод часто используют в качестве базового метода при сравнении различных методов машинного обучения.

Алгоритм k-ближайших соседей это метрический алгоритм для автоматической классификации объектов поиска. Данный метод, в отличие от других, не требует фазы обучения. Для того чтобы найти искомые данные, предположенный результат сравнивают со всеми верными данными из обучающей выборки. При верном соответствии данный результат приравнивается к верному, что укрепляет обучающую выборку. Данный метод показывает довольно высокую эффективность, но требует довольно больших вычислительных затрат на этапе сравнивания.

Классификатор Роше является наиболее простым классификатором для поиска данных, основанных на векторной модели. Для каждого верного поиска строится взвешенный центроид по формуле. В дальнейшем это помогает лучше и быстрее найти нужные данные. Классификатор обладает полезной особенностью: взвешенные центроиды можно быстро пересчитать при добавлении новых косвенных данных. Эта особенность полезна, например, когда пользователь указывает какие сайты система нашла правильно, а какие нет. Существует множество различных модификаций данного метода.

Нейронные сети — это большой класс систем, архитектура таких систем построена по аналогии с нервной тканью из нейронов. Нейроны являются наборами данных, соединенных между собой. Каждый такой нейрон по сути обычный преобразователь, получающий на вход некие данные и выводящий результат. Выходные результаты вычисляются как

функция о входных сигналах. Т.е. подавая некий набор параметров на вход сети, мы получаем какой-то набор чисел на выходе. В результате работа сети сводиться в преобразование входного вектора данных в выходной вектор, причем это преобразование задается весами на преобразователях. Как уже указывалось выше, нейронные сети имеют очень высокий спектр применения, но минус в том, что на обучение тратится очень много времени, связанно это с введением большого количества узлов для задач высокой размерности.

Деревья решений используются в поиске информации, при помощи обучения. Существует несколько алгоритмов для обучения, например, CLS. Данный алгоритм циклически разбивает обучающие примеры на классы в соответствии с переменной, имеющей наибольшую классифицирующую силу. В ходе такой разбивки образуется дерево решений. Но большой проблемой является сложность интерпретации результатов поиска таким методом.

Построение булевых функций строит правила выбора достоверной информации в виде «если выполняется формула, то данные А». Построение таких функций строится на методе деревьев решений, что касается построения правил вывода. Также, к построению правил классификации можно отнести методы ограниченного перебора для правил заданного вида. Данный метод доказал свою эффективность, но не подходит для нашей задачи. В дальнейшем во второй главе мы опишем разработанный алгоритм поиска нужной информации, использующий частичный перебор вариантов.

Метод опорных векторов разработан В. Вапником на основе принципа структурных минимизаций риска – одновременного контроля количества ошибок классификации на множестве для обучения и степени обнаружения зависимостей. Данный метод позволяет работать с абстрактной векторной моделью данных. Это позволяет применять его для решения различных задач машинного обучения. Преимущественно используется для задач распознавания образов, распознавания речи, классификации текстов [4].

В данной работе необходимо реализовать описание, математическую модель и программный продукт «поисковика» официальных сайтов компаний и организаций по их косвенным данным. Такой «поисковик» может является инструментом, который работает через взаимодействия с сервисами поиска сайтов по тексту. Данный программный продукт может быть своего рода «посредником» между интернет-поисковиками и пользователями. Предполагается, что «Поисковик» будет находить официальные сайты путем анализа выдачи, в частности выданных сайтов. При проектировании продукта необходимо использовать технологии машинного обучения, а в частности машиннообученные формулы. В процессе предусматривается возможность интеграции инструмента с MapReduce-системой. Тем самым переводя задачу поиска официального сайта в MapReduce-задачу. Предполагается, что помимо этого, «поисковик» будет работать и локально [3].

Библиографический список

1. Артюхин В.В Применение методов машинного обучения при работе с литературными источниками. М.: Доклад, 2015. 17 с.
2. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич.В., СапинА.С. Автоматическая обработка текстов на естественном языке и анализ данных. М.: Изд-во НИУ ВШЭ, 2017. 269 с.
3. Hadoop: Подробное руководство. СПб.: Питер, 2013. 672 с.
4. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques / Morgan Kaufman Publishers, 2011.
5. How Machine Learning in Search Works: Everything You Need to Know https://www.searchenginejournal.com/how-machine-learning-in-search-works/257837/
6. How Search Engines Use Machine Learning: 9 Things We Know for Sure https://www.searchenginejournal.com/how-search-engines-use-machine-learning/224451/
7. The 10 Algorithms Machine Learning Engineers Need to Know https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article