§ 5 СИСТЕМНЫЙ АНАЛИЗ, ПОИСК, АНАЛИЗ И ФИЛЬТРАЦИЯ ИНФОРМАЦИИ

Менщиков А.А., Комарова А.В., Гатчин Ю.А., Полев А.В.

РАЗРАБОТКА СИСТЕМЫ АВТОМАТИЧЕСКОГО КАТЕГОРИРОВАНИЯ ТЕМАТИКИ СТРАНИЦ ВЕБ-РЕСУРСА

Аннотация: *В данной статье рассматриваются вопросы автоматической обработки содержимого веб-ресурсов. Поскольку скорость устаревания передаваемой во всемирной сети информации очень велика, актуальной темой становится своевременное извлечение необходимых данных из сети интернет. Объектом исследования являются веб-ресурсы, содержащие в себе неадаптированный к автоматизированной обработке текст. Предметом исследования является набор программных средств и методов. Особое внимание уделяется определению категорий объявлений, расположенных на специализированных сайтах. Также рассматриваются прикладные аспекты разработки универсальной архитектуры систем сбора информации. В ходе данного исследования использовались следующие методы: аналитический обзор основных принципов разработки систем автоматизированного сбора информации и анализа естественных языков. Для получения практико-ориентированного результата использовались методы синтеза и анализа. Особым вкладом авторов в исследование темы является разоработка автоматизированной системы сбора, обработки и классификации информации, содержащейся на веб-ресурсе. Новизна исследования заключается в использовании нового подхода к решению данной проблемы на основе учета семантики и структуры характерной для конкретных сайтов. Основными выводами проведенного исследования являются применимость и эффективность используемого метода классификации для решения данной задачи.*

Ключевые слова: *парсинг, анализ текста, категоризация веб-сайтов, система классификации, сбор информации, веб-роботы, машинное обучение, обработка информации, краулинг, большие данные*

Abstract: *This article reviews the problems of automatic processing of web content. Since the speed of obsolescence of information in the global network is very high, the problem of prompt*

DOI: 10.7256/2305-6061.2016.4.21438

extraction of the necessary data from the Internet becomes more urgent. The research focuses on the web resources that contain text, unadapted to the automated processing. The subject of the research is a set of software and methods. A particular attention is paid to the categorization of ads placed on specialized websites. The authors also review practical aspects of the development of a universal architecture of information-gathering systems. The following methods were used during this study: analytical review of the main principles of development of systems of automated information gathering and analysis of natural languages. For obtaining practice-oriented methods of synthesis and analysis results were used. A special contribution of the authors of the study is in developing an automated system for collecting, processing and classification of the information contained on the web-site. The novelty of the research is to use a new approach to solve this problem by taking into account the semantics and structure characteristic for specific sites. The main conclusions of the study are the applicability and effectiveness of the classification method for solving this problem.

Keywords: *machine learning, web robots, information collection, classification system, web-sites categorization, text analisis, parsing, data processing, crawling, big data*

Введение

Веб-краулер (веб-робот или парсер) – это специальная автоматизированная компьютерная программа, которая исследует веб-ресурсы на предмет интересующей ее владельцев информации, собирает данные и анализирует их в соответствии с внутренними правилами [1, 2]. В настоящее время информация в интернете имеет очень высокую скорость устаревания, а новый контент на веб-ресурсах появляется чуть-ли не ежечасно [3]. В современном мире скорость реакции на события имеет первостепенное значение. В таких условиях актуальной является задача своевременного и автоматизированного сбора информации. Однако, зачастую при сборе контента, имеющего неструктурированный формат (например, данные, генерируемые пользователями ресурса), возникает задача адаптивного анализа текста. Одной из важных ее составляющих является автоматизированное категорирование и классифицирование содержимого веб-страницы [4, 5].

Классификация текстов является одной из задач информационного поиска и состоит в отнесении текста (в нашем случае содержимого веб-страницы) к одной из нескольких категорий. Данная классификация может быть автоматической, ручной или полуавтоматической [6]. В контексте данного исследования наибольший интерес представляют автоматические и полуавтоматические методы. Существует множество методов решения задачи классификации, таких, как Naive Bayes, TWCNB, k-nearest-neighbour, Support Vector Machine, N-Gram based classification и др. Данные алгоритмы имеют свои сильные и слабые стороны, различаются по быстродействию и точности классификации [7].

Описание задачи

Большинство исследовательских работ по тематике классификации и категорированию веб-ресурсов опираются на методы классификации текстов и не учитывают семанти-

DOI: 10.7256/2305-6061.2016.4.21438

ческие особенности присущие веб-страницам [8, 9]. Так, на веб-странице часть текста не имеет отношение к категории основного его содержания (общие навигационные блоки, рекламные блоки, блоки рекомендаций и т.д.). Отсюда встает дополнительная задача учета структуры веб-ресурса при парсинге его содержания, один из подходов к решению которой описан в [8]. В решаемой нами задаче, мы рассматривали категорирование сайтов каталогов и объявлений для классификации конкретного продукта по тематике. Нас интересовало адаптивное извлечение характеристик товара, что невозможно без предварительного определения его типа. Сайты данной категории имеют похожую структуру, что позволяет сфокусировать краулера на тех блоках контента, которые с наибольшей вероятностью содержат релевантные описания содержания страницы. Данный подход широко используется при парсинге тематических сайтов, поскольку помогает избежать проблем характерных для алгоритмов, учитывающих лишь содержательные характеристики [10-12].

К сожалению, слабым местом существующих исследований по данной тематике является отсутствие подробностей реализации предлагаемых алгоритмов и описания методов решения проблем, возникающих при апробации данных методов. Стоит отметить, что используемые подходы к классификации сайтов с простейшей структурой (сайты объявлений) в дальнейшим могут быть обобщены на более сложные ресурсы, такие как блоги и новостные порталы, информация с которых может быть полезна при решении задачи автоматизированной оценки событий, общественного мнения и обеспечения информационной безопасности организации.

Описание алгоритма работы системы

Опорное слово – это лексическая единица на веб-странице, которая имеет значение при формировании категории. В рамках текущего исследования мы рассматривали в качестве таковых слова из полей meta-keywords и meta-description, а также заголовки внутри content полей веб-страницы.

Процесс категорирования мы разделили на несколько этапов в соответствии с функциональным наполнением (Рис. 1.).

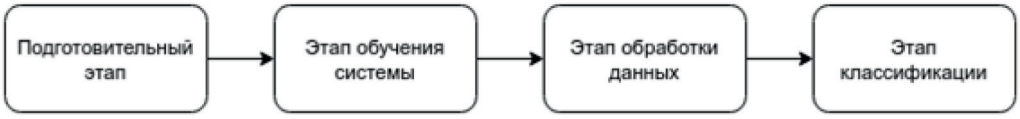

Рис. 1. Задачи подсистемы анализа

На подготовительном этапе происходит выборка из словаря категорий, а также подготовка тренировочного набора веб-страниц с размеченной классификацией и наличием определенного числа опорных слов. Обучающий набор формируется на базе двух процессов: ручное классифицирование с использованием специального программного продукта, который упрощает работу оператора и по сути превращает процесс в полуавтоматический; автоматическая выборка описаний с веб-ресурсов, которые имеют четкую

DOI: 10.7256/2305-6061.2016.4.21438

структуру классификации страниц (категории, теги, и т.д.). Для выделения лексических единиц и приведения их к начальной форме, мы используем морфологический анализатор из бибиотеки pymorphy [13].

На этапе обучения происходит выборка рассматриваемых опорных слов из тренировочного набора и обновление векторов классификации для каждого опорного слова.

```
for keyword in total_words:
    test_keyword = self.db.keywords.find_one({'word': keyword})
    total = test_keyword['total'] + 1
    cats = test_keyword['categories']
    found = False
    for cat in cats:
        if cat['name'] == category_name:
            cat['count'] += 1
            found = True
            break

    if not found:
        cats.append({'name': category_name, 'weight':
                                                0.0, 'count': 1})

    for cat in cats:
        cat['weight'] = float(cat['count']) / float(total)
```

На этапе обработки происходит нормирование значений, устранение статистических выбросов, формирование вспомогательных структур данных и словарей, иными словами – подготовка данных к дальнейшему использованию.

```
total_words = self._get_key_words(html_page)
match_cats = {}

for keyword in total_words:
    test_keyword = self.db.keywords.find_one
                                            ({'word': keyword})
    if test_keyword:
        cats = test_keyword['categories']
        for cat in cats:
            cat_name = cat['name']
            match_cats[cat_name] += cat['weight']

    if len(match_cats) > 0:max_cat = max(match_cats.iteritems(),
                                key=operator.itemgetter(1))[0]
    return {url, max_cat}
```

На этапе классификации происходит постраничный парсинг веб-ресурса из рабочей

DOI: 10.7256/2305-6061.2016.4.21438

выборки, выделение опорных слов, суммирование векторов категорий опорных слов и формирование результирующего вектора, компоненты которого представляют веса отнесения его к той или иной категории.

Функциональная схема работы системы

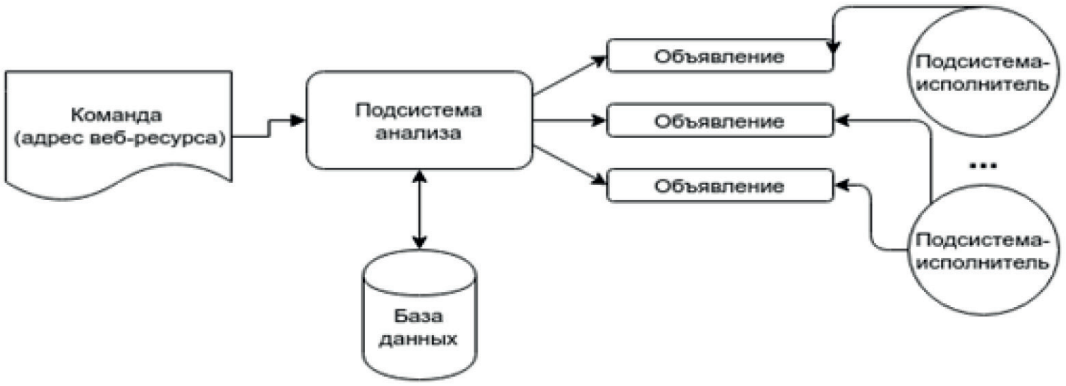

Рис. 2. Общая структура системы парсинга

Работа системы парсинга начинается с команды на запуск, полученной от администратора или от подсистемы управления (например, запуск по таймеру). Подсистема анализа начинает свою работу с веб-сайтом, собирая необходимые страницы.

Во время своей работы она выделяет определенные объявления и передает их соответствующим парсерам для дальнейшего извлечения конкретных параметров и содержания объявления. Сама же подсистема анализа во время выполнения должна получить общие для всех категорий данные о каждом объявлении.

Её работа сводится к последовательному ответу на следующие вопросы:

- является ли данная страница страницей с объявлением. Ответ да – продолжить анализ; нет – пропустить;
- определить тематику страницы. Необходимо классифицировать информацию, понять, соответствует ли она определенным категориям. Сложность здесь состоит в том, что в рамках одной категории может быть дополнительное подразделение на уровни;
- определить местоположение объекта объявления;
- определить ключевые параметры. Важным является определения «рамок» параметра, например, адекватной цены;
- определить дату публикации. Самые «свежие» данные будут обладать наибольшей ценностью;
- определить контакты автора объявления, которые являются очень важной информацией, при этом, наиболее тяжело извлекаемой в общем виде, на произвольном сайте.

DOI: 10.7256/2305-6061.2016.4.21438

Описание результатов классификации

Разработанная система может работать в двух режимах: обучение и рабочий режим классификации. В обоих случаях, необходимо знать, какую веб-страницу считать объявлением, а какую – лишь набором дополнительных ссылок. В настоящее время, это делается в ручном режиме – для каждого сайта указывается, какого формата ссылка будет считаться объявлением. В будущем этот вопрос будет также решён в общем виде. То есть, нужно просто переходить по всем возможным ссылкам в рамках данного домена и как только была обнаружена «интересующая» страница, необходимо организовать её обработку.

Для обучения за основу был взят региональный веб-сайт с каталогом объявлений и аукционом «24au.ru», так как категории данного сайта достаточно полно охватывают все возможные направления сбора информации, к тому же, они располагаются в определённой иерархии. За 10 часов многопоточным краулером, работающим в 3 потока, было обработано более 149 тысяч страниц, выделено 8115 ключевых слов по 1382 различным категориям.

Далее производилось тестирование обученной системы. Для проверки работы системы, ей на вход подавались реальные объявления с трех подобных веб-сайтов объявлений («irr.ru», «24auto.ru», «avito.ru») и оценивалась корректность классификации. В большинстве случаев, система верно определяла категорию (важно отметить, что не только базовую, а полную иерархию нескольких категорий) в случае с наиболее часто продаваемой электроникой (ноутбуки, смартфоны). В ситуации с автомобилями, базовая категория определялась всегда верно, а далее система часто относила объявление к запчастям, нежели чем к продаже авто.

Для оценки качества автоматической классификации использовалась метрика F_1, вычисляемая по формуле

$$F_1 = 2 * \frac{p*r}{p+r}, \qquad (1)$$

r – это отношение R/Q, p – это отношение R/L. В свою очередь, R – количество правильно отнесенных к категории объявлений, Q – общее количество объявлений, которые должны быть отнесены к данной категории, L – общее количество объявлений, отнесенных системой парсинга к данной категории.

В таблице 1 приведены количество собранных объявлений по каждому из рассматриваемых сайтов и результирующая F_1 метрика. Стоит отметить, что на некоторых сайтах мы столкнулись с использованием систем противодействия автоматизированному сбору информации [14], что сказалась как на количестве собранных объявлений, так и на результатах классификации.

Таблица 1. Количественные показатели парсинга

Веб-ресурс	Количество страниц	F1 метрика
24au.ru	149 тыс.	- (обучение)
24auto.ru	168 тыс.	0.82
avito.ru	3 тыс.	0.73

DOI: 10.7256/2305-6061.2016.4.21438

Таким образом, была определена дополнительная задача – доработать интерфейс, в котором можно будет осуществлять дальнейшее обучение системы, для наиболее точного определения категории. Также, необходимо решить, как алгоритмически будет реализовано «дообучение» системы экспертом, чтобы она могла после указания ошибочности принятого решения и правильного варианта, верно относить данное и похожие объявления к нужной категории. В процессе сбора информации были выявлены попытки блокировки запросов со стороны ресурсов, что вызвало необходимость использовать прокси, заметим, что достаточно мало сайтов используют механизмы противодействия, изучения данных механизмов также является актуальной задачей [14].

Выводы

Была рассмотрена задача автоматической классификации тематики текста страниц веб-ресурса, показан новый подход к решению данной проблемы с учетом семантики и структуры характерной для сайтов объявлений. Была разработана автоматизированная система сбора, обработки и классификации информации, содержащейся на веб-ресурсе. Проведенные исследования показали применимость и эффективность используемого метода классификации для решения данной задачи.

Библиография :

1. Liu H. and Milios E. (2012), PROBABILISTIC MODELS FOR FOCUSED WEB CRAWLING. Computational Intelligence, 28: 289–328.
2. Менщиков А.А., Гатчин Ю.А. Методы обнаружения автоматизированного сбора информации с веб-ресурсов // Кибернетика и программирование. – 2015. – № 5. – С.136-157.
3. Razniewski Simon, and Werner Nutt. Long-term Optimization of Update Frequencies for Decaying Information // Proceedings of the 18th International Workshop on Web and Databases. ACM. – 2015.
4. Pant Gautam, and Padmini Srinivasan Learning to crawl: Comparing classification schemes // ACM Transactions on Information Systems (TOIS) 23.4 (2005): 430-462.
5. Kim Jin Young, et al. Characterizing web content, user interests, and search behavior by reading level and topic // Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012.
6. Паутов К. Г., Попов Ф. А. Информационная система анализа и тематической классификации веб-страниц на основе методов машинного обучения // Современные проблемы науки и образования. 2012. №6.
7. Aggarwal Charu C., and ChengXiang Zhai. "A survey of text classification algorithms." Mining text data. Springer US, 2012. 163-222.
8. Chen Yu, Wei-Ying Ma, and Hong-Jiang Zhang. "Detecting web page structure for adaptive viewing on small form factor devices." Proceedings of the 12th international conference on World Wide Web. ACM, 2003.
9. Агеев Михаил Сергеевич, Добров Борис Викторович, Лукашевич Наталья Валентиновна Автоматическая рубрикация текстов: методы и проблемы // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. 2008. №4.

DOI: 10.7256/2305-6061.2016.4.21438

10. Eswaran Dhivya, Paul N. Bennett, and Joseph J. Pfeiffer III. "Modeling Website Topic Cohesion at Scale to Improve Webpage Classification." Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015.
11. Martinez-Alvarez Miguel et al. "Document Difficulty Framework for Semi-automatic Text Classification" DAWAK (2013).
12. Tripathi Nandita, Michael Oakes, and Stefan Wermter. "A Scalable Meta-Classifier Combining Search and Classification Techniques for Multi-Level Text Categorization." International Journal of Computational Intelligence and Applications 14.04 (2015).
13. Морфологический анализатор pymorphy [Электронный ресурс]. – Режим доступа: https://pythonhosted.org/pymorphy/, свободный (дата обращения: 30.09.2016).
14. Менщиков А.А. Методы обнаружения автоматизированного сбора информации с веб-ресурсов // Альманах научных работ молодых ученых Университета ИТМО-2016. – Т. 3. – С. 230-232.

References:

1. Liu H. and Milios E. (2012), PROBABILISTIC MODELS FOR FOCUSED WEB CRAWLING. Computational Intelligence, 28: 289–328
2. Menshchikov A.A., Gatchin Yu.A. Metody obnaruzheniya avtomatizirovannogo sbora informatsii s veb-resursov // Kibernetika i programmirovanie. – 2015. – № 5. – S.136-157.
3. Razniewski Simon, and Werner Nutt. Long-term Optimization of Update Frequencies for Decaying Information // Proceedings of the 18th International Workshop on Web and Databases. ACM. – 2015.
4. Pant Gautam, and Padmini Srinivasan Learning to crawl: Comparing classification schemes // ACM Transactions on Information Systems (TOIS) 23.4 (2005): 430-462.
5. Kim Jin Young, et al. Characterizing web content, user interests, and search behavior by reading level and topic // Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012.
6. Pautov K. G., Popov F. A. Informatsionnaya sistema analiza i tematicheskoi klassifikatsii veb-stranits na osnove metodov mashinnogo obucheniya // Sovremennye problemy nauki i obrazovaniya. 2012. №6.
7. Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text classification algorithms." Mining text data. Springer US, 2012. 163-222.
8. Chen, Yu, Wei-Ying Ma, and Hong-Jiang Zhang. "Detecting web page structure for adaptive viewing on small form factor devices." Proceedings of the 12th international conference on World Wide Web. ACM, 2003.
9. Ageev Mikhail Sergeevich, Dobrov Boris Viktorovich, Lukashevich Natal'ya Valentinovna Avtomaticheskaya rubrikatsiya tekstov: metody i problemy // Uchen. zap. Kazan. un-ta. Ser. Fiz.-matem. nauki. 2008. №4.
10. Eswaran, Dhivya, Paul N. Bennett, and Joseph J. Pfeiffer III. "Modeling Website Topic Cohesion at Scale to Improve Webpage Classification." Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015.
11. Martinez-Alvarez, Miguel et al. "Document Difficulty Framework for Semi-automatic Text Classification" DAWAK (2013).

DOI: 10.7256/2305-6061.2016.4.21438

12. Tripathi, Nandita, Michael Oakes, and Stefan Wermter. "A Scalable Meta-Classifier Combining Search and Classification Techniques for Multi-Level Text Categorization." International Journal of Computational Intelligence and Applications 14.04 (2015).
13. Morfologicheskii analizator pymorphy [Elektronnyi resurs]. – Rezhim dostupa: https://pythonhosted.org/pymorphy/, svobodnyi (data obrashcheniya: 30.09.2016).
14. Menshchikov A.A. Metody obnaruzheniya avtomatizirovannogo sbora informatsii s veb-resursov // Al'manakh nauchnykh rabot molodykh uchenykh Universiteta ITMO-2016. – T. 3. – S. 230-232.

DOI: 10.7256/2305-6061.2016.4.21438