УДК 004.072.7

А.О. Исхакова

Модель процесса формирования инвариантов классов текстов

Предложена модель процесса формирования инвариантов классов текстов, основанная на использовании качественных и уточняющих их количественных характеристик текста, при этом формирование перечня характеристик основывается на лингвистических особенностях исследуемых классов текстов. Модель применялась на примере формирования инвариантов для двух классов: естественных и искусственных текстов. Результатом является перечень характеристик, обладающих различительной способностью при классификации таких текстов, а также инварианты указанных классов. Представленные данные могут быть использованы для автоматического определения происхождения текста на основе обученной нейронной сети или оценке статистических параметров указанных характеристик текстов.

Ключевые слова: текст, инвариант класса текстов, характеристика текста, модель, искусственный текст, естественный текст.
doi: 10.21293/1818-0442-2016-19-3-76-80

Для современного общества сеть Интернет – это основа многих социальных и деловых взаимодействий, уникальная совокупность локальных, региональных, национальных и общемировых компьютерных сетей. Важнейшим символом глобальной сети является технология мгновенного обмена данными между миллионами пользователей.

Объем распространяемой в Интернете информации увеличивается с каждым годом, вместе с ним усиливается конкурентная борьба за внимание пользователя. Для повышения позиций ресурса в поисковых системах и, как следствие, популярности вебсайтов используются различные методы оптимизаций. Часто для этого используются методы автоматического порождения текстов, которые позволяют создавать множество уникальных версий некоторого исходного экземпляра [1]. Уникальность достигается благодаря использованию специальных алгоритмов, связанных с изменением текстообразующих элементов.

Массово порожденные тексты используются:

– для привлечения читателя к веб-ресурсу (для этого генерируются «копии» текстового контента, освещающего популярные темы);

– для распространения в сети контента определенной направленности (для этого «копии» текста распространяются по различным информационным ресурсам).

Такие интернет-ресурсы как социальные сети или информационные порталы являются для многих пользователей основным источником сведений о событиях, изменениях в мире, аналогом средств массовой информации. При таком положении массовое порождение текстов можно рассматривать как инструмент для формирования общественного мнения через интернет-СМИ и социальные сети различного уровня. Такой инструмент может быть использован для пропаганды определенных идей, в том числе преступных, а также введения в заблуждение населения или парализации работы электронных ресурсов [2].

Множество исследовательских работ посвящено выявлению отличительных свойств искусственно созданных текстов, представляющих собой поисковый спам и направленных на обман алгоритмов работы поисковых систем [3–5]. Данные тексты имеют ряд особенностей: обилие ключевых слов, определенным образом выстроенные ссылки, наличие скрытого текста и др. [6]. Поисковый спам практически не несет смысловой нагрузки и предназначен для манипулирования работой поисковых алгоритмов.

Автоматические генераторы также используются для создания контента, предназначенного для прочтения пользователем. Такой класс текстов изучен в меньшей степени. В частности на сегодняшний день не выделены отличительные свойства, которые бы позволили определять их происхождение. В связи с этим задача исследования характеристик автоматически сгенерированных текстов, которые представляют собой информационный контент, является актуальной.

Модель процесса формирования инвариантов классов текстов

В рамках решаемой задачи выделим 2 класса текстов: естественные и искусственные. Под первыми понимаются тексты, созданные человеком, под вторыми – созданные автоматически с помощью специального программного алгоритма. Для отнесения входного текста к одному из указанных классов необходимо сформировать соответствующие инварианты.

В классической задаче атрибуции – установлении авторства – инвариант, на основе которого идентифицируется автор, представляет собой набор значений характеристик текста определенного лица [7]. Для создания такого набора существует несколько подходов. В случае с идентификацией искусственных текстовых произведений инвариантом является набор значений характеристик текста, с помощью которых может быть установлена причастность данного генератора к происхождению входного текста [8].

Многими учеными предпринимались попытки смоделировать подход к формированию набора характеристик текста, составляющих инвариант. В ранних работах, посвященных обработке и классификации текстовых произведений, в основе выбора

характеристик лежал либо интуитивный подход, либо случайный перебор [9]. Исследователями в области обработки текста сформированы наборы характеристик, которые чаще всего используются для расчета инварианта. На использовании таких наборов основываются модели создания инвариантов текстов при решении задач классификации.

В работе А.С. Романова [10], посвященной определению авторства текстов, приводится методика, в которую включена модель процесса создания инвариантов. Процесс описан следующим образом: на вход подаются доступные признаки текста, которые пользователь формирует в некоторую группу признаков текстов, данная группа используется для формирования модели авторского стиля, то есть инварианта. Модель процесса основана на использовании известных наборов характеристик вне зависимости от особенностей решаемой прикладной задачи. При таком подходе значительно возрастает вычислительная сложность расчетов, так как количество всевозможных характеристик может составлять несколько тысяч, а также приводит к риску упущения каких-либо характеристик текста, которые отсутствуют в стандартных наборах, но в конкретном случае могут обладать различительной способностью. Особенность формирования инвариантов должна состоять в том, что наборы исследуемых характеристик зависят, в первую очередь, от непосредственно решаемой задачи, а процесс формирования набора отталкивается от задачи классификации.

Автором была предложена модель процесса формирования инвариантов классов текстов (рис. 1), основанная на использовании качественных и уточняющих их количественных характеристик текста. Таким образом, при формировании набора исследователь основывается на лингвистических особенностях рассматриваемых классов текстов.

На вход модели подаются классы текстов и наборы текстов данных классов. На 1-м этапе с учетом лингвистических особенностей языка формируется перечень характерных для исследуемых классов качественных признаков. Ими могут выступать эмоциональная окраска, соответствие стилю и времени, связность текста и т.п. Далее пошагово происходит уточнение выделенных признаков до формирования набора количественных характеристик: уточняются текстовые свойства, которые определяют проявление выделенных качественных признаков. Затем для каждого свойства формируется набор текстовых характеристик, позволяющих оценить проявление этих свойств в тексте. На дальнейших шагах происходит проверка выделенных характеристик на различительную способность и взаимную зависимость.

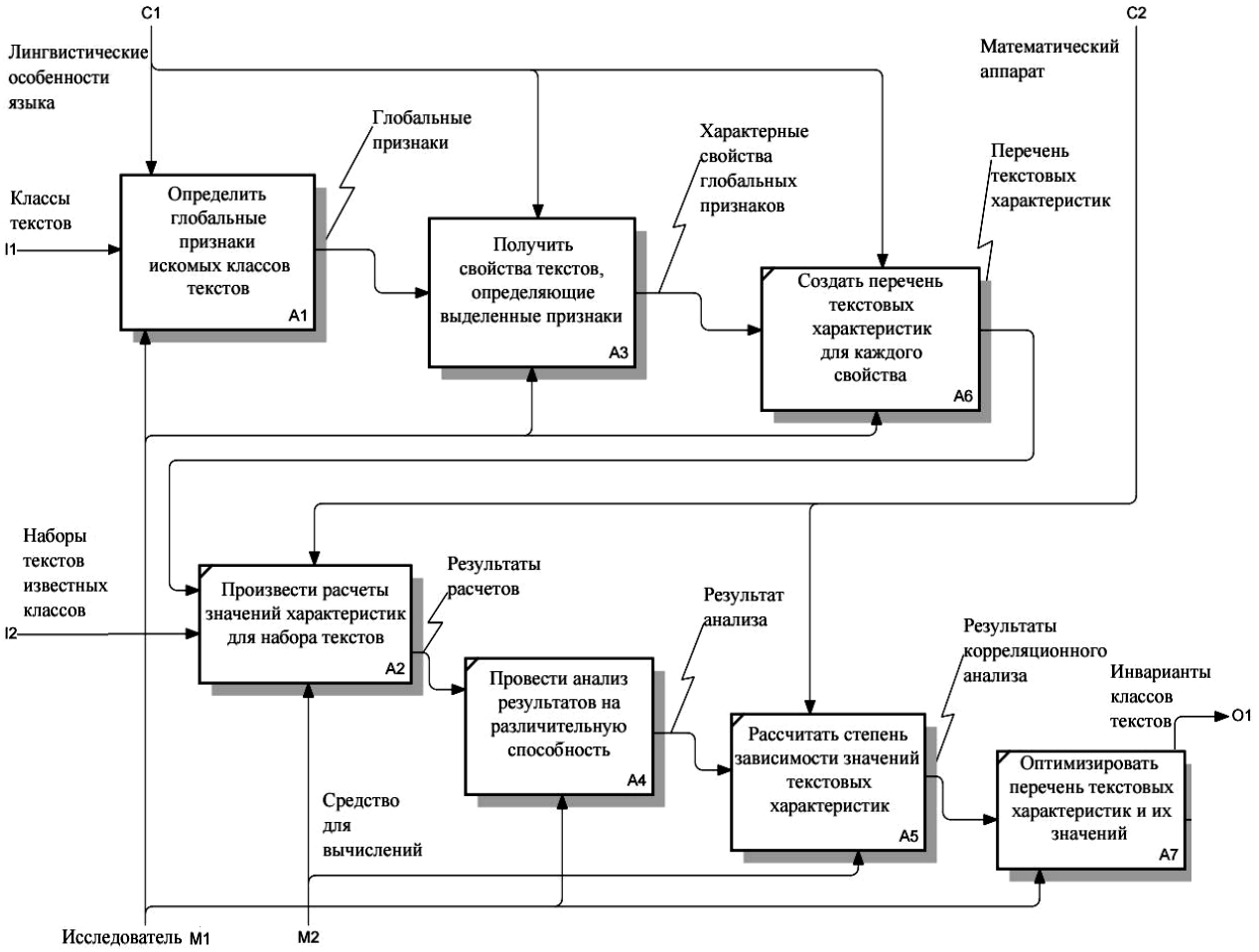

Рис. 1. Модель процесса формирования инвариантов классов текстов

Формирование инварианта искусственных и естественных текстов

Входные данные. Для формирования инвариантов естественных и искусственных текстов были исследованы 3210 текстов и созданные на их основе автоматические сгенерированные экземпляры. Данные тексты представляли собой публицистические статьи информационного характера длиной от 1000 до 5700 символов. В качестве генератора был использован синонимизатор со словарем, содержащим синонимы к 700 000 словам. В общей сложности при оценке численных значений характеристик текстов двух классов были использованы 3210 естественных текстов общим объемом 10,7 млн символов и 3210 искусственных текстов (12,3 млн символов). Указанные объемы считаются достаточными для обучения в соответствии с опытом формирования инвариантов при анализе текстов различных авторов, приведенных в [11].

Формирование перечня исследуемых характеристик текстов

Согласно предложенной модели на основе лингвистических особенностей языка были определены качественные (глобальные) характеристики, определяющие важнейшее различие между указанными классами текстов. Отличием естественных текстов от искусственных является их связность в рамках межфразовых единств, а также цельность, то есть наличие глобальной связи компонентов текста на содержательном уровне [12, 13]. Таким образом, связность и цельность являются непременными лингвистическими признаками текста, которые проявляются в целесообразно построенном человеком тексте и отличают его от массово порожденных экземпляров.

Далее на основе работ по теории лингвистики [12] были определены свойства текста, обеспечивающие его связность и цельность:

– символьные (связанные с наличием символов и их сочетаний);

– лексические (связанные с наличием слов и словосочетаний);

– синтаксические (связанные с конструкциями предложений);

– семантические (связанные с оценкой мер семантического сходства и связанности);

– тематические (связанные с соответствием используемых средств тематике текста).

Также сформирован перечень количественных характеристик для каждого из свойств:

– средняя длина слов;

– среднее количество знаков пунктуации в предложении;

– частота 100 популярных биграмм букв;

– частота служебных слов;

– частота неопределенных местоимений;

– частота коротких слов (менее 4 символов);

– частота длинных слов (более 7 символов);

– количество уникальных слов;

– среднее число слов в предложении;

– количество грамматических ошибок;

– количество предложений в тексте;

– количество сложноподчиненных предложений;

– доля сложноподчиненных предложений;

– количество вопросительных предложений;

– количество восклицательных предложений;

– доля эмотизированных предложений;

– частота 100 популярных слов;

– частота 100 популярных 2-грамм слов;

– частота 100 популярных 3-грамм слов;

– количество слов в семантическом ядре;

– наличие единства тематики в разных частях текста;

– наличие единства жанра в разных частях текста.

Анализ результатов на различительную способность характеристик текста

По предложенной модели на следующем шаге необходимо оценить различительную способность выделенных характеристик, прежде произведя расчеты значений для текстовых выборок. Условием различительной способности текстовой характеристики по [14] была выбрана мера, определяющая превосходство разности математических ожиданий для двух классов над суммой их среднеквадратических отклонений:

$$|M_1 - M_2| > \sigma_1 + \sigma_2,$$

где M_1, M_2 – математические ожидания величины значения текстовой характеристики для двух выборок текстов; σ_1, σ_2 – среднеквадратические отклонения величины значения текстовой характеристики для двух выборок текстов (индексы совпадают).

В соответствии с проведенными вычислениями был сделан вывод, что для исследуемых классов текстов различительной способностью не обладают следующие характеристики, которые были удалены из набора:

– средняя длина слова;

– частота длинных слов;

– доля сложноподчиненных предложений;

– доля восклицательных и вопросительных предложений.

Оценка корреляции значений

Оценка корреляции рассчитанных значений позволяет выделить закономерно изменяющиеся характеристики внутри одной выборки. Пары, имеющие сильную корреляционную зависимость, должны быть разбиты, одна из характеристик удалена из набора. Это позволит снизить вычислительные затраты для расчета значений и классификации, а также увеличить различительную способность инварианта в целом.

Оценка корреляции значений текстовых характеристик внутри каждой выборки осуществлялась с помощью метода Пирсона (метод квадратов):

$$r_{xy} = \frac{\sum_{i=1}^{k}(d_{x_i} \cdot d_{y_i})}{\sqrt{\sum_{i=1}^{k} d_{x_i}^2 \cdot \sum_{i=1}^{k} d_{y_i}^2}},$$

где d_{xi}, d_{yi} – отклонение i-го числового значения от среднего значения своего вариационного ряда; k – количество элементов вариационных рядов (количество текстов в наборе).

По итогам расчета коэффициента корреляции была обнаружена сильная корреляция ($|r_{xy}| \geq 0,7$) у ряда пар характеристик. Из набора были исключены следующие из них:

– частота неопределенных местоимений;
– частота 100 популярных 3-грамм слов;
– количество коротких слов;
– количество восклицательных предложений.

Инварианты классов текстов

В результате были сформированы два инварианта: для естественных текстов, то есть созданных человеком, и для искусственных – созданных с помощью синонимизации. Инвариант a_i представляет собой вектор значений характеристик. Размерность таких векторов соответствует количеству отобранных характеристик текста:

$$a_i = (a_{i1}, a_{i2}, \ldots, a_{im}),$$

где a_{ij} – усредненное информативное значение j-й текстовой характеристики i-го инварианта, $i = 1...n$, $j = 1…m$; n – количество инвариантов (соответствует количеству классов текстов); m – количество используемых характеристик текста в инварианте.

Характеристики текста, составившие инварианты текстов, разделенных на классы по своему происхождению:

– среднее количество знаков пунктуации в предложении;
– частота 100 популярных биграмм букв;
– частота служебных слов;
– количество уникальных слов;
– среднее число слов в предложении;
– количество грамматических ошибок;
– количество предложений;
– количество сложноподчиненных предложений;
– количество вопросительных предложений;
– частота 100 популярных слов;
– частота 100 популярных 2-грамм слов;
– количество слов в семантическом ядре;
– наличие единства тематики в разных частях текста.

Ниже приведены полученные векторы численных значений характеристик текста (a_1 – инвариант класса естественных текстов; a_2 – инвариант класса искусственных текстов, сгенерированных с помощью синонимизации):

a_1 = (31,742; 201,269; 34,691; 64,804; 9,113; 0,01; 109,812; 68,655; 1,414; 49,001; 9,1; 66,025; 1,7);
a_2 = (29,035; 112,562; 25,702; 101,659; 9,987; 6,215; 100,2; 62,082; 1,358; 32,882; 3,554; 95,645;0,6).

Таким образом, с помощью предложенной модели был получен набор характеристик текстов, обладающих различительной способностью в решении задачи идентификации происхождения текста, а именно – определения, написан ли текст человеком или создан автоматически с помощью программного генератора. На основе проведенных расчетов средних были сформированы инварианты исследуемых классов текстов.

Заключение

Предложенная модель процесса формирования инвариантов классов текстов была применена для создания инвариантов естественных и искусственных текстов. Данная модель основывается на классических вариантах представления этого процесса, однако в выборе характеристик текста предлагается основываться на лингвистических особенностях текста, что позволяет поэтапно сформировать перечень количественных характеристик. Такой подход в моделировании процесса позволяет снизить вычислительные затраты на проведение расчетов, а также выделить все необходимые характеристики, в том числе если их нет в стандартных наборах.

Задачи, связанные с атрибуцией текста, носят междисциплинарный характер, поэтому исследования в области лингвистики при создании набора характеристик для инварианта являются основополагающими. Учитывая данный факт, можно заключить, что предложенная модель универсальна и может быть использована в решении любой задачи, связанной с классификацией текстовых произведений.

Литература

1. SEO-копирайтинг: как приручить поисковик [Электронный ресурс]. – Режим доступа: http://onedesign.pro/upload/books/11_Kak_priruchit.pdf, свободный (дата обращения: 06.08.2016).

2. Управление ООН по наркотикам и преступности. Использование Интернета в террористических целях [Электронный ресурс]. – Режим доступа: https://www.unodc.org/documents/terrorism/Publications/Use_of_Internet_for_Terrorist_Purposes/Use_of_the_internet_for_terrorist_purposes_Russian.pdf, свободный (дата обращения: 24.08.2016).

3. Павлов А.С. Методы обнаружения поискового спама, порожденного с помощью цепей Маркова / А.С. Павлов, Б.В. Добров // Тр. XI Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции. – Петрозаводск: Изд-во Карельского научного центра РАН, 2009. – Т. 1. – С. 311–317.

4. Поиск неестественных текстов / Е.А. Гречников, Г.Г. Гусев, А.А. Кустарев, А.М. Райгородский // Труды XI Всерос. конф. «Цифровые библиотеки: продвинутые методы и технологии, цифровые коллекции» – RCDL'2009, Петрозаводск. – Петрозаводск: Изд-во Карельского научного центра РАН, 2009. – С. 306–308.

5. A reference collection for web spam / C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, S. Vigna // ACM Sigir Forum 2006. – 2006. – Vol. 40, Issue 2. – P. 11–24.

6. Романов А.С. Обобщенная методика идентификации автора неизвестного текста / А.С. Романов, А.А. Шелупанов, С.С. Бондарчук // Доклады ТУСУРа. – 2010. – № 1(21), ч. 1. – С. 108–112.

7. Зайцева А.А. Метод оценки качества текстов в задачах аналитического мониторинга информационных ресурсов / А.А. Зайцева, С.В. Кулешов, С.Н. Михайлов // Труды СПИИРАН. – 2014. – Вып. 37. – С. 144–155.

8. Шумская А.О. Выбор параметров для идентификации искусственно созданных текстов // Доклады ТУСУРа. – 2013. – № 2(28). – С. 126–128.

9. Батура Т.В. Формальные методы определения авторства текстов // Вестник НГУ. Сер.: Информационные технологии. – 2012. – Т. 10, вып. 4. – С. 81–94.

10. Романов А.С. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста / А.С. Романов, А.А. Шелупанов, Р.В. Мещеряков. – Томск: В-Спектр, 2011. – 188 с.

11. Романов А.С. Методика идентификации автора текста на основе аппарата опорных векторов // Доклады Том. гос. ун-та систем управления и радиоэлектроники. – 2009. – № 1(19), ч. 2. – С. 36–42.

12. Валгина Н.С. Теория текста. – М. : Логос, 2003. – 191 с.

13. Николина Н.А. Филологический анализ текста: учеб. пособие. – М.: Изд. центр «Академия», 2003. – 256 с.

14. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999. – 270 с.

Исхакова Анастасия Олеговна
Аспирантка каф. комплексной информационной безопасности электронно-вычислительных систем ТУСУРа
Тел.: +7-913-814-28-24
Эл. почта: shumskaya.ao@gmail.com

Iskhakova A.O.
Model to set up the texts class invariants

The paper proposes a model to form the texts class invariants based on the use of qualitative and quantitative characteristics. The setting up the characteristics list is based on the texts linguistic features. The model was used on the example of the generating invariants for two classes: original and artificial texts. The result is a list of features, distinguished in the classification of such texts, as well as the invariants of these classes. The presented data can be used to identify automatically generated texts based on taught neural network or to evaluate statistical text characteristics.

Keywords: text, texts class invariant, text characteristic, model, artificial text, original text.