УДК 00.004.9
DOI 10.33764/2618-8031-2020-1-130-134

МОДЕЛИ И МЕТОДЫ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ЗАПРОСОВ В СИСТЕМЕ ИДЕНТИФИКАЦИИ ПАТТЕРНОВ ПОЛИЯЗЫЧНЫХ ТЕКСТОВ

Владимир Павлович Куликов
Северо-Казахстанский государственный университет им. М. Козыбаева, 150000, Республика Казахстан, г. Петропавловск, ул. Пушкина, 86, кандидат физико-математических наук, доцент кафедры информационно-коммуникационных технологий, e-mail: qwertyra@mail.ru

Валентина Петровна Куликова
Северо-Казахстанский государственный университет им. М. Козыбаева, 150000, Республика Казахстан, г. Петропавловск, ул. Пушкина, 86, кандидат технических наук, доцент кафедры кафедры информационно-коммуникационных технологий, e-mail: v4lentina@mail.ru

Елена Михайловна Крылова
Сибирский государственный университет геосистем и технологий, 630108, Россия, г. Новосибирск, ул. Плахотного, 10, кандидат технических наук, доцент кафедры высшей математики, тел. (383)343-25-77, e-mail: redikarceva@ssga.ru

Гулнур Туратайкызы Еркебулан
Северо-Казахстанский государственный университет им. М. Козыбаева, 150000, Республика Казахстан, г. Петропавловск, ул. Пушкина, 86, докторант, e-mail: erkgulnur@mail.ru

Описывается схема классификации текстовых документов, состоящая из пяти шагов: предобработка, индексация, выбор признаков, построение и обучение классификатора, оценка качества. Приведен сравнительный анализ классификационных методов по точности и времени обучения классификатора. Сделан вывод об оптимальности методов классификации.

Ключевые слова: классификация текста, оценка качества классификатора, нейронные сети, метод Байеса, метод опортных векторов.

MODELS AND METHODS OF CLASSIFICATION OF TEXT REQUESTS IN THE SYSTEM OF PATTERN IDENTIFICATION OF POLYLINGUAL TEXTS

Vladimir P. Kulikov
North Kazakhstan State University n.a. M. Kosybaev, 86, Pushkina St., Petropavlovsk, 150000, Kazakhstan Republic, Ph. D., Associate Professor, Department of Information and Communication Technologies, e-mail: qwertyra@mail.ru

Valentina P. Kulikova
North Kazakhstan State University n.a. M. Kosybaev, 86, Pushkina St., Petropavlovsk, 150000, Kazakhstan Republic, Ph. D., Associate Professor, Department of Information and Communication Technologies, e-mail: v4lentina@mail.ru

Elena M. Krylova
Siberian State University of Geosystems and Technologies, 10, Plakhotnogo St., Novosibirsk, 630108, Russia, Ph. D., Associate Professor, Department of Higher Mathematics, phone: (383)343-25-77, e-mail: redikarceva@ssga.ru

Gulnur T. Yerkebulan
North Kazakhstan State University n.a. M. Kosybaev, 86, Pushkina St., Petropavlovsk, 150000, Kazakhstan Republic, Postdoc, e-mail: erkgulnur@mail.ru

A classification scheme for text documents consisting of five steps is described: preprocessing, indexing, selection of features, construction and training of a classifier, quality assessment. Two comparative analyzes by classification methods are considered. Conclusions are drawn about models and classification methods regarding implementation efficiency.

Key words: text classification, classifier quality assessment, neural networks, Bayes method, export vector method.

В разрабатываемой авторами системе идентификации паттернов полиязычных текстов необходимо формирование поискового индекса англоязычных документов. В дальнейшем планируется формирование поисковых индексов и на других языках. Предполагается, что готовая система должна решить следующую задачу: на вход поступает текст на любом языке, а на выходе формируется отчет о том, что какая-то часть текста заимствована из второго языка, а какая-то из третьего языка.

Для проверки работы системы на ее вход поступает русскоязычный проверяемый документ, который переводится на английский язык и сравнивается на схожесть с документами из поискового индекса англоязычных документов.

Работа с поисковыми индексами подразумевает использование классификаторов. Благодаря классификации поисковый индекс разделен на тематические каталоги. Классификация нужна, чтобы сравнивать поступающий текст не со всем индексом, а только с той его частью, к которой относится предметная область текста. В статье рассматриваются наиболее распространенные модели и методы классификации текстовых запросов (поступающие на вход проверяемые документы).

Классификация документов (рис. 1) – это одна из задач поиска информации, которая включает в себя присвоение документу одной из нескольких категорий на основе содержания документа [1].

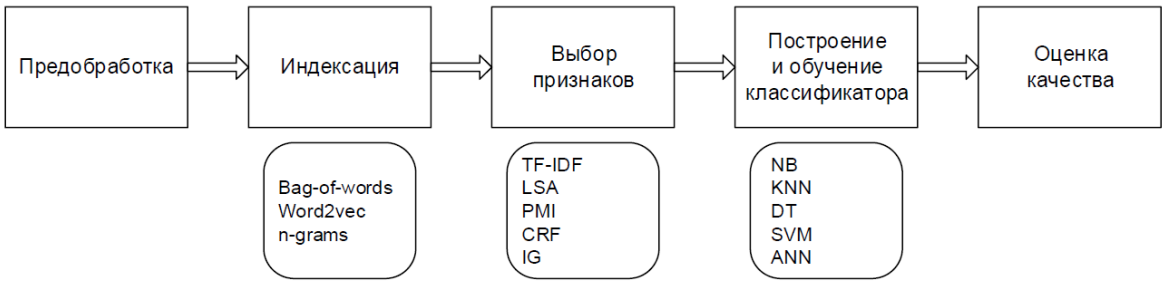


Рис. 1. Общая схема классификации текстового запроса [2]

Необходимо отличать классификацию текстов от кластеризации, в последнем случае тексты также группируются в соответствии с некоторыми критериями, но заранее предопределенных категорий нет [1].

Предобработка и индексация документов. Сначала необходимо выполнить разбиение текста на слова, удалить незначительные слова, например, артикли, союзы и прочие. Затем выполнить определение морфологических характеристик слов (установить к каким частям речи относятся слова) с систематизацией. Таким образом, признаками текста будут наиболее значимые слова.

Следующим шагом классификации является индексация текста, т. е. представление документа в виде числовой модели.

Модель «bag-of-words» («мешка слов») представляет документ в виде вектора в многомерном пространстве, координаты которого соответствуют номерам слов, а значения координат – значениям весов.

Модель «Word2vec» представляет каждое слово в виде вектора, который содержит информацию о контекстных (сопутствующих) словах.

Модель «*n*-грамм» – это наборы из последовательности символов. Самыми распространенными считаются униграммы и биграммы. Кроме того, используются символьные n-граммы, когда текст разбивают на наборы символов определенный длины. В каждом конкретном случае лучше экспериментировать, выявляя наиболее эффективную модель «n-грамм» [3].

Выбор признаков. Количество признаков напрямую связано со скоростью вычислений, поэтому для ускорения классификатора используют уменьшение числа терминов.

Вычислительная сложность различных методов классификации связана с размерностью пространства признаков. Поэтому для эффективной работы классификатора часто прибегают к сокращению числа используемых признаков (терминов). На практике для оценки важности слова в контексте документа часто используют статистическую меру TF-IDF. Наибольший вес преобретают термины, которые редко встречаются в других документах, но часто в пределах анализируемого текста.

Построение и обучение классификатора с использованием машинного обучения. Существуют следующие методы классификации:

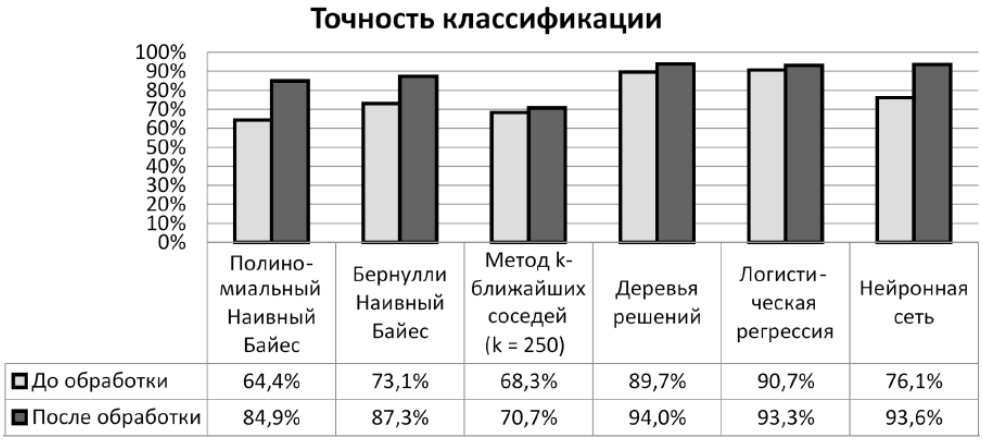
– линейные (метод опорных векторов; логистическая регрессия);

– вероятностные (метод Байеса);

– логические (метод деревьев решений);

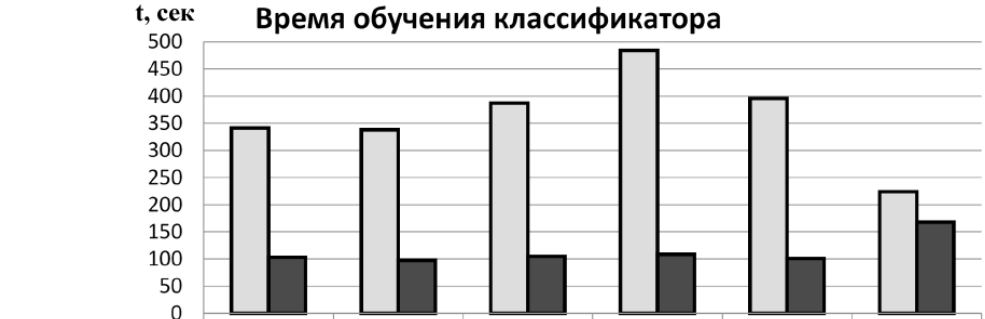
– метрические (метод k ближайших соседей);

– методы на основе искусственных нейронных сетей (нейронные сети прямого распространения, рекуррентные нейронные сети, динамические нейронные сети, сверточные нейронные сети). Каждый метод имеет свои плюсы и минусы и предпочтителен для определенной конкретной задачи. Примером сравнительного анализа методов классификации могут быть графики, приведенные на рис. 2, 3, полученные опытным путем из 3 000 документов [4]. Под обработкой, в данном случае, подразумевается прохождение входящего текствого запроса через предобработку, индексацию и уменьшение размерности пространства признаков.



	Полиномиальный Наивный Байес	Бернулли Наивный Байес	Метод k-ближайших соседей (k = 250)	Деревья решений	Логистическая регрессия	Нейронная сеть
До обработки	64,4%	73,1%	68,3%	89,7%	90,7%	76,1%
После обработки	84,9%	87,3%	70,7%	94,0%	93,3%	93,6%

Рис. 2. Сравнение точности классификационных методов [4]



	Полиномиальный Наивный Байес	Бернулли Наивный Байес	Метод k-ближайших соседей (k = 250)	Деревья решений	Логистическая регрессия	Нейронная сеть
До обработки	341,253	337,991	387,09	484,079	395,827	224
После обработки	103,333	97,306	104,627	108,564	100,879	168

Рис. 3. Сравнение времени обучения разных методов классификации [4]

Оценка качества классификации. Для оценки качества классификации используют обучающую и тестовую выборки. Дополнительно привлекают эксперта. Кроме этого, оценку также выполняют, используя кросс-валидацию.

Оценку вычисляют с помощью сочетания точности и полноты (*F*-мера).

Точность – это количество элементов, правильно отнесенных к некоторому классу, деленное на общее число элементов, отнесенных к этому классу.

Полнота классификации – количество элементов, правильно отнесенных к некоторому классу, деленное на общее число элементов, реально относящихся к этому классу [5].

Результаты. Для предварительной обработки и индексации документа обычно применяется одна из трех моделей: модель «мешка слов», Word2vec и модель, основанная на учете *n*-грамм. Первые две модели требуют компетентности в синтаксисе и морфологии конкретного языка. Модель *n*-грамм не ставит ограничения на использование определенного языка, поэтому порой является оптимальной.

Эффективность классификационных методов оценивают по точности и полноте. На основе проведенного исследования [2] можно сделать вывод, что самые лучшие показатели обоих критериев достигаются при использовании методов опорных векторов (точность 80–85 %, полнота 83–87 %) и сверточных нейронных сетей (точность 90–95 %, полнота 80–85 %). Вместе с тем наивысшая скорость наблюдается у метода Байеса, при этом точность для разных экспериментальных наборов значительно различается (71–90 %). При одновременной обработке текстовых запросов классификация должна выполняться вместе с поступением их из источника. Поэтому предпочтительно выбирать инкрементальные алгоритмы, к которым относятся сверточные нейронные сети и метод опорных векторов.

Качество классификации связано с обучающей выборкой, размер которой необходимо подбирать. Правильный подбор характеристик алгоритмов для обучения является по-прежнему затруднительным делом. Из проведенных экспериментов стало ясно, что для обучения и тестирования классификатора с использованием метода опорных векторов на русском языке и достижения точности 80–85 % нужна размеченная коллекция текстов объемом 1 000–2 000 документов. Для обучения и тестирования классификатора с применением метода сверточных нейронных сетей и достижения точности 90–95 % необходимо собрать и подготовить коллекцию текстов на русском языке объемом около 1 000 000 документов [2].

При сравнительном анализе [4] авторы не обнаружили единственно верного и оптимального метода классификации документов. В каждом частном случае необходимы испытания на конкретных наборах исходных данных исходя из вычислительных мощностей.

В результате проведенного анализа планируется провести классификацию документов на основе нейронной сети с учетом опыта построения классификации текста с помощью нейронной сети на Java [3].

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Википедия. Классификация документов [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Классификация_документов (дата обращения: 01.02.2020).

2. Батура Т. В. Методы автоматической классификации текстов [Электронный ресурс]. – Режим доступа: http://swsys.ru/index.php?page=article&id=4252 (дата обращения: 02.02.2020).

3. @i11. Классификация текста с помощью нейронной сети на Java [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/332078/ (дата обращения: 03.02.2020).

4. Сравнительный анализ методов машинного обучения для решения задачи классификации документов научно-образовательного учреждения [Электронный ресурс] / М. Н. Краснянский, А. Д. Обухов, Е. М. Соломатина, А.А. Воякина. – Режим доступа: http://www.vestnik.vsu.ru/pdf/analiz/2018/03/2018-03-19.pdf (дата обращения: 03.02.2020).

5. Глоссарий важных терминов по машинному обучению [Электронный ресурс]. – Режим доступа: https://docs.microsoft.com/ru-ru/dotnet/machine-learning/resources/glossary#precision (дата обращения: 03.02.2020).

© В. П. Куликов, В. П. Куликова, Е. М. Крылова, Г. Т. Еркебулан, 2020