УДК 004.442: 519.25

ИНФОРМАЦИОННАЯ СИСТЕМА АНАЛИЗА И ТЕМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ВЕБ-СТРАНИЦ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Паутов К.Г., Попов Ф.А.

Бийский технологический институт (филиал) Федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Алтайский государственный технический университет им. И.И. Ползунова» (Россия, 659305, Алтайский край, г. Бийск, ул. Трофимова, 27), pautov@bti.secna.ru

В статье рассматриваются вопросы проектирования и разработки информационной системы анализа и тематической классификации веб-страниц с использованием аппарата опорных векторов и нейронных сетей. В начале статьи приводится общая характеристика подходов к классификации веб-страниц, основанных на анализе их текстового содержимого, гиперссылок и метаинформации. Авторы ссылаются на работы исследователей, доказывающие низкую эффективность методов классификации, основанных лишь на этих текстовых признаках веб-страницы. В работе указывается, что качество классификации можно повысить благодаря предварительному сегментированию веб-страницы на семантические блоки и, таким образом, отделения содержательного контента от информационного шума. Авторами приводится архитектура информационной системы, реализующей предложенный метод. Дается описание составляющих систему функциональных подсистем и модулей. Затрагиваются вопросы представления текстовых и пространственных признаков в базе данных. Дается общая оценка методов машинного обучения в их применимости к решению задачи.

Ключевые слова: классификация веб-страниц, машинное обучение, извлечение информации из веб-страниц, пространственные признаки.

AN INFORMATION SYSTEM OF WEB-PAGE CLASSIFICATION BASED ON MACHINE LEARNING METHODS

Pautov K.G., Popov F.A.

Biysk Technological Institute (branch) of the federal government budget of educational institutions of higher education "Altai State Technical University. of I.I. Polzunov" (Russia,659305, Altay territory, Biisk, street Trofimova, 27), pautov@bti.secna.ru

The article describes a design of an information system of the analysis and subject classification of web pages with use of the Support Vector Machine and Neural networks. At the beginning of article the total characteristic of approaches is given to classification of the web pages based on the analysis of their text contents, hyperlinks and metadata. Authors refer to the operations of researchers proving a low performance of classification methods, based only on these text features of the web page. In operation it is specified that quality of classification can be increased thanks to preliminary segmentation of the web page on semantic units and, thus, separations of an informative content from information noise. Authors give a structure of the information system implementing the offered method. The description of the functional subsystems and modules is given. Special attention is paid to process a textual and spatial attributes representation in the database. The general assessment of methods of machine training in their applicability to the solution of the task is given.

Keywords: webpage classification, machine learning, web mining, spatial features.

Стремительные темпы развития Интернета и быстрый рост количества информации привели к усложнению задачи информационного поиска. Для того чтобы систематизировать и описать множество веб-сайтов в 90-е годы были разработаны системы интернет-каталогов, в которых сайты были классифицированы в соответствии с их тематической принадлежностью. Благодаря этому пользователи могли с успехом просмотреть большую часть веб-сайтов по интересующей их тематике. Однако в современном мире сайты стали политематичны, и их количество увеличилось на несколько порядков. Ручное

рубрицирование, проводимое экспертами и направленное на поддержание каталогов в актуальном состоянии, стало в современных условиях невозможным. В связи с этим остро встала необходимость в разработке методов автоматической классификации веб-страниц.

Одним из перспективных направлений применения данных методов являются системы фильтрации интернет-трафика. Использование автоматических методов анализа содержимого позволяет определять принадлежность веб-страницы к конкретному тематическому классу (классам). Причем для каждого класса и группы пользователей может быть выработана отдельная политика доступа к ресурсу.

Рассмотрению методов автоматической классификации веб-страниц посвящен ряд публикаций. Наиболее распространенным является метод построения классификатора на основе текстового содержимого страницы. Однако данный метод имеет низкую точность, прежде всего из-за того, что обучающее подмножество, на котором строится классификатор, содержит элементы навигации, рекламные объявления и прочую, не относящуюся к основной тематике информацию. В [2; 7] показано, что точность классификации можно повысить, включив в анализ тексты входящих и исходящих ссылок страницы, а также текст и контекст ссылки. Существует ряд других методов, рассмотрение которых выходит за рамки данной статьи.

Предлагаемый нами метод заключается в том, чтобы проводить классификацию не по всему тексту веб-страницы, а только по ее содержательной части. Для этого производится предварительная сегментация веб-страниц на семантические блоки. Как показала практика, количество таких блоков для подавляющего большинства веб-страниц лежит в диапазоне от трех до восьми единиц. На основе вычисленных пространственных характеристик блока и анализа его содержимого делается предположение о том, является ли блок содержательным или «шумовым». Содержащийся в блоках текст подвергается графематическому и морфологическому анализу.

На рисунке 1 представлена архитектура предлагаемой нами информационной системы классификации веб-страниц. Стрелками обозначена последовательность обработки информации компонентами ИС.

В основу информационной системы положена модульная структура, позволяющая производить ее реализацию в виде отдельных функциональных модулей и подсистем. Каждый блок имеет законченную функциональность, работая с определенным набором входных параметров и выходных данных. Рассмотрим функции каждого из блоков более подробно.

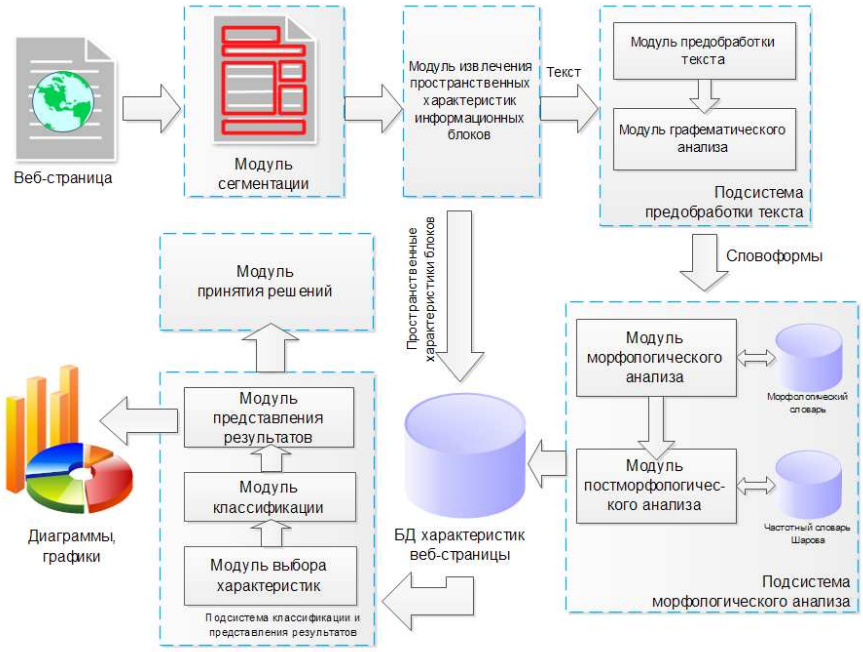

Рис. 1. Архитектура информационной системы.

Модуль сегментации предназначен для разделения веб-страниц на отдельные семантические блоки. Эта процедура позволяет впоследствии исключить из рассмотрения блоки, содержащие элементы навигации, рекламу и прочий информационный шум.

Наиболее эффективным с точки зрения авторов является подход к сегментированию веб-страниц, основанный на анализе ее визуального представления [9]. Подразумевается, что веб-дизайнеры изначально, при разработке страниц сайта стремятся организовать контент таким образом, чтобы семантически однородная (связанная) информация находилась в одном или рядом стоящих (смежных) блоках. Блоки при этом отделяются друг от друга при помощи визуальных разделителей, например пустых строк, рамок, изображений и т.п. Таким образом, пользователь визуально способен определить для себя границы семантически однородного текста.

Модуль извлечения пространственных характеристик информационных блоков предназначен для определения расположения блока относительно страницы, вычисления размеров занимаемой им области, порядка, занимаемого им в дереве. Размеры занимаемой области и координаты блока являются пространственными признаками, позволяющими статистически оценить его важность. Здесь также анализируется содержимое блоков. Вычисляются статистические характеристики, такие как длина текста в символах, количество изображений, таблиц и гиперссылок. Полученные таким образом характеристики записываются в базу данных. Текстовое содержимое блока передается подсистеме предобработки текста.

Подсистема предобработки текста состоит из двух модулей: модуль предобработки текста и модуль графематического анализа.

Функциями *модуля предобработки текста* являются: снятие теговой разметки; приведение всех символов текста к одному регистру (например, верхнему). Немаловажной задачей является определение кодировки текста, т.к. одним и тем же печатным символам в электронном представлении соответствуют различные коды. Для русскоязычных веб-страниц может быть использована одна из следующих кодировок: Unicode, CP-1251, Koi-8r, DOS, Mac. Ввиду того что символы-разделители (знаки препинания, пробелы и т.д.) кодируются одинаково, достаточно просто выделить из текста отдельные слова. Далее эти слова-кандидаты сравниваются с уже имеющимися в словаре словами, заранее известными для каждой кодировки.

Модуль графематического анализа осуществляет первоначальный анализ естественного текста, представленного в виде цепочки ASCII-символов, и выполняет разделение текста на слова. При этом разделители, такие как знаки препинания, символы пробела, табуляции и перевода каретки удаляются. Результатом работы подсистемы предобработки текста является список словоформ, расположенных в порядке их следования в тексте.

Подсистема морфологического анализа предназначена для определения лемм и морфологических характеристик словоформ, полученных на этапе предобработки текста.

Модуль морфологического анализа реализован на основе декларативного подхода и использует в своей работе словарь всех возможных словоформ для каждого слова [5]. Для каждой словоформы приводится описание значений ее морфологических категорий (род, число, падеж и т.д.) и нормальная форма (лемма). Задача данного модуля состоит в определении единственно верной морфологической интерпретации слова. Однако зачастую эту задачу решить достаточно сложно ввиду возникновения ситуаций морфологической неоднозначности – когда на одну словоформу находится несколько возможных лемм. Такая ситуация в нашей программе разрешается при помощи использования нового частотного словаря русской лексики [1] в *модуле постморфологического анализа*. Для анализа выбирается лемма с наибольшей частотой.

Морфологические характеристики словоформ, леммы и пространственные характеристики блоков помещаются в базу данных для удобства последующей работы с ними. В нашей работе использована СУБД Oracle 11g Release 2, ввиду обеспечения последующей интеграции с существующей ИС [3].

Информация о веб-странице хранится в базе данных в виде набора связанных таблиц, содержащих информацию о самой странице, семантических блоках и содержимом этих

блоков. Текст веб-страницы представлен в виде связанного списка лемм с указанием их позиций. Такое представление удобно для дальнейшей работы и позволяет учитывать контекст. Что, в свою очередь, позволяет говорить о представлении смысла предложения как суперпозиции смыслов слов, его составляющих. Следовательно, и смысл текста в целом можно представить как суперпозицию смыслов входящих в него предложений [4].

Модуль выбора характеристик отвечает за формирование запросов к базе данных, и выборку из нее характеристик, необходимых для работы классификатора. Это дает возможность использовать в работе несколько классификаторов и выбирать произвольные характеристики. При выборе характеристик мы руководствовались следующими параметрами:

1) влияние модели на качество классификации;
2) вычислительная ресурсоемкость при классификации и требуемый объем памяти для хранения модели веб-страницы;
3) временные затраты на классификацию одной веб-страницы. Данный параметр является очень важным ввиду того, что получившаяся модель может обладать высокой точностью классификации, но при этом время, затрачиваемое на определение класса, будет столь продолжительным, что не позволит в дальнейшем использовать модель на практике.

Модуль классификации реализует один из наиболее популярных методов классификации. Для построения классификаторов было принято решение использовать в качестве отправной точки метод опорных векторов (SVM – Support Vector Machine) [8] и многослойный перцептрон (MLP – MultiLayer Perceptron) [6], впоследствии выбрав наиболее точный. При использовании метода опорных векторов задача многоклассовой классификации приводится к нескольким бинарным задачам: последовательного отделения первого класса от остальных, второго класса от оставшихся и т.д. После решения этих бинарных задач получается несколько обученных классификаторов на базе SVM, соответствующих каждому классу. При проведении процедуры классификации каждый из классификаторов возвращает коэффициент принадлежности объекта к классу. Класс объекта определяется по максимальному значению этого коэффициента.

Многослойный перцептрон (MLP) – архитектура нейронной сети прямого распространения, которая на сегодняшний день является наиболее популярной среди исследователей. В качестве алгоритма обучения используется алгоритм обратного распространения ошибки. Среди достоинств данного метода можно отметить его изученность и простоту реализации. К недостаткам – необходимость подбора архитектуры нейронной сети, что является сложной задачей. В сравнении с методом опорных векторов нейронные сети имеют более низкую скорость обучения и склонны к переобучению.

Модуль представления результатов. Результаты классификации должны быть обработаны и приведены к виду, удобному для их дальнейшего использования исследователем или программой (представление результатов классификации в виде таблиц и графиков). При использовании результатов классификации на практике их необходимо преобразовывать в формат, удобный для взаимодействия с другими компонентами или ИС. Например, в качестве такого компонента может выступать *модуль принятия решений,* который на основании результатов работы классификатора вырабатывает определенное управляющее воздействие (разрешает или ограничивает доступ к ресурсу).

Предложенный в данной работе подход позволяет сделать классификатор более эффективным и точным по сравнению с рассмотренными выше методами, так как позволяет исключить влияние информационного шума на результаты классификации и учитывает контекст. Кроме того, данный подход менее ресурсоемок, так как оперирует меньшими объемами данных.

Предложенный в статье метод использован при построении модулей информационной системы учета и контроля интернет-трафика [3]. Такой метод позволяет сделать фильтрацию веб-страниц более гибкой и эффективной по сравнению с веб-фильтрами, использующими «черные» и «белые» списки URL-адресов, а также избежать затрат, связанных с поддержанием этих списков в актуальном состоянии (оплата труда экспертов, проводящих классификацию).

Список литературы

1. Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). – М. : Азбуковник, 2009.

2. Маслов М. Ю., Пяллинг А.А., Трифонов С.И. Автоматическая классификация веб-сайтов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Десятой всерос. науч. конф. «RCDL'2008», Дубна, 7-11 окт. 2008.

3. Паутов К.Г., Парахин В.В. Система учета и контроля интернет-трафика пользователей БТИ АлтГТУ // Информационные технологии в науке, экономике и образовании : материалы Всероссийской научной конференции (16-17 апреля 2009 г.). В 2-х ч. / под ред. О.Б. Кудряшовой ; Алт. гос. техн. ун-т, БТИ. – Бийск : Изд-во Алт. гос. техн. ун-та, 2009. – Ч. 2. – С. 50-53.

4. Паутов К.Г., Попов Ф.А., Данилюк Ю.С. Извлечение семантической информации из веб-страниц с использованием данных о расположении в них информационных блоков // Материалы XIX Всероссийской научно-методической конференции «Телематика'2012» (25–

28 июня 2012 г., Санкт-Петербург) / Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики. – СПб., 2012. – С. 228-230.

5. Сокирко А.В. Морфологические модули на сайте www.aot.ru [Электронный ресурс]. – Режим доступа: http://www.aot.ru/docs/sokirko/Dialog2004.htm (дата обращения: 12.10.2012).

6. Хайкин С. Нейронные сети: полный курс. Neural Networks: A Comprehensive Foundation. – 2-е изд. – М. : Вильямс, 2006. – 1104 с.

7. Шабанов В.И., Андреев А.М. Метод классификации текстовых документов, основанный на полнотекстовом поиске // Труды первого российского семинара по оценке методов информационного поиска / под ред. И.С. Некрестьянова. – СПб. : НИИ Химии СПбГУ, 2003. – 132 с.

8. Burges C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition // Data Mining and Knowledge Discovery. – 1998. – Vol. 2. – № 2. – P. 121-167.

9. Giuseppe Della Penna, Daniele Magazzeni and Sergio Orefice. A spatial relation-based framework to perform visual information extraction // Knowledge and Information Systems. – V. 30. – № 3. – P. 667-692.

Рецензенты

Старовиков Михаил Иванович, д.п.н., к.ф.-м.н., доцент, и.о. зав. кафедрой физики ФГБОУ ВПО «Алтайская государственная академия образования им. В.М. Шукшина», г. Бийск.

Попок Николай Иванович, д.т.н., профессор, нач. лаб. ОАО «ФНПЦ «Алтай», г. Бийск.