§ 5 СИСТЕМНЫЙ АНАЛИЗ, ПОИСК, АНАЛИЗ И ФИЛЬТРАЦИЯ ИНФОРМАЦИИ

Батура Т.В.

МЕТОДЫ ОПРЕДЕЛЕНИЯ АВТОРСКОГО СТИЛЯ ТЕКСТОВ И ИХ ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Аннотация: *Статья представляет собой обзор формальных методов атрибуции текстов. Задачи определения авторства текстов встречаются в различных областях и представляют интерес для филологов, литературоведов, историков, юристов. При решении задачи атрибуции наибольший интерес и наибольшую сложность представляет анализ синтаксического, лексико-фразеологического и стилистического уровней текста. В некотором смысле более узкой задачей является задача сентимент-анализа (определения тональности текста). Методы ее решения могут оказаться полезными при определении автора текста. К сожалению, экспертный анализ авторского стиля является трудоемким и длительным процессом. Целесообразно создание новых подходов, позволяющих хотя бы частично автоматизировать деятельность экспертов. Поэтому в статье уделяется внимание именно формальным методам идентификации авторов текстов и программной реализации этих методов. В настоящее время для атрибуции текстов применяются алгоритмы сжатия данных, методы математической статистики и теории вероятностей, алгоритмы нейронных сетей, кластерного анализа и др. В статье приведено описание наиболее известных на сегодняшний день программных систем для определения авторского стиля текстов на русском языке, предпринята попытка произвести их сравнительный анализ, выявить особенности и недостатки рассмотренных подходов. Среди проблем, затрудняющих исследования в области атрибуции, можно выделить проблему выбора лингвостилистических параметров текста и проблему составления выборки эталонных текстов. Необходимо проводить дальнейшие исследования, направленные на поиск новых или совершенствование уже имеющихся методов атрибуции текстов, на поиск характеристик, позволяющих четко разделять стили авторов, в том числе на коротких текстах и на малых объемах выборки.*

Ключевые слова: *атрибуция текста, определение авторства, формальные параметры текста, авторский стиль, классификация текстов, машинное обучение, стастистический анализ, компьютерная лингвистика, идентификация стиля автора, анализ текстовой информации.*

DOI: 10.7256/2305-6061.2014.2.11705

Введение.

Для определения автора текста зачастую приходится обращаться к экспертам. Эксперты могут идентифицировать автора неизвестного текста или определить принадлежность произведения другому автору при помощи характерных языковых особенностей, стилистических приемов.

Важно отметить, что задача установления авторства текстов (задача атрибуции) встречается в различных областях и представляет интерес для филологов, литературоведов, историков, юристов, криминалистов. Несомненно, экспертный анализ авторского стиля является трудоемким процессом. Поэтому с развитием компьютерных технологий все больше появляется необходимость в создании формальных методов решения задачи атрибуции, создании различных программных приложений для автоматизации деятельности экспертов. По этой причине в данной статье уделяется внимание именно формальным методам определения авторов текстов.

С другой стороны, помимо художественной и научной литературы бывает необходимо установить автора текстов новостного, рекламного или другого характера. В последнее время стремительно увеличивается количество форумов, блогов, онлайн-обзоров и отзывов, сообщений в социальных сетях. Это тоже расширяет круг применения задачи атрибуции. В некотором смысле более узкой задачей можно считать задачу определения тональности текста, когда требуется определить по тексту эмоциональную оценку мнения автора или выявить в исходном тексте эмоционально окрашенную лексику. Поэтому создание новых формальных методов сентимент-анализа также представляет интерес. При автоматизации обеих задач возможно применение методов классификации.

В настоящее время для атрибуции текстов применяются подходы из теории распознавания образов, математической статистики и теории вероятностей, алгоритмы нейронных сетей и кластерного анализа и многие другие. Программные продукты, существующие на сегодняшний день, позволяют учитывать и варьировать различные лингвостатистические параметры, характеризующие текст с разных сторон. В статье приведен обзор различных формальных методов определения авторства текстов, предпринята попытка выявить особенности и недостатки рассмотренных методов, сравнить программные продукты по атрибуции текстов на русском языке.

2. Лингвистические и статистические параметры текстов.

Наиболее полная классификация основных формальных методов атрибуции текстов содержится, например, в работе [1]. Согласно ей все формальные методы можно разделить на статистические и методы машинного обучения. В случае одномерного статистического анализа применяются: критерий Стьюдента, хи-квадрат Пирсона, двусторонний критерий Фишера, QSUM. В случае многомерного обычно используются: критерий Колмогорова-Смирнова, цепи Маркова, метод главных компонент, энтропийный подход, алгоритмы кластерного анализа и др. Среди методов машинного обучения эффективными оказы-

DOI: 10.7256/2305-6061.2014.2.11705

ваются: Байесовский классификатор, нейронные сети, деревья решений, метод опорных векторов и некоторые другие.

Формальные методы чаще всего основаны на сравнении вычислимых характеристик текстов, как в теории распознавания образов. Применение теории распознавания образов в задаче атрибуции текстов можно встретить, например, в [2] и [3]. В общем случае текст представляется в виде вектора параметров, каждый из которых объективно характеризует некоторый набор особенностей текста. При такой формализации автор также может быть представлен в виде аналогичного вектора параметров, ассоциированных с текстами, которые написаны данным автором.

В большинстве случаев в качестве характеризующих параметров текста выбираются те или иные его статистические характеристики: количество использования определенных частей речи, некоторых конкретных слов, знаков препинания, фразеологизмов, архаизмов, редких и иностранных слов, количество и длина предложений (измеренная в словах, слогах, знаках), объем словаря, количество полнозначных и служебных слов, средняя длина предложения, отношение числа глаголов к общему количеству словоупотреблений в тексте и т.д.

Основная проблема формальных методов анализа авторства состоит как раз в выборе параметров. Как было отмечено Марковым А.А. [4], существует целый ряд формальных статистических характеристик текстов, непригодных для определения авторства в силу одного из двух недостатков.

Отсутствие устойчивости. Разброс значений параметра для текстов одного и того же автора настолько велик, что диапазоны возможных значений для разных авторов перекрываются. Очевидно, данный параметр не поможет различать авторов, а при использовании в составе группы параметров лишь сыграет роль дополнительного шума.

Отсутствие различающей способности. Параметр может принимать близкие значения для всех или большинства авторов, поскольку его значение определяются свойствами языка, на котором написаны тексты, а не индивидуальными особенностями создателя текста.

Поэтому параметры, используемые в формальных методиках определения авторства, должны предварительно исследоваться на устойчивость и различающую способность, желательно на текстах большого количества различных авторов. В работе [5] выделены три условия применимости формального параметра: условие массовости, устойчивости и различающей способности. Примером формальной характеристики, удовлетворяющей всем трем условиям, является так называемый *авторский инвариант*. Он вычисляется как процент содержания служебных слов (союзов, предлогов, частиц) в тексте. Правда недостатком такого подхода является низкая разделительная способность в случае большого количества авторов (потенциально метод может разделять только 10 авторских стилей).

Массовость. Параметр должен опираться на те характеристики текста, которые слабо контролируются автором на сознательном уровне. Это условие необходимо, чтобы устранить возможность сознательного искажения автором характерного для него стиля или имитации стиля другого автора.

DOI: 10.7256/2305-6061.2014.2.11705

Устойчивость. Параметр должен сохранять постоянное значение для одного автора. Естественно, в силу случайных причин некоторое отклонение значений от среднего неизбежно, но оно должно быть достаточно мало.

Различающая способность. В наилучшем случае, параметр должен принимать существенно различные значения для разных авторов, превышающие колебания, возможные для одного автора. Необходимо отметить, что выбрать параметры, которые гарантированно разделяют двух любых авторов, очень трудно. Какими бы ни были параметры, всегда существует вероятность того, что два или более авторов окажутся по данным параметрам близки. Поэтому на практике считается достаточным, чтобы параметр позволял уверенно различать между собой разные группы авторов, то есть существовало достаточно большое количество групп авторов, для которых средние значения параметра существенно различаются. Параметр, очевидно, не поможет различить тексты авторов из одной группы, но позволит уверенно различать тексты авторов, попавших в разные группы. Различать тексты авторов одной группы можно за счет использования одновременно достаточно большого вектора различных по характеру параметров. В этом случае вероятность случайного совпадения станет заметно меньше. Для уверенного вывода в отношении текстов, для которых формально вычисленное параметрическое расстояние мало, требуется дополнительное исследование экспертными методами.

2.1. Методы распознавания образов.

Проблема атрибуции текстов часто сводится к сравнению двух коллекций: текстов, автор которых известен заранее (эталонный набор текстов), и текстов неизвестного автора (экспериментальный набор текстов). В общем случае задача идентификации авторства может быть представлено в следующем виде (см. рис. 1).

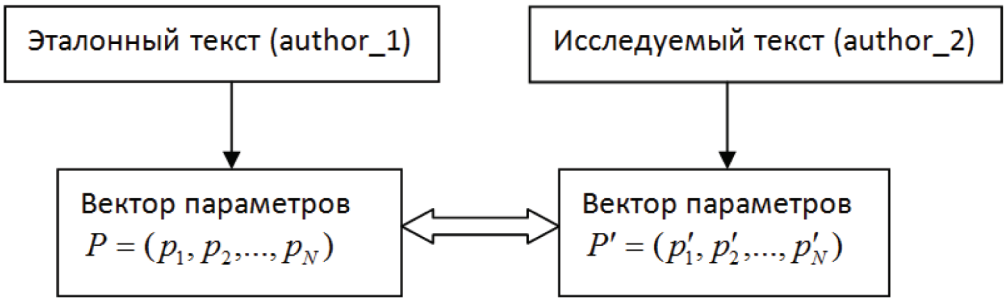

Рис. 1. Сравнение векторов параметров

Расстояние является интегральной характеристикой различия текстов. Если расстояние $\rho(P, P')$ велико, то считаем, что авторы текстов разные; если расстояние мало, то автор тот же самый (автор_1 = автор_2).

Например, в работах [2] и [3] для атрибуции были выбраны 13 комедий Мольера в стихах. Существует гипотеза, что автором большинства стихотворных пьес, приписывае-

DOI: 10.7256/2305-6061.2014.2.11705

мых Мольеру, является П. Корнель и некоторые другие французские драматурги. Поэтому основной целью исследования стало применение математических методов для решения проблемы «Корнель–Мольер».

Количество предложений в выбранных произведениях варьировалось от 72 до 1293. Для описания эталонных текстов был взят 51 параметр. Из полученного априорного словаря выбрали небольшое количество информативных параметров при помощи схемы М.М. Бонгарда, предусматривающей двухступенчатое свертывание параметрического пространства. На первом этапе производилось автоматическое разбиение априорного набора информативных параметров на два подмножества, релевантных и нерелевантных для различения эталонных классов. На втором этапе релевантность параметров определялась по t-критерию Стьюдента (см. раздел 4). Если наблюдаемое значение критерия было больше порогового, то параметр относился к числу информативных, в противном случае он исключался из дальнейшего рассмотрения. Информативный набор состоял из пяти параметров: число элементарных и сочиненных предложений, число спрягаемых форм глагола, подлежащих и местоимений-подлежащих.

В результате эксперимента были получены следующие результаты. Автором 6 пьес (из 13 рассмотренных) с точностью 95% является П. Корнель, 4 другие пьесы с точностью 63–73% были также отнесены к произведениям П. Корнеля. Одна пьеса с точностью 68% была отнесена к авторству Ф. Кино. Две оставшихся пьесы были отнесены к отдельному апостериорному классу.

3. Алгоритмы сжатия данных.

3.1. Система «Лингвоанализатор».

Ряд исследований был проведен Д.В. Хмелёвым [6], [7], результатом которых явился вывод об эффективности применения алгоритмов сжатия данных для задачи определения авторства. Также был сделан вывод о том, что простейший подход с использованием цепей Маркова первого порядка показывает хорошие результаты на файлах большого объема и плохие по сравнению с другими методами на отрывках длиной в 2000–5000 символов. Этот метод был реализован в системе «Лингвоанализатор» (URL: http://www.rusf.ru/books/analysis).

Существенное преимущество метода энтропийной классификации (с помощью сжатия) состоит в отсутствии предварительной обработки текста. Суть метода в том, чтобы добавлять текст, автор которого неизвестен, к тексту, принадлежащему конкретному автору, и смотреть, насколько хорошо сжимается эта «добавка». Правильный исходный класс документа – это тот, на котором он сжимается лучше всего.

Метод, предложенный Хмелёвым Д.В., основан на применении относительной энтропии. Есть несколько способов вычислить эту характеристику: «шельфовый» (off-the-shelf) алгоритм, метод предсказания по частичному совпадению (PPM – Prediction by Partial Matching) и использование индекса повторяемости.

Среди алгоритмов сжатия данных без потерь наиболее часто встречающимися явля-

DOI: 10.7256/2305-6061.2014.2.11705

ются: кодирование Хаффмана, арифметическое кодирование, метод Барроуза-Уилера и множество вариаций метода Лемпеля-Зива (LZ). К алгоритмам, специально ориентированным на сжатие текста, относятся следующие: PPM использует Марковскую модель небольшого порядка и DMC (Dynamic Markov compression) использует динамически изменяемую Марковскую модель. В рамках подхода PPM правильный исходный класс документа – это тот, на чьей модели получается наилучшее сжатие. Каждый алгоритм имеет большое число модификаций и параметров (например, существует динамическое кодирование Хаффмана, варьируется объем используемого словаря и пр.). Кроме того, существует множество «смешанных» алгоритмов, где текст, сжатый, например, с помощью алгоритма PPM, дополнительно кодируется с помощью кода Хаффмана.

Все эти алгоритмы реализованы в различных программах, которых в настоящий момент существует довольно много. Каждая из них реализует разные варианты алгоритмов сжатия данных. Дополнительное разнообразие возникает из-за того, что у многих программ имеется несколько версий, которые также имеют разные алгоритмы сжатия. В работе [6] приведены некоторые результаты эксперимента по сравнению точности определения авторства текста с использованием алгоритмов сжатия данных.

Ряд экспериментов проводился на массиве новостей агентства Рейтерс (Reuters Corpus Volume 1). Было отобрано 50 авторов с наибольшим объемом статей, всего 1813 статей. Выборка случайно была разбита на 10 равных частей, одна из которых использовалась для тестирования. Лучший результат был получен для метода с применением программы **rar** (точность 89,4%).

Другой ряд экспериментов был проведен на корпусе текстов, состоящем из 385 текстов 82 писателей. Тексты подверглись предварительной обработке. Во-первых, были склеены все слова, разделенные переносом. Далее были выкинуты все слова, начинавшиеся с прописной буквы. Оставшиеся слова помещены в том порядке, в каком они находились в исходном тексте с разделителем из символа перевода строки. У каждого из писателей было отобрано по контрольному произведению. Остальные тексты были объединены в обучающие тексты. Объем каждого контрольного произведения составлял не менее 50000–100000 символов. Проведенные исследования показали, что программы сжатия угадывают истинных писателей весьма часто на текстах большого объема. Особенно хорошо проявляет себя программа **rarw** (точность 71%), результаты применения которой превосходят реализацию других подходов в этой области. Тем не менее, остаются и открытые вопросы. Например, почему использование программы **rarw**, применяющей модификацию алгоритма LZ, на файлах большого объема опережает многие другие методы, также применяющие модификацию LZ.

4. Методы теории вероятностей и математической статистики.

В некотором роде продолжение данного исследования нашло себя в работе [8]. Предлагаемый метод основан на учете статистики употребления пар элементов любой природы, идущих друг за другом в тексте (букв, морфем, словоформ и т.п.), т.е. на формальной математической модели последовательности букв (и любых других элементов)

DOI: 10.7256/2305-6061.2014.2.11705

текста как реализации цепи Маркова. По тем произведениям автора, которые достоверно им созданы, вычислялась матрица переходных частот употребления пар элементов (букв, грамматических классов слов и т.п.). Она служила оценкой матрицы вероятности перехода из элемента в элемент. Для каждого автора строилась матрица переходных частот и оценивалась вероятность того, что именно он написал анонимный текст (или фрагмент текста). Автором анонимного текста считался тот, для кого вычисленная оценка вероятности больше.

Исходный корпус текстов в результате предварительной обработки был представлен в следующих четырех вариантах:

1. пары букв в их естественных последовательностях в тексте – в словах (в той форме, в которой они употреблены в тексте) и пробелах между ними;
2. пары букв в последовательностях букв в приведенных (словарных, лемматизованных или исходных) формах слов; например, предыдущее предложение в таком случае предстает в виде «пара буква в последовательность буква приведенный словарный лемматизованный или исходный форма слово»;
3. пары наиболее обобщенных грамматических классов слов в их последовательностях в предложениях текста. К таким классам слов относят части речи (существительные, глаголы, прилагательные и т.п.) и некоторые условных категории вроде «конец предложения», «сокращение» и др.
4. пары менее обобщенных грамматических классов слов. К ним относятся такие семантико-грамматические разряды, как одушевленные существительные, неодушевленные существительные, прилагательные качественные, относительные, притяжательные и т.п.

В процессе предварительной обработки отбрасывались все слова, для которых не удалось автоматически определить грамматический класс, все знаки препинания, все слова с заглавной буквы, склеивались все слова, разделенные переносом. Каждый символ кодировался числом.

Была произведена перекрестная проверка метода на материале 385 текстов 82 авторов. Показателем точности метода являлся процент правильно определенных произведений. Для варианта 1) была достигнута точность 73%, для 2) – 62%, для 3) – 61%. На материале варианта 4) получены существенно худшие результаты 4%.

В работе [9] показано, что последовательность символов текста не обладает свойствами простой цепи Маркова. Таким образом, гипотеза, выдвинутая в [7] и [8] опровергнута. Тем не менее, на основе проведенных в [8] экспериментов был сделан вывод, что использование пар подряд идущих в тексте букв, дает более точные результаты, чем использование таких языковых категорий, как одиночные грамматические классы слов и их пары. Поэтому выдвинуто предположение, что в буквенных парных структурах частично отображаются полные структуры морфем словоформ текста – префиксальные, корневые, суффиксальные и флективные. Тем самым, довольно большой объем словоизменительной и словообразовательной информации о структуре русских слов оказывается отображенным в статистике парной встречаемости букв, что и определяет довольно высокий

DOI: 10.7256/2305-6061.2014.2.11705

уровень эффективности использования этой статистики для определения авторства текста. Другими словами, подсчет частот употреблений пар букв позволяет учесть информацию о словаре, который используется автором, а также, косвенно, информацию о предпочитаемых им грамматических конструкциях.

4.1. Система «Атрибутор».

Как продолжение развития подхода, использующего в качестве стилевых признаков бинарные буквосочетания, А.Н. Тимашев [10] предложил применять трехбуквенные сочетания – триады. При таком методе анализу поддаются однобуквенные и двухбуквенные служебные слова, а это значительная часть наиболее частотных предлогов, союзов, частиц и междометий, которые традиционно считаются значимыми стилеметрическими показателями. По этой причине двухбуквенные, четырех- и более буквенные цепочки менее показательны, что и было доказано в процессе исследования.

На основе данных рассуждений был создан программный продукт для автоматического сравнения и классификации текстов по параметрам индивидуального авторского стиля под названием «Атрибутор» (URL: http://www.textology.ru/web.htm).

База этой программы содержит произведения 103 авторов и использует экспертную обработку текстов. В эталонную выборку, на которой происходило обучение «Атрибутора», попали в основном романы и повести отечественных писателей XIX–XX веков. Пополнение шло за счет ресурсов известных электронных библиотек, наибольшее количество текстов было получено в библиотеке М. Мошкова (URL: http://lib.ru). Выборка подбиралась таким образом, чтобы тексты разных писателей в максимальной степени различались друг от друга, а тексты одного писателя были максимально близки. Те случаи, когда известный писатель в какой-то период своего творчества резко менял стиль изложения, отсеивались.

Для обработки текста «Атрибутором» необходимо, чтобы его длина была не меньше 6 страниц. Ограничение на длину текста накладывается для того, чтобы избежать ошибок, связанных со сравнением статистически несопоставимых объектов. В обработку попадают все слова текста за исключением имен собственных.

4.2. Система «СМАЛТ».

Еще одна система атрибуции текстов «СМАЛТ» (Статистические методы анализа литературного текста) описана в [11] и [12]. Система основана на алгоритмах автоматизации морфологического и синтаксического анализа текстов. Обработка текстов в этой системе производится в несколько этапов. На первом шаге выполняется автоматизированное разбиение исходного текста на лексические единицы, среди которых выделяются части (или разделы), абзацы, предложения, слова. На втором этапе осуществляется автоматическая обработка текста и его морфологический разбор. На базе построенного морфологического разбора производится третья стадия обработки текста – синтаксический анализ.

DOI: 10.7256/2305-6061.2014.2.11705

Базой данных литературных произведений для проведения исследований явилась 81 публицистическая статья 60–70 гг. XIX в. журналов «Время» и «Эпоха». Цель исследования – установить, является ли действительным автором выбранных статей Ф.М. Достоевский.

В работе [12] была выдвинута гипотеза об эффективности выполнения некоторых методов для анализа текстов: метода проверки гипотез с помощью критерия Стьюдента, критерия Колмогорова-Смирнова на согласованность с заданным распределением, методов кластерного анализа, методики «сильный граф», в которой в качестве основной характеристики текстов рассматривалась матрица частот парной встречаемости грамматических классов слов.

Для проведения эксперимента с помощью t-критерия Стьюдента в качестве параметров были взяты следующие величины: средняя длина слова в буквах, средняя длина предложения в словах, индекс разнообразия лексики (отношения числа разных словоформ к числу словоупотреблений). Проводилось исследование с выборками разных объемов: в 200, 300, 400, 500 и 600 слов. В итоге, были получены числовые значения критерия Стьюдента для всех статей. Для случая независимых выборок статистика критерия равна:

$$t = \frac{\bar{x} - \bar{y}}{\sigma_{x-y}}, \text{ где}$$

$\bar{x}$ – среднее арифметическое в экспериментальной группе;

$\bar{y}$– среднее арифметическое в эталонной группе;

σ_{x-y} – стандартная ошибка разности средних арифметических и вычисляется по формуле:

$$\sigma_{x-y} = \sqrt{\frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}, \text{ где}$$

n_1 – размер первой выборки;

n_2 – размер второй выборки.

Среди группы статей Ф.М. Достоевского выявлялась статья с максимальным значением t-характеристики. Среди группы атрибутируемых статей и статей других авторов исключались статьи со значением t-характеристики, большим фиксированного.

При работе с такими параметрами, как общее распределение длины слова, общее распределение длины предложения, лексический спектр текста на уровне словаря и лексический спектр текста на уровне текста ставилась задача определения вероятности того, что распределения длин слов в буквах в двух статьях, одна из которых объединение статей Ф.М. Достоевского, взяты из одной и той же «генеральной совокупности» и могут рассматриваться как управляемые одними и теми же закономерностями. Для этого использовался непараметрический критерий Колмогорова-Смирнова. Использовались частотные словари на каждые 500 слов текста. Все словоформы распределились в группы по 1, 2, …, 10 раз встречаемости в выборке. Далее определялось число словоформ в каждой группе, что означает распределение частот на уровне словаря, и покрываемость текста, что означает распределение частот на уровне текста.

В результате экспериментов не удалось установить, является ли автором рассматри-

DOI: 10.7256/2305-6061.2014.2.11705

ваемых статей Ф.М. Достоевский, т. к. обе гипотезы (о том, что Ф.М. Достоевский – автор, и о том, что не автор) не верны. Причем была доказана независимость результатов исследования от видов текстов (авторская или современная орфография и пунктуация).

В методе иерархической кластеризации использовались две меры расстояния между объектами: Евклидова мера и мера Чебышева. Для определения расстояния между кластерами использовались методы ближнего и дальнего соседа. Исследование проводилось на основе двух наборов признаков: основного, состоящего их частей речи (16 признаков) и расширенного, с подключением дополнительных морфологических параметров, например падеж, род и т.п. (156 признаков). Применение методов корреляционных плеяд и иерархической кластеризации показало неэффективность использования формально-грамматических параметров для классификации исследуемых статей с целью решения задачи атрибуции. Более того, было доказано, что увеличение числа этих параметров не улучшает результаты исследования.

Применение методики «сильный граф», основанной на изучении закономерностей расположения частей речи в рамках предложения по определенным параметрам, не позволило четко и однозначно ответить на вопрос о принадлежности ряда статей Ф.М. Достоевскому.

Еще один недостаток предложенных методов состоит в том, что задачу определения авторства приходится сводить к задаче построения качественного и быстрого синтаксического анализатора. Последняя из задач является не менее трудной и, на сегодняшний день, до сих пор не решена на требуемом уровне.

4.3. Система «Антиплагиат».

Среди программных продуктов для определения авторства текстов можно выделить систему «Антиплагиат» [13]. Этот интернет-сервис предлагает осуществить проверку текстовых документов на наличие заимствований из общедоступных сетевых источников. Система позволяет проводить атрибуцию текстов на различных языках.

На первом этапе система собирает информацию из различных источников: загружает из Интернета и обрабатывает сайты, находящиеся в открытом доступе, базы научных статей и рефератов. Загруженные документы проходят процедуру фильтрации, в результате которой отбрасывается бесполезная с точки зрения потенциального цитирования информация (например, HTML-страницы с большим количеством рекламы, новостные заголовки и т.д.).

На следующем этапе каждый из полученных таким образом текстов определенным образом форматируется и заносится в системную базу данных. Кроме того, в общую базу текстов поступают документы, загруженные на проверку пользователем, если такая возможность была разрешена им во время процедуры загрузки. Все пользовательские документы, загружаемые для проверки, ставятся в очередь на обработку.

Поиск совпадений осуществляется методом сравнения последовательностей символов без учета языковых особенностей и речевых взаимосвязей. За счет этого достигается высокая, в несколько секунд, скорость поиска совпадений. Проверка документа, напри-

DOI: 10.7256/2305-6061.2014.2.11705

мер, реферата среднего размера, занимает несколько секунд. После проверки документа, пользователь получает отчет, в котором представляются результаты. Структура отчета позволяет выделять в проверяемом тексте заимствованные части как по всем источникам, так и по их любому подмножеству.

Все программные алгоритмы, используемые в «Антиплагиате», являются коммерческой тайной компании «Форексис», и открытого доступа к ним нет. К недостаткам системы можно отнести невозможность «отлавливать» заимствованный текст при условии, что в каждом из предложений текста добавлено или убрано всего лишь одно слово. На данный момент существуют программы, например «Антиплагиат киллер» (URL: http://hakcity2.ru/file/411), позволяющие «обходить» систему «Антиплагиат».

5. Гибридный подход с использованием нейронных сетей.

5.1. Система «Стилеанализатор».

Проблему атрибуции текстов в работах [9] и [14] предлагается решать при помощи нейронных сетей и методов иерархической кластеризации. В качестве меры сравнения матриц частот появления признаков в исследовании использовалась мера Кульбака и мера хи-квадрат. В работе также показано, что мера Хмелёва из [7] является частным случаем меры Кульбака, вычисляемой по формуле

$$\rho(x,y)=\sqrt{\sum_i\left(x_i\log_2\frac{2x_i}{x_i+y_i}+y_i\log_2\frac{2y_i}{x_i+y_i}\right)}.$$

Под частотным признаком понимается любой признак стиля текста, допускающий возможность нахождения частоты его появления в тексте (например, число появления абзацев в тексте). Мера хи-квадрат вычисляется по формуле

$$\chi^2=\frac{(f_o-f_e)^2}{f_e},\ \text{где}$$

f_o – фактические (наблюдаемые) частоты событий;
f_e – ожидаемые частоты событий.

Для экспериментов был разработан программный комплекс «Стилеанализатор». Проводились [9] исследования зависимости качества классификации текстов (по авторству, по жанровым типам и источникам) от объемов текстовых фрагментов. Для этого применялись: метод Хмелёва, деревья решений и метод с использованием нейронных сетей. В экспериментах было взято два набора текстов: художественных произведений (156 текстов, три подмножества: 30, 20 и 10 авторов) и газетных статьей (5697 текстов, 57 журналистов за 2003–2004 гг.). Рассмотрены количественные признаки трех уровней: уровня букв, уровня слов и уровня предложений. Всего 14 различных наборов признаков.

Было обнаружено, что для разных текстов, с разным числом классов, для разных наборов признаков существует примерно постоянное минимальное значение объема

DOI: 10.7256/2305-6061.2014.2.11705

фрагментов для приемлемой классификации (точность 90–98%). Оно составляет примерно 30000–40000 символов, или 5000–6000 слов, или 400–600 предложений.

Использовались нейронные сети, обучающиеся без учителя и предназначенные для обработки больших массивов многомерной информации – самоорганизующиеся карты Кохонена (Self-organizing map – SOM). За последние годы это направление является одним из наиболее развивающихся. С помощью SOM-сетей решаются многие проблемы классификации, обработки естественного языка, изображений, тестирования и обучения. Несмотря на широкое использование, SOM-сетям не хватает теоретической обоснованности – они опираются в основном на эмпирические результаты.

В итоге был получен вывод о том, что в случае удачного нахождения универсального набора характеристик можно обрабатывать любое число авторов и текстов (большие массивы информации). Достаточно постоянно модифицировать карту, добавляя новые произведения и оценивать, как они взаимодействуют с ранее присутствующими.

Одним из серьезных недостатков метода является невозможность прогнозирования успешного результата. Генетический поиск на заданном наборе текстов может никогда не найти хороший вариант для разделения характеристик. Нет никакого критерия того, в правильном ли направлении движется поиск, верно ли он делает скачки, нужную ли скапливает информацию об исследуемом пространстве. Исследователь сам должен производить мониторинг поиска и следить за всеми «поворотами событий». Кроме того, нет механизмов, определяющих, сколько времени осталось до конца работы алгоритма, до того момента, когда дальнейший поиск не принесет своих результатов.

Другой проблемой метода является его трудоемкость. Число загруженных текстов, которое напрямую влияет на качество поиска, требует больших ресурсов от вычислительной системы (большой объем памяти и мощный процессор). Для нахождения по-настоящему универсальных характеристик необходимо обработать большие корпуса текстов, чтобы можно было с уверенностью заявить об их универсальности.

Проведенные эксперименты показали, что метод Хмелёва и его модификации выигрывают как в скорости обучения, так и в качестве классификации. Нейронные сети дают сопоставимое качество, но сильно проигрывают в скорости. Деревья решений обеспечивают наихудшее качество классификации, но при этом дают наглядный вид решения и по ходу производят отбор самых информативных признаков.

5.2. Система «Авторовед».

Продолжение исследований по применению нейронных сетей в сочетании с методом опорных векторов при установлении авторства текстов нашло отражение в работах [1] и [15]. Если задачу определения авторства сформулировать как задачу классификации, то одним из широко применяемых выходов является построение бинарного классификатора. Все тексты, включая обучающую часть выборки, разворачиваются в очень большой вектор, индексируемый словами. После этого имеется два множества точек из обучающей выборки в многомерном пространстве: принадлежащие данному автору и не принадлежащие автору. Для того, чтобы разделить эти множества, нужно поделить

DOI: 10.7256/2305-6061.2014.2.11705

пространство на две части. Самый простой способ сделать это – построить гиперплоскость. Такую гиперплоскость можно построить с помощью метода опорных векторов (SVM – Support Vector Machines). После этого для классификации текста с неизвестным автором достаточно проверить, в какую часть пространства он попал.

Методы классификации с помощью SVM значительно превосходят кластерную и наивную Байесовскую классификацию [15]. Кластерная классификация характеризуется тем, что документ считают принадлежащим к ближайшему множеству, и в конечном итоге все зависит от определения расстояния до множества. Имеется большое количество вариантов этого метода: средневзвешенный, k ближайших соседей и др. Байесовская классификация предполагает, что частоты слов в тексте являются независимыми случайными величинами.

В качестве характерных признаков текста для описания авторского стиля предлагается брать наиболее частые триграммы символов и наиболее частые слова русского языка.

В качестве инструментов для атрибуции текстов в работе [1] были выбраны искусственные нейронные сети архитектуры многослойный перцептрон (MLP), сети каскадной корреляции (CCN) и аппарат машины опорных векторов (SVM). CCN позволяют снизить временные затраты на обучение по сравнению с перцептроном за счет алгоритма автоматического построения топологии сети. SVM является наиболее точным из существующих в настоящее время методов классификации и в то же время наименее затратным по времени. Итоговое решение об авторе текста принимается ансамблем классификаторов по принципу мажоритарного голосования.

Основные результаты проведенных исследований были получены на корпусе, состоящем из 215 прозаических текстов 50 русских писателей. Тексты взяты из электронной библиотеки М. Мошкова. Размер каждого текста составлял более 100000 символов. Использовались выборки объемом 1000–100000 символов (200–20000 слов). Количество обучающих примеров каждого автора бралось равным 3, для тестирования использовалось по 1 выборке автора.

Эксперименты для случая 2, 5 и 10 авторов показали, что наиболее информативными авторским признаками являются ограничения в 300–700 наиболее частотных триграмм и 500 наиболее частых слов. Автора можно определить с точностью в среднем 95%–98% при объеме текстовой выборки 20000–25000 символов. При этом начиная с 10000 символов, машина опорных векторов показывает лучшие из трех исследуемых классификаторов результаты. Установлено, что использование при идентификации автора комбинации частот букв русского языка, знаков пунктуации, наиболее частых триграмм символов и наиболее частых слов увеличивает точность идентификации в среднем на 6%–12% на объемах текста до 10000 символов.

Полученные методики были применены на практике для идентификации авторов коротких электронных сообщений во время внедрения разработанного метода и программного комплекса, названного «Авторовед», в деятельность воинской части 51952. Результаты показали, что авторство коротких текстов длиной 100 символов можно определить с точностью до 76 ± 11% в случае двух потенциальных авторов. При решении

DOI: 10.7256/2305-6061.2014.2.11705

частной задачи по определению автора сообщения интернет-форума была достигнута точность 89 ± 8%. Таким образом, предложенный метод дает довольно хорошие результаты на коротких электронных сообщениях, что выгодно отличает его от других ранее предложенных методов.

6. Определение эмоциональной оценки мнения автора как подзадача определения автора текста.

Задача определения тональности текста в некотором смысле похожа на задачу атрибуции. Вторая является более обширной. При автоматизации этих задач иногда применяются похожие подходы, в частности, методы машинного обучения или методы классификации. Существует довольно много работ, например, [16], [17], [18], посвященных автоматическому определению эмоциональной составляющей в тексте. В большинстве из них рассматривается упрощенное эмотивное пространство, например, выделяется два класса оценок: позитивная и негативная. Но бывают случаи, когда эмоциональную окраску текста не всегда можно определить однозначно. Тогда применяют многополосное шкалирование, т.е. словам, связанным с тональностями, ставятся в соответствии числа по шкале, например, от -5 до 5 (от самого отрицательного к самому положительному).

Помимо оценки существует множество других типов высказываний, например, благодарность, извинение, поздравление и прочие. Это связано с тем, что в процессе говорения (по-латински in locutio) человек одновременно совершает еще и некоторое действие, имеющее какую-то внеязыковую цель: он благодарит, поздравляет, просит, приказывает и т.п. Такие действия Дж. Остин [19], [20] назвал *иллокутивными*. В иллокутивном речевом действии выделяют две основных составляющих: иллокутивную функцию (F) и пропозицию (P) и обобщенно представляют его в виде формулы $F(P)$. Например, смысл высказывания «*Я обещаю сесть за уроки*» в типичной ситуации его употребления состоит из пропозиции «*я сяду за уроки*» и иллокутивной функции «обещание»; смысл высказывания «*Он обещает сесть за уроки*» в типичной ситуации его употребления складывается из пропозиции «*он обещает сесть за уроки*» и иллокутивной функции «сообщение».

В различных языках существуют специальные формальные средства, прямо или косвенно указывающие на иллокутивную функцию речевого действия. Признаками иллокутивной функции являются: контекст, наклонение глагола, множество перформативных глаголов (*я прошу / обещаю / советую* и т.п.), порядок слов, пунктуация (на письме), ударение (в устной речи), интонационный контур (в устной речи). Пунктуация и порядок слов мало могут помочь, т.к. вопросительный и восклицательный знаки – слишком скудные средства [19], а порядок слов в русском языке относительно свободный. При применении данной теории для компьютерной обработки, устная речь обычно не рассматривается. Часто в реальных речевых ситуациях иллокутивную функцию высказывания проясняет контекст, и формализовать соответствующий показатель функции затруднительно [21]. Поэтому из перечисленных признаков остается учитывать наклонение глаголов, множество перформативных глаголов и контекст.

DOI: 10.7256/2305-6061.2014.2.11705

6.1. Перформативные глаголы.

На семантическом уровне перформативное предложение отличается от обычного повествовательного тем, что последнее используется с целью представления некоторого положения дел, т.е. с целью описания, сообщения, утверждения и т.п., а перформативное предложение служит не для описания действия, которое совершает говорящий, а для пояснения того, какое именно действие он совершает. Например, в обычном повествовательном предложении «*Я рисую вас*» описывается некоторая ситуация, существующая независимо от речевого акта, а в перформативном предложеним «*Я приветствую вас*» при нормальном для него употреблении говорится о самом речевом акте его употребления. Обычное повествовательное предложение, будучи употребленным, становится высказыванием, которое можно оценить как истинное или ложное, тогда как к перформативным предложениям в типовом контексте их употребления этот вид оценки не может быть применен. В нормальном случае употребления такого предложения вопрос об истинности или ложности слов говорящего не встает. Соответствующее высказывание может оцениваться только как уместное или неуместное, но не как истинное или ложное.

Классическая форма перформативного предложения имеет следующий вид.

Подлежащее, выраженное личным местоимением 1-ого лица единственного числа, + согласованное с ним сказуемое в форме изъявительного наклонения настоящего времени активного (действительного) залога.

Например, «*(Я) обещаю вам исправиться*».

Для русского языка можно добавить следующие случаи [20]:

1. лицо может быть не только первым, но и третьим, например, в тексте официального послания глагол в третьем лице *благодарят* употреблен перформативно: *Президент и посол Республики Мали благодарят сотрудников аэропорта...* (сообщение);
2. число может быть множественным;
3. время может быть будущим: *Напомню вам, что завтра заканчивается срок паспорта* (напоминание);
4. залог может быть пассивным (страдательным): *Вы назначаетесь моим заместителем* (назначение);
5. наклонение может быть сослагательным: *Я посоветовал бы вам уйти* (совет).

Кроме того, для перформативного употребления глагола не обязательно даже, чтобы он был синтаксической вершиной предложения, например: *Хотелось бы поблагодарить присутствующих за теплые слова* (благодарность); *Спешу поздравить вас с рождением дочери* (поздравление) и т.п.

6.2. Контекст.

Контекст можно учитывать при помощи тематических тезаурусов, составленных для каждой иллокутивной функции. Этим функциям будут соответствовать предикаты. На данный момент нами было выделено 14 предикатов: аргументация (объяснение), благо-

DOI: 10.7256/2305-6061.2014.2.11705

дарность, вопрос, извинение, инструкция, мнение, напоминание, обещание, поздравление, приглашение, приказ, просьба, рекомендация (совет), сообщение (повествование).

Например, $\mathrm{Explain}(w_1,\ldots,w_n,t)$ – предикат истинен на предложении (или тексте) t, если $w_1,\ldots,w_n$ – набор слов, входящих в высказывание (или текст), иллокутивная функция которого является аргументацией (объяснением). Тезаурус включает в себя слова: *так как, потому что, ввиду того что, из-за,* поэтому и некоторые другие.

$\mathrm{Opinion}(w_1,\ldots,w_n,t)$ – предикат истинен на t, если $w_1,\ldots,w_n$ – набор слов высказывания, иллокутивная функция которого является мнением. Тезаурус будет содержать: *на мой взгляд, на наш взгляд, по-моему, по-нашему, думаю, думаем, считаю, считаем* и пр.

Позже можно будет расширить набор предикатов и рассматривать, в том числе редко встречающиеся иллокутивные функции. Предложенный метод позволяет учитывать смысловую нагрузку высказывания, а значит, получать более тонкую классификацию текстов. При разделении на классы подобным способом становится возможным учитывать эмоциональную составляющую. Высказанные эмоции автора, его оценка происходящих событий косвенным образом отражают его восприятие и жизненный опыт. Эта информация является ценной и может применяться для выявления особенностей лексики автора в контексте различных ситуаций.

7. Заключение.

В данной статье рассмотрены формальные методы атрибуции, т.е. определения авторов текстов. Большинство из них используют определенные параметры в качестве признаков авторского стиля. Довольно хорошие результаты при решении задачи атрибуции демонстрируют статистические методы, алгоритмы сжатия данных и нейронные сети.

В основе формальных методов атрибуции текстов лежит представление о том, что с возрастанием объема текста параметры, характеризующие авторский стиль, становятся устойчивыми с вероятностной точки зрения, что позволяет устанавливать авторство по стабильно повторяющимся формальным характеристикам текста. Поэтому более высокое качество атрибуции достигается для текстов большого объема, и менее точный результат получается для текстов маленького объема.

Открытым остается вопрос о выборе авторского инварианта (набора формальных параметров текста). Часто на практике решается ограниченный круг задач для предварительно заданного набора текстов. Настройка, тестирование и демонстрация инструментов анализа ориентирована только на эти тексты, и нет никакой гарантии, что методы будут эффективно справляться с задачей на других данных. Иными словами, для построения универсального и независящего от текстов авторского инварианта необходимо искать новые пути формирования характеристик.

Установив набор характеристик, исследователь сталкивается с проблемой их структуризации, в чем существенную помощь могут оказать классические статистические методы. С помощью факторного анализа и анализа главных компонент можно установить вклад той или иной характеристики в процесс распознавания автора, иерархический кластер-

DOI: 10.7256/2305-6061.2014.2.11705

ный анализ позволяет осуществлять объединение отдельных характеристик в подгруппы, подгрупп в группы и так далее. Немалую помощь можно получить от нейронных сетей прямого распространения, если попытаться обучить сеть на наборе примеров, взяв в качестве входов отдельные характеристики, а затем оценивать, какое влияние оказывает тот или иной вход на систему выходов.

Недостаточно исследованы зависимости качества классификации различными методами от объемов фрагментов и от числа классов. Наконец, имеющиеся программы анализа текстов не ориентированы на комплексное исследование и сравнение стилей текстов (для разных задач анализа стилей текстов с использованием различных методов их решения, различных частотных признаков, различного текстового материала и т.д.). Тем не менее, на сегодняшний день лучшие результаты показывают программы «Стилеанализатор» и «Авторовед». На русскоязычных текстах получена точность 90–98%. Наиболее удачное сравнение доступных программных средств для идентификации авторства текстов можно найти в [1] и приведено в таблице 1.

Название	Методы	Измен. параметров метода	Средства анализа текстов	Расширен. перечня характеристик	Необходимый объем текста	Точность, %	Примен. к решению реальных задач
«Лингвоанализатор»	Энтропийный подход, марковские цепи	Нет	Графем., стат. анализ	Нет	40000-100000 символов	84–89	Нет
«Атрибутор»	Марковские цепи	Нет	Стат. анализ	Нет	>20000 символов	Не изв.	Нет
«СМАЛТ»	Критерии Стьюдента, Колмогорова-Смирнова, кластерный анализ	Нет	Графем., морф., синт., стат. анализ, поддержка дореволюционной орфографии	Нет	500 слов для определения однородности	Не изв.	Да
«Стилеанализатор»	Марковские цепи, нейронные сети, деревья решений, меры расстояния	Да	Графем., стат. анализ, работа с размеченными текстами	Да	30000-40000 символов	90–98	Да
«Авторовед»	Нейронные сети, метод опорных векторов, QSUM	Да	Графем., морф., стат. анализ	Да	20000-25000 символов	95–98	Да
					100 символов	76	

Таблица 1. Сравнение программных средств атрибуции текстов

К проблемам, затрудняющим исследования в области атрибуции текстов, относится также проблема составления выборки эталонных текстов. Желательно, чтобы произведения были подобраны следующим образом: тексты разных писателей в максимальной степени различались друг от друга, а тексты одного писателя были максимально близки. Но существует немало случаев, когда известный писатель в какой-то период своего творчества менял стиль изложения, или произведения были написаны в соавторстве. Эти факты создают дополнительные сложности при решении задачи установления авторства.

Необходимо проводить дальнейшие исследования, направленные на поиск новых или совершенствование уже имеющихся методов атрибуции текстов, а также на про-

DOI: 10.7256/2305-6061.2014.2.11705

ведение экспериментов, целью которых является поиск характеристик, позволяющих четко разделять стили авторов, в том числе и на малых объемах выборки. Представляет также интерес создание новых методов автоматического определения эмоциональной тональности текстов. Следует заметить, что разбиение на классы должно быть более детальным, нежели при одномерном эмотивном пространстве. Тогда тип высказывания, оценка автора происходящему могут служить косвенным признаком авторского стиля и, как следствие, принадлежности текста конкретному автору.

Библиография :

1. Романов А.С. Методика и программный комплекс для идентификации автора неизвестного текста: Автореф. дис. канд. тех. наук. Томск, 2010. 26 с.
2. Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов. Л.: ЛГУ, 1990. 164 с.
3. Родионова Е.С. Методы атрибуции художественных текстов // Структурная и прикладная лингвистика: Межвузовский сборник. СПб.: СПбГУ, 2008. Вып. 7. С. 118–127.
4. Марков А.А. Об одном применении статистического метода // Известия Императорской Академии наук. Сер. 6. 1916. Т. 10, № 4. С. 239–242.
5. Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов // Новая хронология Греции: Античность в Средневековье. М.: МГУ, 1995. 422 с.
6. Хмелёв Д.В. Классификация и разметка текстов с использованием методов сжатия данных // Всё о сжатии данных, изображений и видео. 2003. URL: http://compression.ru/download/articles/classif/intro.html (дата обращения: 17.04.2014)
7. Хмелёв Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вестник МГУ. Сер. 9: Филология. 2000. №2. С. 115–126.
8. Кукушкина О.В., Поликарпов А.А, Хмелёв Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. М.: Наука, 2001. Т. 37. № 2. С. 96–108.
9. Шевелёв О.Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений: Автореф. дис. канд. тех. наук. Томск, 2006. 18 с.
10. Тимашев А.Н. Атрибутор // Текстология. ru. 1999–2007. URL: http://www.textology.ru/atr_resum.html (дата обращения: 17.04.2014)
11. Информационная система «Статистические методы анализа литературного текста». 2004. URL: http://smalt.karelia.ru (дата обращения: 16.04.2014) .
12. Рогов А.А., Сидоров Ю.В., Король А.В. Автоматизированная система обработки и анализа литературных текстов СМАЛТ // Труды и материалы II-го Международного конгресса исследователей русского языка «Русский язык: исторические судьбы и современность». М: МГУ, 2004. С. 485–486.
13. Антиплагиат. 2005–2014. URL: http://www.antiplagiat.ru (дата обращения: 16.04.2014)

DOI: 10.7256/2305-6061.2014.2.11705

14. Шевелёв О.Г. Методы автоматической классификации текстов на естественном языке: Учебное пособие. Томск: ТМЛ-Пресс, 2007. 144 с.
15. Романов А.С., Мещеряков Р.В. Идентификация автора текста с помощью аппарата опорных векторов / А.С. Романов, Р.В. Мещеряков // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». М.: РГГУ, 2009. Вып. 8, №15. С. 432–437.
16. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. Vol. 2, No 1-2. 2008. P. 1–135.
17. Пазельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: сб. научных статей. М.: Изд-во РГГУ, 2011. Вып. 10, №17. С. 510–522.
18. Yi J., Nasukawa T., Bunescu R., Niblack W. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques // Proc. of the Third IEEE International Conference on Data Mining (ICDM 2003), 2003. P. 427–434.
19. Остин Дж. Слово как действие // Новое в зарубежной лингвистике. М.: Прогресс, 1986. Вып. 17. С. 22–130.
20. Онлайн энциклопедия «Кругосвет». 1997–2014. URL: http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/RECHEVO_AKT.html (дата обращения: 15.04.2014)
21. Серль Дж. Что такое речевой акт? // Новое в зарубежной лингвистике. М., 1986. Вып. 17. С. 151–169.

References:

1. Romanov A.S. Metodika i programmnyi kompleks dlya identifikatsii avtora neizvestnogo teksta: Avtoref. dis. kand. tekh. nauk. Tomsk, 2010. 26 s.
2. Marusenko M.A. Atributsiya anonimnykh i psevdonimnykh literaturnykh proizvedenii metodami teorii raspoznavaniya obrazov. L.: LGU, 1990. 164 s.
3. Rodionova E.S. Metody atributsii khudozhestvennykh tekstov // Strukturnaya i prikladnaya lingvistika: Mezhvuzovskii sbornik. SPb.: SPbGU, 2008. Vyp. 7. S. 118–127.
4. Markov A.A. Ob odnom primenenii statisticheskogo metoda // Izvestiya Imperatorskoi Akademii nauk. Ser. 6. 1916. T. 10, № 4. S. 239–242.
5. Fomenko V.P., Fomenko T.G. Avtorskii invariant russkikh literaturnykh tekstov // Novaya khronologiya Gretsii: Antichnost' v Crednevekov'e. M.: MGU, 1995. 422 s.
6. Khmelev D.V. Klassifikatsiya i razmetka tekstov s ispol'zovaniem metodov szhatiya dannykh // Vse o szhatii dannykh, izobrazhenii i video. 2003. URL: http://compression.ru/download/articles/classif/intro.html (data obrashcheniya: 17.04.2014)
7. Khmelev D.V. Raspoznavanie avtora teksta s ispol'zovaniem tsepei A.A. Markova // Vestnik MGU. Ser. 9: Filologiya. 2000. №2. S. 115–126.
8. Kukushkina O.V., Polikarpov A.A, Khmelev D.V. Opredelenie avtorstva teksta s ispol'zovaniem bukvennoi i grammaticheskoi informatsii // Problemy peredachi informatsii. M.: Nauka, 2001. T. 37. № 2. S. 96–108.
9. Shevelev O.G. Razrabotka i issledovanie algoritmov sravneniya stilei tekstovykh proizvedenii: Avtoref. dis. kand. tekh. nauk. Tomsk, 2006. 18 s.

DOI: 10.7256/2305-6061.2014.2.11705

10. Timashev A.N. Atributor // Tekstologiya. ru. 1999–2007. URL: http://www.textology.ru/atr_resum.html (data obrashcheniya: 17.04.2014)
11. Informatsionnaya sistema «Statisticheskie metody analiza literaturnogo teksta». 2004. URL: http://smalt.karelia.ru (data obrashcheniya: 16.04.2014)
12. Rogov A.A., Sidorov Yu.V., Korol' A.V. Avtomatizirovannaya sistema obrabotki i analiza literaturnykh tekstov SMALT // Trudy i materialy II-go Mezhdunarodnogo kongressa issledovatelei russkogo yazyka «Russkii yazyk: istoricheskie sud'by i sovremennost'». M: MGU, 2004. S. 485–486.
13. Antiplagiat. 2005–2014. URL: http://www.antiplagiat.ru (data obrashcheniya: 16.04.2014) 14.
14. Shevelev O.G. Metody avtomaticheskoi klassifikatsii tekstov na estestvennom yazyke: Uchebnoe posobie. Tomsk: TML-Press, 2007. 144 s.
15. Romanov A.S., Meshcheryakov R.V. Identifikatsiya avtora teksta s pomoshch'yu apparata opornykh vektorov / A.S. Romanov, R.V. Meshcheryakov // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog 2009». M.: RGGU, 2009. Vyp. 8, №15. S. 432–437.
16. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. Vol. 2, No 1-2. 2008. P. 1–135.
17. Pazel'skaya A.G., Solov'ev A.N. Metod opredeleniya emotsii v tekstakh na russkom yazyke // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: cb. nauchnykh statei. M.: Izd-vo RGGU, 2011. Vyp. 10, №17. S. 510–522.
18. Yi J., Nasukawa T., Bunescu R., Niblack W. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques // Proc. of the Third IEEE International Conference on Data Mining (ICDM 2003), 2003. P. 427–434.
19. Ostin Dzh. Slovo kak deistvie // Novoe v zarubezhnoi lingvistike. M.: Progress, 1986. Vyp. 17. S. 22–130.
20. Onlain entsiklopediya «Krugosvet». 1997–2014. URL: http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/RECHEVO_AKT.html (data obrashcheniya: 15.04.2014)
21. Serl' Dzh. Chto takoe rechevoi akt? // Novoe v zarubezhnoi lingvistike. M., 1986. Vyp. 17. S. 151–169.

DOI: 10.7256/2305-6061.2014.2.11705