Мы предлагаем создание Всероссийского информационного ресурса, где по предложенным критериям оценки туристической привлекательности, каждый желающий сможет проставить свои оценки. В итоге можно будет получить объективную экспертную оценку туристической привлекательности региона, обозначить «проседающие» параметры и на основе этой оценки принимать управленческие решения для улучшения туристической привлекательности того или иного региона.

**Литература**

1. Туристической привлекательности региона [Электронный ресурс]. – Режим доступа: http://vestnik.uapa.ru/en/issue/2013/01/14/ (дата обращения: 08.02.2020)

2. Методы оценки туристической привлекательности региона [Электронный ресурс]. – Режим доступа: https://moi-portal.ru/blogi/19692-metody-otsenki-turisticheskoy-privlekatelnosti-regiona/ (дата обращения: 08.02.2020)

3. Першина Н.В. К вопросу о туристической привлекательности региона [Электронный ресурс] / Н.В. Першина, С.В. Угрюмова. К вопросу о туристской привлекательности территории // Молодой ученый. — 2016. — №16. — С. 187-189. — URL https://moluch.ru/archive/120/33147/ (дата обращения: 12.02.2020).

4. Ларичев О.И. Теория и методы принятия решений [Электронный ресурс]. – Режим доступа: https://www.studmed.ru/larichev-oi-teoriya-i-metody-prinyatiya-resheniy_82b487700ad.html (дата обращения: 10.03.2020)

5. Семочкина, Л.Д. Информационная система поддержки путешественника (на территории полуострова Крым) [Текст]/ Л.Д. Семочкина, А.В. Тимофеев // Перспективные информационные технологии (ПИТ 2019). – 2019. –С–842–845.

Я.В. Соловьева, Ю.В. Шабанова

## ИССЛЕДОВАНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

(Самарский университет)

На сегодняшний день в условиях стремительного роста текстовой информации в электронном виде и в связи с потребностью в ней ориентироваться, все более актуальной становится проблема построения универсального классификатора текстов, предоставляющего возможность распределения исходного набора статей по нескольким заранее установленным тематикам в соответствии с их смысловым содержанием [1]. Использование такого классификатора позволит сократить трудозатраты на поиск необходимой информации, представленной электронными текстами, а также ограничить поиск относительно небольшим подмножеством документов.

Различные решения данной задачи находят свое практическое применение в таких областях, как составление тематических каталогов, фильтрация спама, классификация сайтов по тематическим каталогам, обработка документооборота и т.д. В настоящее время примерами классификатов текстов являются такие системы как NNCS (Neural Network Classification & Search), TextAnalyistPro, TextCat, CBTReader, а также проект ДИАЛИНГ, который был разработан специалистами факультета лингвистики РГГУ. Однако все они имеют ряд недостатков: во-первых, это коммерческие проекты, стоимость которых достаточно высока, а во-вторых, эти проекты рассчитаны на профессионального пользователя, следовательно, только обучение использованию предлагаемых пакетов займет слишком много времени.

Наиболее распространенными методами решения данной задачи являются методы машинного обучения и методы, основанные на знаниях. Перспективным направлением на сегодняшний день также считается использование метода опорных векторов и метода деревьев решений в качестве основы подобного рода классификатора [2]. Основным преимуществом данных методов является возможность выявления зависимостей, не поддающихся обнаружению при использовании других подходов обработки информации. Методы опорных векторов и деревьев решений в анализе текстовой информации обладают достаточным быстродействием и не зависят от языка предметной области, но при этом, в отличие от многих алгоритмов обработки текстов дают хорошие результаты.

Целью данной работы является исследование возможностей метода деревьев решений и метода опорных векторов в решении задачи классификации текстов в соответствии с их смысловым содержанием, проектирование и реализация классификатора текстов на естественном языке, а также сравнение результатов, полученных при реализации данных методов.

В качестве входных данных для каждого метода классификации было выбрано 7 тематик, содержащих 150 текстовых фрагментов.

Подход к решению задачи классификации текстов без предварительной обработки данных дает плохие результаты, так как задача обладает рядом особенностей:

1. разреженность пространства;
2. высокая размерность пространства;
3. нестатистический характер данных;
4. большой объем данных.

В данной работе была выполнена предобработка входных векторов-документов в виде векторизации данных TF-IDF, что позволило учесть описанные выше особенности пространства и сделать признаки более информативными.

TF — отношение числа вхождения некоторого слова к общему количеству слов документа. Таким образом, оценивается важность терма t в пределах отдельного документа d.

IDF — инверсия частоты, с которой некоторый терм встречается в документах коллекции. IDF учитывает тот факт, что если терм встречается во многих документах множества, то он не может являться существенным критерием принадлежности документа рубрике и наоборот [3].

В модели TF-IDF каждому терму t документа d ставится в соответствие величина:

$$\mathrm{TF} * \mathrm{IDF} = \mathrm{TF}(d,t) \cdot \log_2\left(\frac{N}{\mathrm{df}(t)}\right)$$

где *tf(d,t)* – это частота терма *t* в документе *d*, *N* – число документов в корпусе, *df(t)* – количество документов в которых встречается терм *t*.

В данном исследовании классификатор на базе деревьев решений представляет собой дерево, узлами которого являются термы $t_k$, каждое ребро обозначено условием $\geq v_k$или $<v_k$, а листья помечены как $c_i$или $\bar{c}_i$. Чтобы классифицировать документ $d_i$в категорию $c_i$или $\bar{c}_i$необходимо пройти по узлам дерева начиная с корня, сравнивая веса терма в документе $w_{kj}$со значениями $v_k$на ребрах (рис. 1).

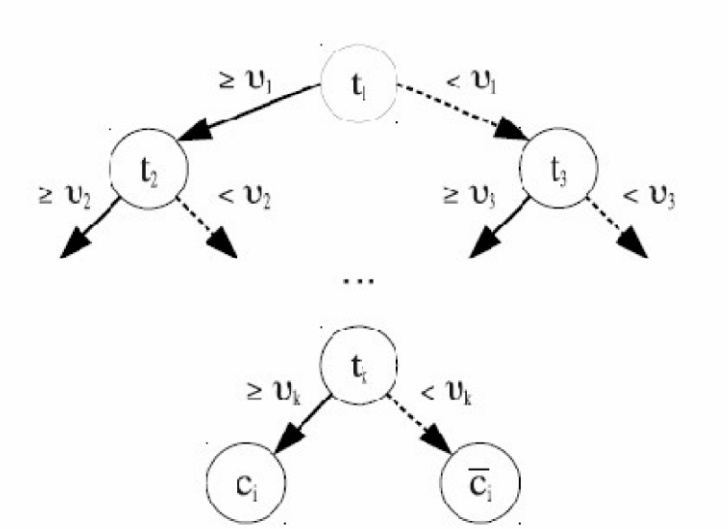


Рис. 1. Дерево решений для категории $c_i$

В данном методе был применен алгоритм CART для усечения дерева, чтобы уменьшить эффект переобучения. В данном алгоритме каждый узел дерева решений имеет двух потомков. На каждом шаге построения дерева правило, формируемое в узле, делит заданное множество примеров на две части – часть, в которой выполняется правило и часть, в которой правило не выполняется [4]. Для выбора оптимального правила используется функция оценки качества разбиения. В алгоритме CART идея неопределенности формализована в индексе Gini:

$$\mathrm{Gini} = 1 - \sum_{i=1}^{n} p_i^2,$$

где T — текущий узел, а $p_i$ — вероятность класса i в узле T.

В таблице 1 приведены результаты точности классификации в зависимости от настраиваемого параметра данного метода — глубины дерева.

Таблица 1. Зависимость точности классификации от глубины дерева

| Глубина дерева | Средняя точность классификации, % |
|---|---|
| 15 | 32 |
| 30 | 40 |
| 45 | 54 |
| 60 | 59 |
| 90 | 61 |
| 100 | 63 |
| 115 | 64 |
| 130 | 62 |

Из таблицы видно, что наилучшие результаты для данной выборки текстов были получены при значении глубины дерева 115. При дальнейшем увеличении глубины наблюдается эффект переобучения.

Метод опорных векторов (SVM) заключается в нахождении гиперплоскости в пространстве признаков, разделяющей его на две части: положительные примеры в одной и отрицательные в другой — у которой минимальное расстояние до ближайших примеров максимально [5]. Некоторая выборка линейно разделима, если в ней возможно получить линейный пороговый классификатор:

$$\text{sign}\left(\sum_{i=1}^{m} w_i * x^i - w_0\right) = sign(\langle w, x\rangle - w_0),$$

где $x = (x^1, \dots, x^n)$ – признаковое описание объекта x; вектор $w = (w^1, \dots, w^n) \in R^n$ и скалярный порог $w_0 \in R$ являются параметрами алгоритма. Таким образом, задача состоит в том, чтобы подобрать значения вектора w такие, при которых функционал, определяющий число ошибок, равен нулю:

$$\sum_{i=1}^{n} [y_i(\langle w, x_i\rangle - w_0) \leq 0] = 0.$$

Наилучшие результаты при использовании метода опорных векторов были получены при установлении предельного параметра ошибки в 1, в качестве ядра была выбрана радиальная базисная функция, коэффициент ядра равен 2. При указанных параметрах средняя точность классификации составила 79%.

В таблице 2 приведены результаты сравнения точности классификации методов опорных векторов и деревьев решений в зависимости от заданных тематик при оптимальных параметрах для каждого метода.

Таблица 2. Результаты классификации текстов

| Тематика | Деревья решений | SVM |
|---|---|---|
| Внешняя экономика | 58% | 80% |
| Налоги | 54% | 69% |
| Медицина | 70% | 83% |
| Туризм | 68% | 95% |
| Недвижимость | 64% | 71% |
| Наука | 47% | 58% |
| Финансы | 69% | 97% |

Полученные результаты позволяют сделать вывод, что метод опорных векторов показал лучшие результаты, чем метод деревьев решений.

Также по результатам исследований было установлено, что один и тот же уровень точности классификации достигается за счет соблюдения баланса между величиной обучающих текстов и их количеством. Чем меньше по размеру обучающие тексты, тем больше их должно быть в обучающем множестве, соответственно, чем фрагменты больше – тем меньшим их количеством можно обойтись.

## Литература

1. Леонтьева, Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы [Текст]: учеб. пособие для вузов/ Н. Н. Леонтьева – М.: Издательский центр «Академия», 2006. – 304 с.

2. Владимир В.В. Математические основы теории машинного обучения и прогнозирования. – МЦМНО, 2013. – 390 с.

3. Губин, М. В. Модели и методы представления текстового документа в системах информационного поиска [Текст]: дис. канд. физ.-мат. наук: защищена 22.03.05: утв. 15.12.05/ М. В. Губин – М., 2005. – 95 с.

4. Деревья решений – CART математический аппарат [Электронный ресурс]. – https://basegroup.ru/community/articles/math-cart-part1(дата обращения 19.03.2018).

5. Воронцов, К.В. Математические методы обучения по прецедентам [Электронный ресурс]. – http://www.ccas.ru/voron/download/SVM.pdf (дата обращения 20.03.2018).