Методы анализа данных в задаче разграничения фольклорных и авторских текстов

Людмила Владимировна Щеголева
Александр Александрович Лебедев[@]
Николай Дмитриевич Москин

Петрозаводский государственный университет, Петрозаводск, Россия;
perevodchik88@yandex.ru

Аннотация: Основной проблемой данного исследования является разграничение фольклорных текстов и текстов, стилизованных под фольклор, при помощи математических методов и компьютерных технологий. Были рассмотрены пять групп текстов: фольклорные песни Заонежья XIX — начала XX века, лужские песни, представляющие собой репертуар Городенского народного хора, стилизованные под фольклор стихотворения Н. А. Клюева, А. К. Толстого и С. А. Есенина. Для сравнения текстов на основе их теоретико-графовых моделей были использованы восемь признаков, с помощью которых в программе R была проведена серия экспериментов с применением пяти методов интеллектуального анализа данных. Все методы показали достаточно высокую среднюю точность распознавания (более 80 %).

Ключевые слова: компьютерная лингвистика, нейронные сети, русская поэзия, фольклор, художественная литература.

Для цитирования: Щеголева Л. В., Лебедев А. А., Москин Н. Д. Методы анализа данных в задаче разграничения фольклорных и авторских текстов. *Вопросы языкознания*, 2020, 2: 61–74.

DOI: 10.31857/S0373658X0008823-4

Methods of data mining in the task of distinguishing between folklore and author's texts

Liudmila V. Shchegoleva
Aleksandr A. Lebedev[@]
Nikolai D. Moskin

Petrozavodsk State University, Petrozavodsk, Russia; perevodchik88@yandex.ru

Abstract: The main problem of the study is the distinction between folklore texts and texts stylized as folklore by means of mathematical methods and computer technologies. Five groups of texts were considered: folklore songs from Zaonezhie of 19th — early 20th century, Luga songs from the repertoire of the Gorodensky folk choir, and poems by N. A. Klyuev, A. K. Tolstoy and S. A. Yesenin stylized as folklore. For comparing texts on the basis of their graph-theoretical models, eight parameters were used. These parameters were used in a series of experiments, carried out in the R environment and involving five methods of data mining. All methods showed a fairly high average recognition accuracy (more than 80 %).

Keywords: computational linguistics, fiction, folklore, neural networks, Russian poetry.

For citation: Shchegoleva L. V., Lebedev A. A., Moskin N. D. Methods of data mining in the task of distinguishing between folklore and author's texts. *Voprosy Jazykoznanija*, 2020, 2: 61–74.

DOI: 10.31857/S0373658X0008823-4

1. Вводные замечания

Проблема разграничения фольклорных текстов и текстов, стилизованных под фольклор, была заявлена сравнительно давно. Одним из первых отечественных классиков-филологов, обратившихся к решению данного вопроса, был М. М. Бахтин, в научном творчестве которого эта тема была затронута в связи с анализом жанра романа, темой хронотопа, а также с рассмотрением фольклорных элементов в творчестве Ф. Рабле [Бахтин 1995]. Также следует выделить труд «Поэтика сказа», в котором проанализирована сказовая форма повествования в целом и сделан вывод о ее значительной жизнеспособности: по мнению авторов, сказ «вновь и вновь возрождается на различных этапах литературного развития» [Мущенко и др. 1978], в частности, на творческом пути авторов XX в.

В целом правомерно поставить вопрос об определении различий между двумя этими группами текстов, причем как с точки зрения рядового читателя, так и с позиции специалиста. Насколько глубоко необходимо проникать в структуру текста, чтобы выявить разницу между настоящим фольклорным текстом и литературным произведением, которое грамотно и умело стилизовано под фольклорное? Можно поставить задачу разработки определенного инструмента, который позволил бы различать между собой тексты разных групп; в идеале распределение текстов на группы должно включать в себя не только работу филолога-эксперта, но и элементы автоматизации, что сделало бы анализ более объективным и упростило бы лингвистическую обработку.

Исследование было построено на материале пяти групп текстов.

1. Фольклорные песни Заонежья XIX — начала XX в. (80 текстов) — материал песен, исполнявшихся во время бесёдных увеселений в Олонецкой губернии, который дан с опорой на монографию Р. Б. Калашниковой [1999] «Бесёды и бесёдные песни Заонежья второй половины XIX века».
2. Лужские песни (50 текстов) — репертуар Городенского народного хора — наследника певучих деревень Лужского уезда, продолжающего их «фольклорную традицию, без разрыва, трансформаций» [Песни 1990: 6].
3. Тексты Н. А. Клюева [1999], стилизованные под фольклор (50 текстов). В творчестве Клюева «народные» элементы и близость к русской культуре традиционно привлекали внимание как читателей, так и исследователей стихотворного текста. Живая народная речь поэзии Клюева, особая образность, смысловая насыщенность текстов сближает авторские тексты с фольклорной традицией.
4. Тексты А. К. Толстого [1969], стилизованные под фольклор (50 текстов). Поэтическое творчество Толстого было обращено к разным жанрам фольклора, содержательно наполнено типичными для народных произведений сюжетами и героями. Близость автора к фольклорной традиции проявлялась как в смысловом, так и в поэтическом аспекте, что сближает его стихотворения с образцами народного творчества.
5. Тексты С. А. Есенина [1995], стилизованные под фольклор (20 текстов). Стихотворения Есенина широко известны отечественному читателю в том числе и потому, что сближены с народной поэзией, а творчество поэта рассматривается исследователями-филологами, в частности, в контексте фольклорной традиции.

Помимо обособленного рассмотрения каждой из групп, вряд ли можно поставить под сомнение противопоставленность фольклорных и литературных текстов как таковую (т. е. текстов групп 1 и 2 в противовес текстам групп 3, 4 и 5): несмотря на талант авторов и их способности к стилизации текста, сама структура стихотворного произведения будет отличаться от структуры фольклорного текста на разных уровнях языка. Всегда ли такая задача, связанная с разграничением текстов, может быть решена филологом, равно как и рядовым читателем? Предлагаемая статья представляет собой одну из попыток определения объективных критериев, отличающих между собой разные группы текстов,

а также фольклорные произведения и тексты-стилизации на синтаксическом и семантическом уровне.

Ряд работ последнего времени также посвящен попыткам фольклорных стилизаций в творчестве литераторов, причем чаще анализу подвергается прозаическое творчество [Хатямова 2006; Завгородняя 2010; Орлов, Осминин 2010; Ястребова 2011]. Если говорить об анализе стилизации в поэзии, то следует отметить статью Т. В. Мануковской [2015], в которой автор приводит набор приемов фольклорной стилизации, наблюдаемых в солдатских песнях Клюева, а также исследование [Барахнин и др. 2017], где с помощью разных алгоритмов классификации анализируется лицейская лирика А. С. Пушкина в контексте жанрово-стилевого соответствия.

Проделанные исследования удерживают в центре внимания в первую очередь лексический уровень языка, а синтаксическая и семантическая составляющие текстов-стилизаций зачастую оказываются на исследовательской периферии. Важной видится задача определения критериев, разграничивающих фольклорные тексты и литературные тексты, стилизованные под фольклор, что позволило бы впоследствии упростить и автоматизировать распределение текстов на две группы с применением компьютерных технологий и разных математических методов.

С точки зрения лингвистического исследования и изучения индивидуально-авторского стиля перспективными и интересными видятся наблюдения над теми произведениями, которые были классифицированы неверно (отнесены к другой группе текстов). Подобное несоответствие может объясняться разными причинами: в частности, возможна авторская стилизация высокого качества, в том числе и на синтактико-семантическом уровне (если неверно атрибутирован литературный текст) либо необычное построение фольклорного текста (если неверно атрибутировано народное произведение). Такие особенности изначально могут не заметить не только рядовые читатели, но и эксперты-филологи; применение математических методов и специальных программ позволяет акцентировать внимание на подобных несоответствиях и дать пищу для размышлений лингвистам.

Полученные результаты могут быть применены и к интернет-фольклору [Алексеевский 2010], изучение которого стало одним из перспективных направлений исследований последних десятилетий, поскольку открывает «возможности для изучения ряда вопросов, которые остаются практически неисследованными в рамках традиционной фольклористики» [Радченко 2011: 418], таких как история развития текста и получение данных о числе воспроизведений текста в тот или иной момент времени.

2. Теоретико-графовая модель текстов и ее характеристики

Как уже отмечалось ранее, для решения поставленной задачи можно использовать математические методы и современные компьютерные технологии. Подобные методы уже нашли свое применение, например, в задаче установления авторства текстов (атрибуции). В [Батура 2012; Романов 2010; Рогов и др. 2014] представлен обзор существующих работ по этой тематике (в них в основном рассмотрены исследования русскоязычных текстов). В зарубежных источниках также можно найти множество работ по атрибуции текстов, см., например, [Stamatatos 2009; Sebastiani 2002].

Следует отметить, что одним из важных исследовательских направлений для Петрозаводского государственного университета является проблема атрибуции текстов, в том числе и с привлечением математических методов (в числе наиболее значимых работ отметим [Захаров и др. 2000; Рогов и др. 2014]; более подробно о проблемах атрибуции текстов Ф. М. Достоевского и проделанных в этой сфере исследованиях см. [Алексеева 2015]). Предлагаемые нами методика и принципы обработки художественного текста, равно как и результаты текущего этапа исследования, могут быть использованы в работах

этого направления (разумеется, при условии уточнения специфики анализируемого текста и спектра применяемых математических методов).

В данной работе рассматривается теоретико-графовый подход для описания синтаксической составляющей фольклорных и авторских текстов. Теоретико-графовая модель представляет собой множество вершин и множество ребер, соединяющих эти вершины. При этом вершины и ребра «нагружены» дополнительной смысловой информацией (атрибутами). Определим более строго теоретико-графовую модель как четверку G = (V, E, α, β), в которой:

— V — конечное непустое множество вершин;
— $E \subseteq V \times V$ — множество ребер;
— α: $V \rightarrow L_V$ — функция, задающая метки вершинам графа;
— β: $E \rightarrow L_E$ — функция, задающая метки ребрам графа,

где L_V и L_E — множества меток и атрибутов объектов и отношений, определенных в некоторой предметной области.

В монографии [Москин 2013: 22–29] подробно описывается подобная модель на материале бесёдных песен Заонежья XIX — начала XX в. В качестве примера рассмотрим, как на основе бесёдной песни «По ельничку, по березничку…» (в записи К. М. Петрова, 1868 г.) строится граф:

По ельничку, по березничку…
По частому молодому олешничку
Оттоль молодец идет,
За собой девку ведет.
Идучи красной девице
Выговаривает,
Выговаривает, сам обманывает:
«Не ходи, красна девица,
На беседушки;
На беседушки,
Ко суседушку.
У суседушка своя бедушка
Своя бедушка… — невестушка,
Она…, она…
Семерых привечат,
Семерых братцев родных,
В восьмых дядюшку,
В девятых лакей,
Во десятых свой мужик
Во постелюшке лежит».

В этом тексте присутствуют следующие объекты: *ельничек*, *березничек*, *частый молодой олешничек*, *молодец*, *красна девица*, *беседушка*, *суседушка*, *бедушка*, *невестушка*, *семеро братцев родных*, *дядюшка*, *лакей*, *мужик*, *постелюшка*. Между объектами устанавливаются связи двух типов: локальные и глобальные. Локальная связь существует в том случае, когда отношение между объектами выражено непосредственно в тексте (глаголом, глагольной формой или прилагательным в позиции предиката). В ситуации, когда глагол пропущен, но предполагается, что он есть (в неполных или эллиптических предложениях), связь существует. Например, между вершинами «*молодец*» и «*красна девица*» существуют три локальные связи, которые соответствуют глаголам «*ведет*», «*выговаривает*» и «*обманывает*». Глобальная связь реализуется в том случае, если связь между объектами не выражена в тексте в виде словоформы, но существует и осознается читающим (отношения принадлежности или месторасположения).

Если объединить одинаковые объекты в одну вершину, то подобную структуру можно представить в виде единого графа синтаксической структуры песни (рис. 1). Здесь локальные связи показаны сплошной линией, а глобальная связь пунктиром.

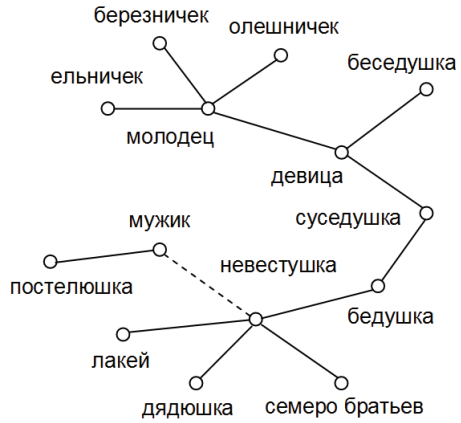

Рис. 1. Теоретико-графовая модель синтаксической структуры песни «По ельничку, по березничку»

Рассмотрим также стихотворение Н. А. Клюева [1999: 217–228] «Я ко любушке-голубушке ходил» 1914 г.:

Я ко любушке-голубушке ходил,
Голубую однорядку износил,
Шубу беличью повыволочил,
Коробейку мелких денег издержал,
Разлюбезной воркованьем докучал:
Я куплю тебе гостинец — скатну нить,
Буду баско оболоченой водить,
Разлюби ты дегтегона-лесника,
Лаптевяза, да Мирона-резчика,
Не подмигивай торговому в ряду,
Не обранивай платочка на ходу,
Протопопу белой ручкой не маши,
Не заглядывай в рыбачьи шалаши,
У калачника не мешкай в куреню,
Не давай овса гонецкому коню,
На гонца крутую бровь не наводи,
Чтобы сердце не кровавилось в груди!
У гонца не застоялая душа, —
В торбе ложка и походная лапша.
Он тебя за белояровый овес
Доведет до неуемных горьких слез,
Что ль до зыбки — непотребного лубка,
До отцовского глухого кулака,
Будет зыбочка поскрипывать,
Красна девушка повздыхивать!

В этом тексте присутствует значительно больше объектов и связей между ними (рис. 2). Как видно из выстроенной теоретико-графовой модели, центральными в тексте

оказываются три объекта: герой, от лица которого ведется повествование; возлюбленная героя; один из потенциальных конкурентов героя в борьбе за руку и сердце девушки. Разделение остальных объектов по трем группам также неслучайно — герой (*я*) вначале рассказывает о том, что он сделал, чтобы добраться до девушки (*любушка*), а затем последовательно объясняет возлюбленной, с кем ей не следует иметь дело (*лесник*, *лаптевяз*, *резчик*, *протопоп* и т. п.), подробно останавливаясь на одном из главных конкурентов (*гонец*) и описывая присущие ему особенности вкупе с теми последствиями, которые ждут возлюбленную, если она полюбит гонца, а не главного героя. Некоторые из объектов связаны с несколькими центральными объектами: например, *любушка*, в соответствии с указаниями, не должна давать овса *коню* (эта связь выражена локально с помощью глагола *давай*); в то же время, *конь* является *гонецким*, т. е. принадлежит *гонцу*, что формирует глобальную связь между двумя этими объектами.

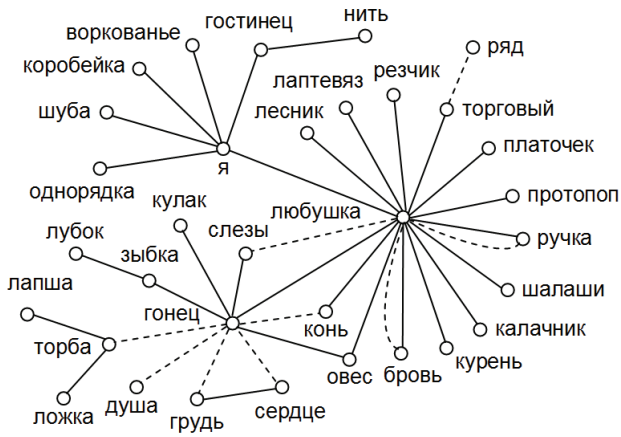

Рис. 2. Теоретико-графовая модель синтаксической структуры стихотворения «Я ко любушке-голубушке ходил»

Определим следующие восемь числовых характеристик текста, формируемых исходя из текста и его теоретико-графовой модели, которые далее будут использоваться для машинного обучения:

1. Число слов (word).
2. Число строк в тексте (string).
3. Число вершин графа *m* (vertex).
4. Число ребер графа *n* (edge).
5. Максимальная степень вершины (max) — это максимальное количество ребер, инцидентных вершине (соединенных с вершиной).
6. Параметр связности *C* (link) находится по формуле $C = \frac{2m'}{m(m-1)}$, где *m* — число вершин в графе, *m′* — число пар вершин, связанных между собой [Москин 2013: 58]. Независимо от кратности ребер пара вершин учитывается только один раз. Параметр *C* принимает значения на отрезке от 0 до 1. При этом *C=0* соответствует нуль-графу, а *C=1* — полному графу.
7. Коэффициенты гиперболической регрессии *a* и *b* [Москин 2013: 64]. Объекты в фольклорных песнях и авторских текстах неравнозначны между собой. В сюжете часто доминируют два основных персонажа (вершины с максимальной степенью), а остальные объекты являются второстепенными, они встречаются в тексте реже, большинство один-два раза. Если сопоставить каждой вершине графа ее степень

и отсортировать их в порядке убывания, то получатся такие диаграммы, как на рис. 3 (построенные на примере песен «Мальчик ты, мальчик» и «Все мужья до жен добры» в записи В. Д. Дашкова). Подобную зависимость можно интерполировать гиперболической кривой вида $y = \frac{a}{x} + b$, предварительно нормировав исходный массив: $y'_i = \frac{y_i}{2m} \cdot 100\ \%$, где m — число вершин в графе, x_i — номер (ранг) вершины в отсортированном по убыванию списке, y_i — степень вершины x_i.

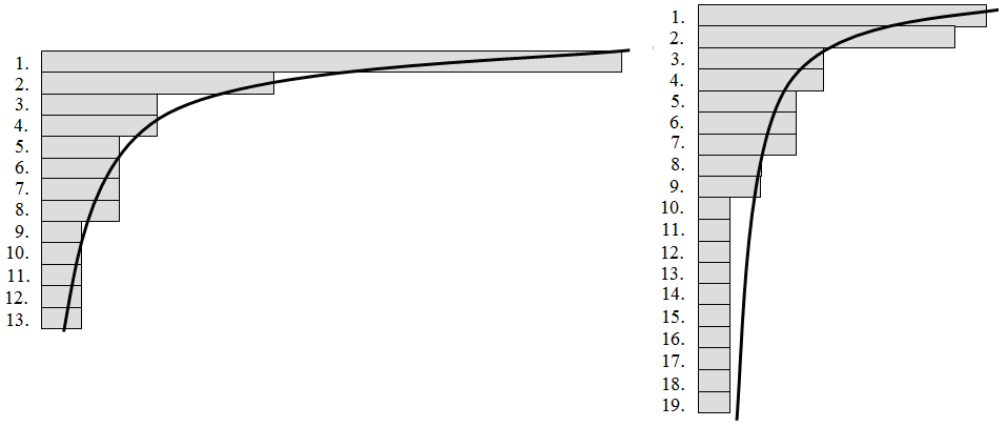

Рис. 3. Распределение степеней вершин по убыванию

Тогда коэффициенты регрессии a и b находятся по формулам

$$a = \frac{m \cdot \sum_{i=1}^{m} \frac{y_i}{x_i} - \sum_{i=1}^{m} \frac{1}{x_i} \cdot \sum_{i=1}^{m} y_i}{m \cdot \sum_{i=1}^{m} \frac{1}{x_i^2} - \left(\sum_{i=1}^{m} \frac{1}{x_i} \right)^2}, \quad b = \frac{\sum_{i=1}^{m} y_i \cdot \sum_{i=1}^{m} \frac{1}{x_i^2} - \sum_{i=1}^{m} \frac{1}{x_i} \cdot \sum_{i=1}^{m} \frac{y_i}{x_i}}{m \cdot \sum_{i=1}^{m} \frac{1}{x_i^2} - \left(\sum_{i=1}^{m} \frac{1}{x_i} \right)^2}.$$

8. Процент глобальных связей от общего числа связей (global).

Для построения теоретико-графовых моделей текстов и подсчета числовых характеристик была использована информационная система «Фольклор» [Москин 2013].

3. Результаты применения методов машинного обучения с учителем

Для построения классификаторов были использованы следующие методы, реализованные в пакете R [Шитиков, Мастицкий 2017]:

— дерево решений;
— дискриминантный анализ;
— метод опорных векторов;
— нейронная сеть;
— случайный лес.

Подобные методы выбраны неслучайно. Они доказали свою эффективность в задачах классификации текстов [Андреев 2003; Ермолаева 2009; Lai et al. 2015; Yang, Liu 1999], машинном переводе [Wu et al. 2016], поиске плагиата [Subroto, Selamat 2014; Engels et al. 2007], фильтрации спама [Мироненко 2012], гендерной идентификации автора произведения [Sboev et al. 2018], определения тональности текстов [dos Santos, Gatti 2014; Socher et al. 2013], определения жанра текстов [Барахнин и др. 2017; Орлов, Осминин 2010] и др.

Для обучения и тестирования классификаторов была использована коллекция из 250 текстов, описанных во введении. Бесёдные песни и лужские песни образуют кластер из 130 фольклорных текстов, а стихотворения трех авторов — кластер из 120 текстов, стилизованных под фольклорные. В качестве признаков были использованы описанные ранее в разделе 2 восемь числовых характеристик текста, обозначенных как word, string, vertex, edge, max, link, a, b, global. Фрагмент исходных данных для пяти текстов представлен в табл. 1.

Таблица 1

Фрагмент исходных данных

Название	Автор	Фольклор	word	string	vertex	edge	max	link	a	b	global
Мальчик ты, мальчик	Бесёдные песни	Да	99	31	11	17	12	0,22	39,86	–1,85	0,18
Кто у нас хороший	Лужские песни	Да	93	34	13	22	16	0,167	40,15	–2,13	0,14
Деревцо мое миндальное	А. К. Толстой	Нет	27	8	7	8	6	0,333	37,78	0,29	0,25
Из-за леса лесу темного	Н. А. Клюев	Нет	106	27	21	20	11	0,095	26,07	0,24	0,45
Сыплет черемуха снегом	С. А. Есенин	Нет	64	16	22	15	7	0,056	21,78	0,89	0,13

Для тестирования качества работы классификаторов был использован подход на основе метода кросс-валидации со случайным выделением одного блока для обучающей коллекции и одного блока для тестовой коллекции.

Было проведено 20 экспериментов, в каждом из которых случайным образом выбирались 25 текстов (10 %) для тестовой коллекции, остальные 225 текстов выступали в качестве обучающей коллекции.

По результатам построения классификатора на основе обучающей коллекции проводилась классификация как обучающей коллекции, так и тестовой коллекции и рассчитывался показатель точности классификации. Результаты представлены в табл. 2.

Все методы показали достаточно высокую среднюю точность на 20 экспериментах — более 80 % для тестовой коллекции, что говорит о применимости показателей теоретико-графовой модели для различия фольклорных и стилизованных текстов.

При этом исключительным свойством дерева решений является то, что этот метод позволяет выделить характерные особенности значений параметров для фольклорных и стилизованных текстов. На рис. 4 представлен один из полученных вариантов дерева решений, где фольклорные тексты обозначены «1», а стилизованные «2». Каждая вершина дерева, обозначенная овалом или прямоугольником, содержит количество текстов, отнесенных классификатором к фольклорным, и количество текстов, отнесенных классификатором к стилизованным, разделенные наклонной чертой. Каждая дуга дерева обозначена меткой в виде неравенства, в левой части которого указано имя признака, а в правой — критериальное значение, определяющее правило, по которому тексты разделяются на две подгруппы.

Таблица 2

Значения точности классификации на обучающей и тестовой коллекциях

	Обучающая коллекция			Тестовая коллекция		
Метод / Точность	**средняя**	**минимальная**	**максимальная**	**средняя**	**минимальная**	**максимальная**
Дерево решений	0,88	0,84	0,91	0,84	0,64	0,96
Дискриминантный анализ	0,87	0,85	0,90	0,87	0,76	1
Метод опорных векторов	0,91	0,88	0,92	0,87	0,76	1
Нейронная сеть	0,99	0,93	1	0,81	0,56	0,96
Случайный лес	1	1	1	0,87	0,72	0,96

На основе дерева решений можно выделить три основных правила классификации на фольклорные и стилизованные тексты:

1. Если (link ≥ 0,129) ∧ (global < 0,4451) ∧ (a ≥ 16,38) ∧ (vertex < 13,5), то текст — фольклорный (80 текстов классифицируются правильно, 6 текстов классифицируются неправильно);
2. Если (link < 0,129) ∧ (string < 27,5), то текст — стилизация (76 классифицируются правильно, 5 текстов — неправильно);
3. Если (link ≥ 0,129) ∧ (global < 0,4451) ∧ (a ≥ 16,38) ∧ (vertex ≥ 13,5) ∧ (string ≥ 17,5), то текст — фольклорный (19 текстов классифицируются правильно, 1 текст — неправильно).

Были проведены эксперименты по кластеризации текстов на две группы методами главных компонент, иерархическими деревьями, сетями Кохонена. Все методы показали, что полученные разбиения не соответствуют разбиению на фольклорные и стилизованные тексты. Это говорит о том, что в целом рассматриваемые характеристики не явно выражены в фольклорных текстах и вместе взятые формируют особый классификационный

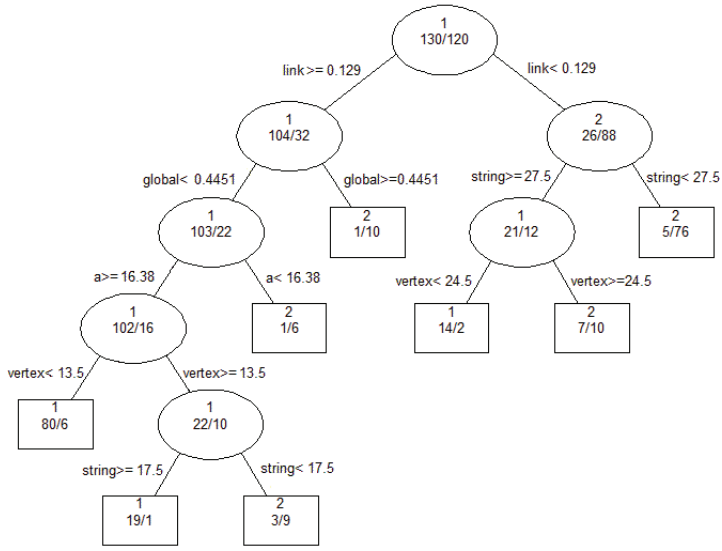

Рис. 4. Дерево решений

профиль каждого произведения. При этом определенные комбинации характеристик, как показали ранее рассмотренные методы, все же позволяют разграничить фольклорные и стилизованные под фольклор тексты.

4. Интерпретация результатов

Как уже было отмечено, использованные в ходе экспериментального анализа методы продемонстрировали высокую среднюю точность, но некоторые из проанализированных текстов определялись классификатором ошибочно при использовании любого из методов. Таким образом, числовые характеристики неверно классифицированных фольклорных текстов совпадали с характеристиками кластера текстов, стилизованных под фольклор; и наоборот, некоторые стилизованные тексты были близки по характеристикам к фольклорному кластеру. Принципиально важна интерпретация результатов, включающая в себя определение особенностей тех текстов, которые распознаются при помощи различных методов неверно — в сочетании с попыткой лингвистической интерпретации подобного рода ошибочных распознаваний такая интерпретация может стать ключом к более совершенному применению методов интеллектуального анализа данных для решения задачи разграничения фольклорных и авторских текстов.

В соответствии со структурой дерева принятий решений видно, что одним из ключевых факторов, влияющих на отнесение группы к фольклорным либо литературным текстам соответственно, является параметр связности (link). Это объяснимо самой структурой повествования — если в центре фольклорного произведения почти всегда стоит только один или два объекта (как правило, парень и девушка), то литературные произведения, даже будучи стилизованными под фольклор, демонстрируют в этом аспекте существенно большее разнообразие.

В дополнение к параметру связности следует упомянуть и процент глобальных связей (global). Фольклорные тексты демонстрируют меньший процент глобальных связей, поскольку в таких произведениях существенно больше глагольных конструкций, все события выражены напрямую в тексте, а число подразумеваемых сравнений невелико. Напротив, литературное произведение может быть наполнено связями, с одной стороны, находящимися в сознании автора, с другой — без особых затруднений интерпретируемыми читателем, но в то же время не выраженными вербально непосредственно в тексте.

Как правило, проанализированные тексты, как литературные, так и фольклорные, соответствуют приведенным выше обоснованиям; некоторым неверно классифицированным примерам можно дать объяснение, связанное с учетом именно двух этих параметров. К примеру, текст «Был я в городе Кронштадте…», принадлежащий к группе фольклорных песен, содержит в себе необычно большое количество глобальных связей, построенных по принципу «человек — часть тела человека»:

На шеюшке много жемчужка,
Во ушках серьги жемчужны,
На головке розовый платок,
На ногах башмачики у ней козловы

Это не слишком типично для фольклорного текста (как в целом, так и в рассматриваемых нами произведениях), что вызывает очевидные классификационные затруднения. Со схожей ситуацией можно столкнуться и при рассмотрении текста лужских песен «Не по сахару речка течет…», где в центре произведения больше двух объектов, а число глобальных связей «человек — часть тела человека», а также «объект — его месторасположение» сравнительно велико.

Определенные литературные стилизации можно принять за фольклорные тексты благодаря высокому качеству подобной стилизации. Например, между стихотворением Клюева «Я сгорела молоденька без огня» и бесёдной песней «Все мужья до жен добры» немало общего в структуре, что, по сути дела, приводит к неправильной классификации текста некоторыми классификаторами.

Также одним из факторов неверного определения может стать объем текста: некоторые произведения, как фольклорные, так и литературные («Там, где капустные грядки…», «Выйду на сини, на сиверочек…», «Деревцо мое миндальное…», «По подлавочью варган, варган, варган…») недостаточно объемны, в них не так много объектов и связей между ними, а потому не столь выразительно проявляются признаки, глобально отличающие литературные тексты от фольклорных.

В то же время некоторые случаи ошибочной классификации видятся не вполне объяснимыми даже при точечном анализе. В частности, тексты Толстого «Ты знаешь, я люблю там, за лазурным сводом…» и «Замолкнул гром, шуметь гроза устала…», а также лужские песни «Как у брода было брода…» и «Сине море на волнах стоит, да…» будут без особых затруднений правильно классифицированы читателем как литературные и фольклорные тексты соответственно, однако классифицируются в ходе разбора неправильно. Присутствие таких ошибочно определенных текстов указывает на необходимость расширения исследования (например, путем введения дополнительных параметров анализа текста, позволяющих на уровне структуры найти отличия между фольклорными и литературными произведениями).

Заключение

В статье была рассмотрена проблема разграничения фольклорных текстов и текстов, стилизованных под фольклор, при помощи математических методов и компьютерных технологий. Были рассмотрены пять групп текстов: фольклорные песни Заонежья, лужские песни, стилизованные под фольклор стихотворения Н. А. Клюева, А. К. Толстого и С. А. Есенина (всего 250 текстов). С помощью информационной системы «Фольклор» экспертом-филологом были построены теоретико-графовые модели, отражающие синтаксическую структуру текстов.

Для сравнения текстов на основе их теоретико-графовых моделей были использованы восемь признаков, с помощью которых в программе R была проведена серия экспериментов с применением пяти методов интеллектуального анализа данных (дерево решений, дискриминантный анализ, метод опорных векторов, нейронная сеть, случайный лес). Все методы показали достаточно высокую среднюю точность классификации (более 80 %). Анализ правильных и ошибочных результатов классификации позволил более четко сформулировать отличительные особенности фольклорных и стилизованных текстов на уровне их синтаксической структуры. Это свидетельствует о применимости показателей теоретико-графовой модели текста для различия фольклорных и стилизованных текстов с использованием методов автоматической классификации.

СПИСОК ИСТОЧНИКОВ

Есенин 1995 — Есенин С. А. *Полное собрание сочинений в 7 т*. Т. 1: Стихотворения. М.: Наука; Голос, 1995.

Клюев 1999 — Клюев Н. А. *Сердце единорога.* Стихотворения и поэмы. СПб.: РХГИ, 1999.

Песни 1990 — *Песни городенского хора* / Сост., предисл., нотация напевов Е. Е. Васильевой. Новгород: ОНМЦ, 1990.

Толстой 1969 — Толстой А. К. *Собрание сочинений*: *в 4 т*. Т. 1. М.: Правда, 1969.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Алексеева 2015 — Алексеева Л. В. Проблемы атрибуции в исследованиях о Ф. М. Достоевском (обзор предложенных решений). *Неизвестный Достоевский*, 2015, 4: 3–10. [Alekseeva L. V. Problems of attribution in studies of Fyodor Dostoevsky: A survey of proposals. *Neizvestnyi Dostoevskii*, 2015, 4: 3–10.]

Алексеевский 2010 — Алексеевский М. Д. Интернет в фольклоре или фольклор в Интернете? (современная фольклористика и виртуальная реальность). *От Конгресса к Конгрессу. Навстречу Второму Всероссийскому конгрессу фольклористов*: Сб. материалов. М.: ГРЦРФ, 2010, 151–166. [Alekseevskii M. D. Internet in the folklore or folklore in the Internet? Modern folklore studies and virtuality. *Ot Kongressa k Kongressu. Navstrechu Vtoromu Vserossiiskomu kongressu fol'kloristov*: Conf. proceedings. Moscow: State Republic Center for Russian Folklore, 2010, 151–166.]

Андреев 2003 — Андреев В. С. Классификация стихотворных текстов методом дискриминантного анализа. *Математическая морфология: электронный математический и медико-биологический журнал*, 2003, 5(1): 58–70. [Andreev V. S. Classifying poetry by means of discriminant analysis. *Matematicheskaya morfologiya: elektronnyi matematicheskii i mediko-biologicheskii zhurnal*, 2003, 5(1): 58–70.]

Барахнин и др. 2017 — Барахнин В. Б., Кожемякина О. Ю., Пастушков И. С., Рычкова Е. В. Автоматизированная классификация русских поэтических текстов по жанрам и стилям. *Вестник НГУ. Сер.: Лингвистика и межкультурная коммуникация*, 2017, 3: 13–23. [Barakhnin V. B., Kozhemyakina O. Yu., Pastushkov I. S., Rychkova E. V. Automatic classification of Russian poetical texts by genre and style. *Vestnik NGU. Seriya: Lingvistika i mezhkul'turnaya kommunikatsiya*, 2017, 3: 13–23.]

Батура 2012 — Батура Т. В. Формальные методы определения авторства текстов. *Вестник НГУ. Сер.: Информационные технологии*, 2012, 10(4): 81–94. [Batura T. V. Formal methods of text attribution. *Vestnik NGU. Seriya: Informatsionnye tekhnologii*, 2012, 10(4): 81–94.]

Бахтин 1995 — Бахтин М. М. Формы времени и хронотопа в романе. Очерки по исторической поэтике. *Вопросы литературы и эстетики*. М.: Художественная литература, 1995, 234–405. [Bakhtin M. M. Forms of time and chronotopos in a novel. Essays in historical poetics. *Voprosy literatury i estetiki*. Moscow: Khudozhestvennaya Literatura, 1995, 234–405.]

Ермолаева 2009 — Ермолаева Ю. Е. Классификация стихотворных текстов методом дискриминантного анализа. *Вестник Тамбовского университета*, 2009, 7(75): 292–296. [Ermolaeva Yu. E. Classifying poetry by means of discriminant analysis. *Vestnik Tambovskogo universiteta*, 2009, 7(75): 292–296.]

Завгородняя 2010 — Завгородняя Г. Ю. Фольклорная стилизация в романе П. И. Мельникова-Печерского «В лесах». *Русская речь*, 2010, 5: 111–114. [Zavgorodnyaya G. Yu. Folklore stylization in Pavel Melnikov-Pechersky's novel 'In the Forests'. *Russkaya rech'*, 2010, 5: 111–114.]

Захаров и др. 2000 — Захаров В. Н., Рогов А. А., Сидоров Ю. В. Поиск грамматического инварианта Ф. М. Достоевского методами статистического анализа. *Труды Петрозаводского государственного университета: Сер. «Прикладная математика и информатика»*, 2000, 9: 67–80. [Zakharov V. N., Rogov A. A., Sidorov Yu. V. Searching for the grammatical invariant of Fyodor Dostoevsky by means of statistical analysis. *Trudy Petrozavodskogo gosudarstvennogo universiteta: Seriya «Prikladnaya matematika i informatika»*, 2000, 9: 67–80.]

Калашникова 1999 — Калашникова Р. Б. *Бесёды и бесёдные песни Заонежья второй половины XIX века*. Петрозаводск: Изд-во ПетрГУ, 1999. [Kalashnikova R. B. *Besedy i besednye pesni Zaonezhya vtoroi poloviny XIX veka* ["Besyody" and "besyodnye pesni" of the second half of 19th century in Zaonezhie]. Petrozavodsk: Petrozavodsk State Univ. Publ., 1999.]

Мануковская 2015 — Мануковская Т. В. Фольклорная стилизация в солдатских песнях Николая Клюева. *Вестник ВГУ. Сер.: Филология. Журналистика*, 2015, 1: 35–40. [Manukovskaya T. V. Folklore stylization in soldier songs of Nikolay Klyuev. *Vestnik VGU. Seriya: Filologiya. Zhurnalistika*, 2015, 1: 35–40.]

Мироненко 2012 — Мироненко А. Н. *Алгоритм контентной фильтрации спама на базе совмещения метода опорных векторов и нейронных сетей*. Дис. … канд. техн. наук. Омск: Омский гос. ун-т им. Ф. М. Достоевского, 2012. [Mironenko A. N. *Algoritm kontentnoi fil'tratsii spama na baze*

sovmeshcheniya metoda opornykh vektorov i neironnykh setei [Algorithm of content-filtration of spam on the basis of support vector machine and neural networks]. Ph.D. diss. Omsk: Dostoevsky Omsk State Univ., 2012.]

Москин 2013 — Москин Н. Д. *Теоретико-графовые модели фольклорных текстов и методы их анализа*. Петрозаводск: Изд-во ПетрГУ, 2013. [Moskin N. D. *Teoretiko-grafovye modeli fol'klornykh tekstov i metody ikh analiza* [Graph-theoretical models of folklore texts and methods of their analysis]. Petrozavodsk: Petrozavodsk State Univ. Publ., 2013.]

Мущенко и др. 1978 — Мущенко Е. Г., Скобелев В. П., Кройчик Л. Е. *Поэтика сказа*. Воронеж: Изд-во Воронежского ун-та, 1978. [Mushchenko E. G., Skobelev V. P., Kroichik L. E. *Poetika skaza* [Poetics of "skaz"]. Voronezh: Voronezh Univ. Publ., 1978.]

Орлов, Осминин 2010 — Орлов Ю. Н., Осминин К. П. Определение жанра и автора литературного произведения статистическими методами. *Прикладная информатика*, 2010, 2: 95–108. [Orlov Yu. N., Osminin K. P. Determining genre and authorship of literature with statistical methods. *Prikladnaya informatika*, 2010, 2: 95–108.]

Радченко 2011 — Радченко Д. А. Сетевой фольклор: перспективы исследования. *Комплексные исследования традиционной культуры в постсоветский период*. Вып. 14. М.: ГРЦРФ, 2011, 417–427. [Radchenko D. A. Web folklore: Research perspectives. *Kompleksnye issledovaniya traditsionnoi kul'tury v postsovetskii period*. No. 14. Moscow: State Republic Center for Russian Folklore, 2011, 417–427.]

Рогов и др. 2014 — Рогов А. А., Седов А. В., Сидоров Ю. В., Суровцова Т. Г. *Математические методы атрибуции текстов*. Петрозаводск: Изд-во ПетрГУ, 2014. [Rogov A. A., Sedov A. V., Sidorov Yu. V., Surovtsova T. G. *Matematicheskie metody atributsii tekstov* [Mathematical methods of text attribution]. Pertozavodsk: Petrozavodsk State Univ. Publ., 2014.]

Романов 2010 — Романов А. С. *Методика и программный комплекс для идентификации автора неизвестного текста*. Дис. … канд. техн. наук. Томск: ТГУ, 2010. [Romanov A. S. *Metodika i programmnyi kompleks dlya identifikatsii avtora neizvestnogo teksta* [Methodics and program complex for identification of text authorship]. Ph.D. diss. Tomsk: Tomsk State Univ., 2010.]

Хатямова 2006 — Хатямова М. А. Фольклорная стилизация в малой прозе Е. И. Замятина. *Вестник ТГПУ*, 2006, 8: 68–75. [Khatyamova M. A. Folklore stylization in Evgeny Zamyatin's minor prose. *Vestnik TGPU*, 2006, 8: 68–75.]

Шитиков, Мастицкий 2017 — Шитиков В. К., Мастицкий С. Э. *Классификация, регрессия и другие алгоритмы Data Mining с использованием R*. Тольятти; Лондон [б. и.], 2017. [Shitikov V. K., Mastitskii S. E. *Klassifikatsiya, regressiya i drugie algoritmy Data Mining s ispol'zovaniem R* [Classification, regression and other algorithms of Data Mining using R]. Tolyatti; London, 2017.]

Ястребова 2011 — Ястребова Н. Г. Фольклорная стилизация в повести Ф. Н. Глинки «Лука да Марья». *Русская речь*, 2011, 3: 110–117. [Yastrebova N. G. Folklore stylization in Fyodor Glinka's 'Luka and Marya'. *Russkaya rech'*, 2011, 3: 110–117.]

dos Santos, Gatti 2014 — dos Santos C. N., Gatti M. Deep convolutional neural networks for sentiment analysis of short texts. *Proc. of COLING 2014, the 25th International Conf. on Computational Linguistics, Dublin, Ireland, August 23-29, 2014*. Dublin, 2014, 69–78.

Engels et al. 2007 — Engels S., Lakshmanan V., Craig M. Plagiarism detection using feature-based neural networks. *Proc. of the 38th SIGCSE Technical Symposium on Computer Science Education*, 2007, 39(1): 34–38.

Lai et al. 2015 — Lai S., Xu L., Liu K., Zhao J. Recurrent convolutional neural networks for text classification. *Proc. of the 29th AAAI Conf. on Artificial Intelligence*, 2015: 2267–2273.

Sboev et al. 2018 — Sboev A., Moloshnikov I., Gudovskikh D., Selivanov A., Rybka R., Litvinova T. Deep learning neural nets versus traditional machine learning in gender identification of authors of RusProfiling texts. *Procedia Computer Science*, 2018, 123: 424–431.

Sebastiani 2002 — Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys*, 2002, 34(1): 1–47.

Socher et al. 2013 — Socher R., Perelygin A., Wu J. Y., Chuang J., Manning C. D., Ng A. Y., Potts C. Recursive deep models for semantic compositionality over a sentiment treebank. *Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing*. Seattle; Washington, 2013, 1631–1642.

Stamatatos 2009 — Stamatatos E. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 2009, 60(3): 538–556.

Subroto, Selamat 2014 — Subroto I., Selamat A. Plagiarism detection through Internet using hybrid artificial neural network and support vectors machine. *TELKOMNIKA*, 2014, 12(1): 209–218.

Wu et al. 2016 — Wu Y., Schuster M., Chen Z., Le Q. V., Norouzi M., Macherey W. et al. *Google's neural machine translation system: Bridging the gap between human and machine translation*. URL: https://arxiv.org/abs/1609.08144 (arXiv preprint).

Yang, Liu 1999 — Yang Y. M., Liu X. A re-examination of text categorization methods. *Proc. of the 22nd International Conf. on Research and Development in Information Retrieval*. Berkeley: Univ. of California, 1999, 42–49.

Получено / received 02.12.2018 *Принято / accepted 17.09.2019*