ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2015 Управление, вычислительная техника и информатика № 2 (31)

УДК 004.912
DOI 10.17223/19988605/31/2

А.В. Глазкова

ОЦЕНКА СТЕПЕНИ БЛИЗОСТИ КАТЕГОРИЙ ТЕКСТОВ ПРИ РЕШЕНИИ ЗАДАЧ КЛАССИФИКАЦИИ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

Предлагается подход к оценке близости категорий текстов при решении задач классификации электронных документов на примере их отнесения к определенной возрастной аудитории. Введены понятия эквивалентности на множестве текстов и меры сходства категорий текстов. Приведен пример решения задачи классификации для взрослой и детской аудиторий.
Ключевые слова: извлечение информации; классификация текстов; математическое моделирование; обработка естественного языка.

Работа посвящена вопросам автоматической классификации документов на естественном языке. Задача классификации неструктурированной текстовой информации актуальна в первую очередь для решения проблем оптимизации информационного поиска в сети Интернет и хранилищах электронных документов. Быстрое увеличение количества информационных ресурсов порождает необходимость усовершенствования механизмов классификации текстов и обусловливает потребность в разработке новых методов и алгоритмов для решения данного рода задач.

При наличии обучающей выборки в существующих классификаторах, применяемых в различных информационных системах, используются методы машинного обучения, преимущественно основанные на байесовской модели и модели векторного пространства. В целях увеличения точности классификации текстов на естественном языке применяется оценка семантической близости текстов [1]. Одной из важных задач, решаемых при построении классификатора, является выбор классификационных признаков. При этом диапазон значений признаков может состоять как из двух значений, так и из конечного упорядоченного или неупорядоченного множества значений или бесконечного множества количественных значений [2–3].

Другой актуальной задачей, которой посвящено наше исследование, является не только отнесение данного текста к определенной категории, но и установление взаимосвязей между категориями.

Также рассматривается задача классификации текстов на примере их отнесения к той или иной возрастной категории адресатов. Возможность классифицировать тексты по возрастным группам их адресатов позволяет, в первую очередь, улучшать релевантность информационного поиска, а также усовершенствовать механизмы исключения из результатов поиска нежелательных запросов, например, сайтов, контент которых рассчитан на пользователя иной возрастной категории. Задача относится к числу слабоформализуемых за счет сложности естественного языка и многообразия его коммуникативных форм, поиск путей ее решения требует построения адекватных математических моделей процесса классификации.

Обсуждается подход к оценке степени близости категорий текстов, позволяющий оценить расстояние между рассматриваемыми категориями.

В контексте решаемой задачи тексты, адресованные одной возрастной группе читателей, должны быть отнесены в процессе классификации к одной категории. Однако на практике задача не решается столь однозначно, и тексты для одной возрастной категории адресатов могут также считаться адресованными другим возрастным аудиториям в том случае, когда они условно соответствуют уровням их коммуникативного развития. Например, тексты, предназначенные «соседним» возрастным группам, часто имеют незначительные отличия, что позволяет говорить о сходстве между ними, а также дает повод с определенной долей уверенности отнести текст, принадлежащий первой категории, ко второй. Также можно говорить о том, что текст, адресованный младшей возрастной категории, понятен и более

старшим читателям. Однако нельзя утверждать, что данный текст является в одинаковой степени интересным и информативным для представителей разных возрастных аудиторий, т.е. что он соответствует уровням коммуникативного развития обеих возрастных групп. Тогда в процессе классификации встает вопрос о величине различий между категориями текстов.

Под возрастной категорией понимается та возрастная группа, для которой данный текст, во-первых, является понятным с точки зрения различных разделов языкознания (лексики, синтаксиса и т.д.), во-вторых, соответствует уровню ее коммуникативного развития, является информативными и представляет интерес для аудитории.

Исходя из специфики поставленной задачи, особый интерес для исследования представляют работы, авторы которых извлекают из текста данные о его авторе или адресате. В ряде статей неоднократно рассматривались вопросы определения характеристик автора текста – его возраста, пола, типа личности и национальной принадлежности [4–6]. В [7] предлагается подход к применению методов распознавания адресанта текста для поиска записей террористической тематики в Интернете. В работах [8–10] рассматривается задача создания диалоговых систем, в контексте которой анализируются признаки, характеризующие текст с точки зрения его ориентации на различных адресатов. В [11] проведена классификация текстов по их автору с использованием потоковых методов классификации.

Подход к классификации поисковых запросов на основании оценки близости терминов предлагается в статье [12]. В [13] вводится метрика для оценивания синтаксического сходства между сверхкороткими текстами.

1. Постановка задачи

В [3] авторами был сформулирован подход к математическому моделированию задачи классификации. Отличие данного подхода от представленных ранее состоит в том, что он позволяет причислить текст к ряду пересекающихся категорий, однако дает возможность учесть то, что различия в уровнях коммуникативного развития представителей различных возрастных категорий не позволяют однозначно отнести текст из категории K_i в категорию K_j, где $i \le j \le n$.

Пусть дан текст T и множество категорий $K = \{K_1,K_2,\ldots,K_n\}$. Необходимо найти подмножество K_I – категории, которым может принадлежать текст:

$$T \sim K_I,\ K_I = \{K_i : T \sim K_i\},$$

где $i = j_1,j_2,\ldots,j_m$, $1 \le j \le n$, и $T \sim K_I$ означает принадлежность текста T к категории K_I.

Тогда категорию K_i можно представить в виде

$$K_i = \{q^K_j, w^K_j\},\ j = \overline{1,L},$$

где q^K_j – классификационный признак, w^K_j – весовой коэффициент классификационного признака, L – общее число классификационных признаков. Таким образом, категория определяется набором поставленных в соответствие классификационным признакам критических значений, а текст, в свою очередь, характеризуется своим признаковым описанием – набором значений классификационных признаков.

Под весовым коэффициентом классификационного признака подразумевается некоторая числовая оценка значимости признака q^K_j в разделении объектов на классы в сравнении с другими признаками, которая может быть определена экспериментально или на основе существующих методик оценивания весовых коэффициентов значимости критериев (например, [14, 15]). Значения весовых коэффициентов признака могут различаться в зависимости от особенностей множества рассматриваемых текстов (тематики, стиля и т.д.). Введение весовых коэффициентов позволяет выполнить масштабирование значений различных классификационных признаков [16], что дает возможность проводить вычисления попарных ковариаций наборов признаков, характеризующих категории, с целью определения степени их близости.

2. Введение отношения эквивалентности на множестве текстов

В контексте данной задачи можно говорить о некотором пороговом значении в различии между признаковыми описаниями двух текстов, которое позволит считать данные признаковые описания до-

статочно близкими и условно совпадающими. Будем называть такие тексты принадлежащими к одному таксономическому виду [17], тогда

$$T_i \cong T_j . \quad (1)$$

Отношение (1) является отношением эквивалентности, поскольку для него выполнены условия рефлексивности, симметричности и транзитивности. Пусть R – отношение эквивалентности на множестве текстов T, где

$$T_i \in T .$$

Тогда множество текстов можно разбить на непересекающиеся классы эквивалентности

$$T_i^{'} \mid R = \left\{ T_i^{'} \in T \mid T_i^{'} R T_i \right\}$$

и построить фактор-множество $T|R$ по отношению к эквивалентности R.

В рамках рассматриваемой задачи классы эквивалентности включают в себя тексты с совпадающими признаковыми описаниями, при этом фактор-множеством $T|R$ является множество всех классов эквивалентности, из чего следует [18]:

1) $T_i \in T_i^{'} \mid R$ для любого $T_i \in T$;
2) $T_i \mid R = T_j \mid R \Leftrightarrow T_i R T_j$ для любых T_i, T_j из T;
3) $T_i \mid R \neq T_j \mid R \Leftrightarrow T_i \mid R \cap T_j \mid R = \varnothing$;
4) $T = \bigcup\limits_{t \in T} t \mid R$.

3. Введение меры близости текстов

Тексты, попавшие в один класс эквивалентности, являются носителями одного признакового описания, которое и позволяет считать их эквивалентными. При этом число текстов, входящих в рассматриваемую выборку и принадлежащих одному классу эквивалентности, служит выражением абсолютного веса данного класса. Поскольку фактор-множество является набором всех возможных классов эквивалентности при заданном отношении эквивалентности, оно включает в себя все возможные классы текстов, подлежащих классификации.

В контексте решаемой задачи преобразование исходного множества текстов в фактор-множество является по своей сути процессом формирования содержимого классов текстов, адресованных определенной возрастной аудитории. Фактически же, как говорилось выше, тексты, адресованные одной возрастной группе читателей, можно в некотором смысле считать адресованными и другим возрастным аудиториям. Кроме того, одна возрастная категория на практике может включать в себя тексты, относящиеся к нескольким классам эквивалентности, которые будут в контексте поставленной задачи иметь незначительные отличия.

В таком случае можно говорить о некой количественной величине различий между категориями текстов, имеющих не совпадающие признаковые описания и относящихся к разным классам эквивалентности. Для описания этой ситуации необходимо задать функцию расстояния (метрику) на множестве текстов [19–21], тем самым сконструировав метрическое пространство. Если значение функции расстояния будет меньше некоторого порогового значения, категории будут считаться достаточно близкими друг другу, признаковые описания входящих в них текстов окажутся схожими. Таким образом, значение функции расстояния $\rho(K_i,K_j)$ является показателем сходства между категориями K_i и K_j, причем чем меньше значение этой функции, тем более схожи классы и, следовательно,

$$\rho\left(K_i, K_j\right) = 0 \Leftrightarrow K_i \cong K_j .$$

Тогда для произвольной категории K_i справедливо неравенство

$$\rho\left(K_i, K_{1j}\right) \le \rho\left(K_i, K_{2j}\right) \le \ldots \le \rho\left(K_i, K_{nj}\right),$$

при этом K_{nj} представляет собой тексты остальных категорий, включенные в некую обучающую выборку текстов из категории K_i размером n+1. Для каждой категории K_i нумерация остальных категорий бу-

дет индивидуальной. Те категории, значения функций расстояния для которых будут невелики, содержат тексты со схожими признаковыми описаниями, которые в некоторых условиях могут рассматриваться как тексты, адресованные одной возрастной аудитории. Пороговое значение, определяющее схожесть категорий и показывающее величину различий между ними, может быть задано двумя способами: на основании экспертной оценки или исходя из экспериментальных данных.

В качестве меры близости категорий может быть принято расстояние Махаланобиса, поскольку признаки объектов, между которыми устанавливается мера сходства, являются статистически зависимыми, а числовая оценка их значимости определяется весовыми коэффициентами. Тогда расстояние между категориями K_i и K_j, представленными в виде векторов, характеризующих их классификационные признаки

$$K_i = (q_{i1}^K, q_{i2}^K, ..., q_{iL}^K),$$
$$K_j = (q_{j1}^K, q_{j2}^K, ..., q_{jL}^K),$$

определяется следующим образом:

$$\rho(K_i, K_j) = \sqrt{(K_i - K_j)^T \Lambda_{\text{cat}}^T C_{\text{cat}}^{-1} (K_i - K_j)},$$

где Λ_{cat} – матрица весовых коэффициентов; C_{cat} – матрица ковариации, т.е. матрица, составленная из попарных ковариаций элементов векторов K_i и K_j.

Попарными ковариациями значений признаков, составляющих вектора K_i и K_j, при этом являются

$$\text{cov}(q_{in}^K, q_{jn}^K) = \frac{1}{n}\sum_{t=1}^{n}(q_{in_t}^K - \overline{q}_i^K)(q_{jn_t}^K - \overline{q}_j^K),$$

где $\overline{q}_i^K = \frac{1}{n}\sum_{t=1}^{n}(q_{in_t}^K)$, $\overline{q}_j^K = \frac{1}{n}\sum_{t=1}^{n}(q_{jn_t}^K)$ – средние значения выборок, $n = 1,\ldots, L$.

4. Пример применения предложенного подхода

Предложенный подход реализован в рамках разработки прототипа программного комплекса для проведения автоматической классификации текстов на русском языке на основании возрастных категорий их адресатов.

В ходе разработки и тестирования использовались тексты, включенные в «Базу данных метатекстовой разметки Национального корпуса русского языка (коллекция детской литературы)» [22]. База состоит из заведомо качественных и максимально разнообразных текстов на русском языке с известным жанром.

Во время проведения эксперимента выделялись две категории: тексты, адресованные взрослым, и тексты, адресованные детям. Это обусловлено соответствующим делением текстов в выборке, используемой для эксперимента. В дальнейшем планируется увеличить число классификационных категорий.

В ходе работы был экспериментально выделен ряд информативных признаков, характеризующих различия между категориями. В данном примере использованы три количественных классификационных признака: средняя длина предложений в тексте, средняя длина слова в тексте, процент многосложных слов (содержащий более трех слогов). Выбор этих признаков основан на работах в области удобочитаемости текстов и обсуждается в [23].

Каждому тексту из выборки (объем выборки – 500 детских и 500 взрослых текстов) было сопоставлено признаковое описание – набор значений признаков и их весовых коэффициентов. Во время эксперимента всем признакам были назначены равные веса. Поскольку в данном случае в выборке представлены тексты только двух категорий, в категории детских текстов были выделены тексты, напечатанные в журналах, целевой аудиторией которых являются дети среднего школьного возраста (выборка 1) и тексты авторов, пишущих для дошкольного и младшего школьного возраста (выборка 2). Обозначив K_{adult} категорию текстов для взрослых, K_{V1} – категорию текстов выборки 1 и K_{V2} – категорию текстов выборки 2, было предположено, что

$$\rho(K_{\text{adult}}, K_{V1}) \le \rho(K_{\text{adult}}, K_{V2}).$$

На графике (рис. 1) визуализированы значения признаков для текстов каждой категории, в целях удобства представления в каждой категории отображены по 30 текстов.

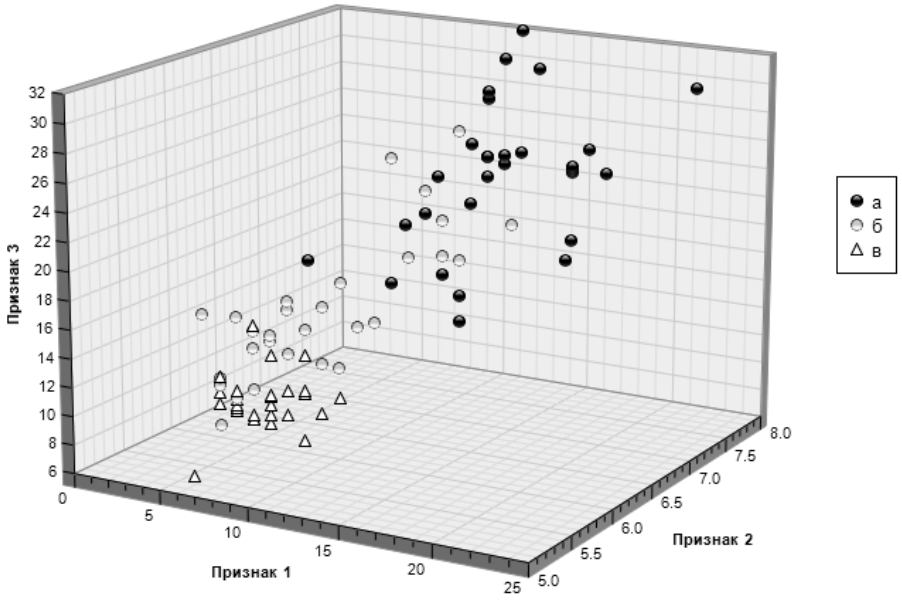

Рис. 1. Взаимное расположение текстов различных категорий: *а* – тексты категории K_{adult}; *б* – тексты категории K_{I1}; *в* – тексты категории K_{I2}

Основываясь на результатах, представленных в таблице, можно сделать вывод о том, что $\rho(K_{adult}, K_{I1}) < \rho(K_{adult}, K_{I2})$.

Полученные расстояния между категориями

Расстояние	Значение
$\rho(K_{adult}, K_{I1})$	1,7484
$\rho(K_{adult}, K_{I2})$	2,1157

Для оценки качества предложенного подхода использовалась процедура скользящего контроля. Функционал качества рассчитывался как сумма попарных внутриклассовых расстояний между текстами. В ходе эксперимента значение оценки скользящего контроля не превысило 7% от значения, полученного на тестовой выборке.

Заключение

В работе предложен и успешно протестирован подход к оценке близости категорий текстов при решении задач классификации электронных документов на примере их отнесения к определенной возрастной аудитории. Результаты применения предложенного подхода могут быть улучшены в ходе сопоставления классификационным признакам весовых коэффициентов, характеризующих их значимость.

ЛИТЕРАТУРА

1. *Нгуен Ба Нгок, Тузовский А.Ф.* Классификация текстов на основе оценки семантической близости терминов // Известия Томского политехнического университета. 2012. № 5(320). С. 43–48.
2. *Колесникова С.И.* Методы анализа информативности разнотипных признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2009. № 1(6). С. 69–80.
3. *Глазкова А.В., Захарова И.Г.* Подход к моделированию задачи автоматической классификации текстов (на примере их отнесения к определенной возрастной аудитории) // Вестник ТюмГУ. 2014. № 7. С. 205–211.
4. *Santosh K., Bansal R., Shekhar M., Varma V.* Author Profiling: Predicting Age and Gender from Blogs // Notebook for PAN at CLEF. Singapore, 2013. P. 119–124.

5. *Гречников Е.А., Гусев Г.Г., Кустарев А.А., Райгородский А.М.* Поиск неестественных текстов // Труды 11-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2009. Петрозаводск, 2009. С. 306–308.
6. *Nguyen D., Smith N., Rose C.* Author Age Prediction from Text using Linear Regression // Proc. of ICASSP. New-York, 2011. P. 267–276.
7. *Choi D., Ko B., Kim H., Kim P.* Text Analysis for Detecting Terrorism-Related Articles on the Web // Journal of Network and Computer Applications. 2013. V. 8, No. 5. C. 37–46.
8. *Akker R. op den, Traum D.* A comparison of addressee detection methods for multiparty conversations // Proc. of methods for multiparty conversations. Amsterdam, 2009. P. 99–106.
9. *Baba N., Huang H.-H., Nakano Y.I.* Addressee identification for human-human-agent multiparty conversations in different proxemics // Proc. 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction. Beijing, 2012.
10. *Lee H., Stolcke A., Shriberg E.* Using out-of-domain data for lexical addressee detection in human-human-computer dialog // Proc. North American ACL/Human Language Technology Conference. Atlanta, 2013. P. 215–219.
11. *Ашуров М.Ф.* Сравнение потоковых методов классификации текстов художественной литературы на основе сжатия информации и подсчета подстрок // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2014. № 4(29). С. 16–22.
12. *Attenberg J., Suel T.* Cleaning search results using term distance features // Proc. of AIRWeb. San Francisco, 2008. P. 21–24.
13. *Oliva J., Serrano J., Castillo M., Iglesias A.* A syntax-based measure for short-text semantic similarity // Journal of Network and Computer Applications. 2013. V. 8, No. 5. P. 37–46.
14. *Колесникова С.И.* О подходах к оцениванию информативности признаков в тестовом распознавании // Известия Томского политехнического университета. 2006. № 8(309). С. 23–27.
15. *Захарова И.Г., Пушкарев А.Н.* Математическое обеспечение динамической интегрированной экспертной системы поддержки принятия решений в маркетинге // Вестник ТюмГУ. 2012. № 4. С. 151–155.
16. *Luo Q., Chen E., Xiong H.* A semantic term weighting scheme for text categorization // Expert Systems with Applications. 2011. No. 38. P. 12708–12716.
17. *Маннинг К., Рагхаван П., Шютце Х.* Введение в информационный поиск. М. : Вильямс, 2011. 528 с.
18. *Дунаев В.В.* Об одной модели классификации // Научно-техническая информация. 1990. Сер. 2. № 3. С. 22–27.
19. *Мангалова Е.С., Агафонов Е.Д.* О проблеме выделения информативных признаков в задаче классификации текстовых документов // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2013. № 1(22). С. 96–103.
20. *Качановский Ю.П., Коротков Е.А.* Предобработка данных для обучения нейронной сети // Фундаментальные исследования. 2011. № 12-1. С. 117–120.
21. *McLachlan G.J.* Discriminant Analysis and Statistical Pattern Recognition. New Jersey : Wiley Interscience, 1992. 552 p.
22. «*База* данных метатекстовой разметки Национального корпуса русского языка (коллекция детской литературы)». 2014.
23. *Глазкова А.В.* Проверка информативности классификационных признаков в задаче автоматической классификации текстов на естественном языке // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2015) : материалы V Междунар. науч.-техн. конф. Минск, 2015. С. 541–544.

Глазкова Анна Валерьевна. E-mail: anya_kr@aol.com
Тюменский государственный университет

Поступила в редакцию 3 февраля 2015 г.

Glazkova Anna V. (Tyumen State University, Russian Federation).
The evaluation of the proximity of text categories for solving electronic documents classification tasks.
Keywords: information extraction; mathematical modeling; natural language processing; text classification.

DOI 10.17223/19988605/31/2

The article deals with the problem of classification of texts by the example of their assignment to a particular age group of recipients. In practice, texts for one age category of recipients can also be considered as addressed to another age when they conditionally correspond to the levels of their communicative development. In this case, we can discuss the magnitude of the differences between text categories.

In previous research authors have formulated an approach to mathematical modeling of the problem of classification. Suppose given a text T and a set of categories $K = \{K_1, K_2, \ldots, K_n\}$. Need to find a subset of K_I, i.e., a category, which may be associated with the text:

$$T \sim K_I, \; K_I = \{K_i : T \sim K_i\},$$

where $i = j_1, j_2, \ldots, j_m$ and $1 \le i \le n$.

So, the category K_i can be presented as

$$K_i = \{q_j^K, w_j^K\}, \; j = \overline{1, L},$$

where q^K_j is classification feature, w^K_j is a weight coefficient of classification feature, L is a total count of classification features.

If the feature descriptions of the two texts are identical, we call these texts belonging to the same taxonomic rank, and then we have

$$T_i \cong T_j.$$

This is an equivalence relation because it satisfies the conditions of reflexive, symmetric, and transitive. Consequently, many texts can be divided into disjoint equivalence classes and one can construct factor set by the equivalence relation.

Texts having an equivalence class are carriers of one of the feature descriptions, which allows us to consider they are equivalent. Converting the original set of texts in the factor set is the process of forming the contents of the classes of texts addressed to a specific age audience. In fact, as mentioned above, we are talking about a certain quantifying the differences between the categories of texts with no matching feature descriptions and belonging to different classes of equivalence:

$$\rho\left(K_i, K_j\right) = 0 \Leftrightarrow K_i \cong K_i .$$

Measure of proximity of categories may be defined as the Mahalanobis distance because the features of objects are statistically dependent and their relevance is determined by the weight coefficient s. Then, the distance between the categories of K_i and K_j represented as vectors characterizing their classification features

$$K_i = (q_{i1}^K, q_{i2}^K, ..., q_{iL}^K),$$
$$K_j = (q_{j1}^K, q_{j2}^K, ..., q_{jL}^K),$$

are defined as

$$\rho\left(K_i, K_j\right) = \sqrt{(K_i - K_j)^T \Lambda_{\text{cat}}^T C_{\text{cat}}^{-1} (K_i - K_j)},$$

where Λ_{cat} is a matrix of weight coefficients, C_{cat} is a matrix of covariance, which is the matrix built by pairwise covariance of the elements in vectors K_i and K_j.

Pairwise covariance of features values for vectors K_i and K_j is:

$$\text{cov}\left(q_{in}^K, q_{jn}^K\right) = \frac{1}{n}\sum_{t=1}^{n}(q_{in_t}^K - \overline{q}_i^K)(q_{jn_t}^K - \overline{q}_j^K)),$$

where $\overline{q}_i^K = \frac{1}{n}\sum_{t=1}^{n}(q_{in_t}^K)$, $\overline{q}_j^K = \frac{1}{n}\sum_{t=1}^{n}(q_{jn_t}^K)$ are average values in the text samples, $n = 1,\ldots,L$.

The proposed approach to the evaluation of the proximity of categories texts is implemented due to the development of the prototype of software system for automatic classification of texts in Russian based on age categories of recipients.

REFERENCES

1. Nguen, B.N., Tuzovskiy, A.F. (2012) Text classification based on estimation of terms semantic similarity. *Izvestiya Tomskogo politekhnicheskogo universiteta – Bulletin of the Tomsk Polytechnic University*. 5 (320). pp. 43-48. (In Russian).
2. Kolesnikova, S.I. (2009) Methods of analysis of different-type features informativity. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 1 (6). pp. 69-80. (In Russian).
3. Glazkova, A.V. & Zakharova, I.G. (2014) Podkhod k modelirovaniyu zadachi avtomaticheskoy klassifikatsii tekstov (na primere ikh otne-seniya k opredelennoy vozrastnoy auditorii) [Approach to modeling of automatic text classification problem (case study of the audience age prediction)]. *Vestnik TyumGU – Tyumen State University Herald*. 7. pp. 205-211.
4. Santosh, K., Bansal, R., Shekhar, M. & Varma, V. (2013) Author Profiling: Predicting Age and Gender from Blogs. *Notebook for PAN at CLEF*. Singapore, 2013. p. 119-124.
5. Grechnikov, E.A., Gusev, G.G., Kustarev, A.A. & Raygorodskiy, A.M. (2009) [Unnatural texts search]. *Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii* [Digital Libraries: Advanced Methods and Technologies, Digital Collections – RCDL'2009]. Proc. of the 11th All-Russian Scientific Conference. Petrozavodsk. pp. 306-308. (In Russian).
6. Nguyen, D., Smith, N. & Rose, C. (2011) Author Age Prediction from Text using Linear Regression. *Proc. of ICASSP*. New-York, 2011. pp. 267-276.
7. Choi, D., Ko, B., Kim, H. & Kim P. (2013) Text Analysis for Detecting Terrorism-Related Articles on the Web. *Journal of Network and Computer Applications*. 8 (5). pp. 37-46. DOI: 10.1016/j.jnca.2013.05.007
8. Akker, R. op den & Traum, D. (2009) A comparison of addressee detection methods for multiparty conversations. *Proc. of Methods for Multiparty Monversations*. Amsterdam. pp. 99-106.
9. Baba, N., Huang, H.-H. & Nakano, Y.I. (2012) Addressee identification for human-human-agent multiparty conversations in different proxemics. *Proc. 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*. Beijing.
10. Lee, H., Stolcke, A. & Shriberg, E. Using out-of-domain data for lexical addressee detection in human-human-computer dialog. *Proc. North American ACL/Human Language Technology Conference*. Atlanta. pp. 215-219.
11. Ashurov, M.F. (2014) Comparison of stream-based fiction text classification methods based on data compression and counting substrings. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 4 (29). pp. 16-22. (In Russian).
12. Attenberg. J. & Suel, T. Cleaning search results using term distance features. *Proc. of AIRWeb*. San Francisco. pp. 21-24.
13. Oliva, J., Serrano, J., Castillo, M. & Iglesias, A. (2013) A syntax-based measure for short-text semantic similarity. *Journal of Network and Computer Applications*. 8 (5). pp. 37-46. DOI: 10.1016/j.datak.2011.01.002
14. Kolesnikova, S.I. (2006) O podkhodakh k otsenivaniyu informativnosti priznakov v testovom raspoznavanii [On the approaches to estimation if feature informativity in the test recognition]. *Izvestiya Tomskogo politekhnicheskogo universiteta – Bulletin of the Tomsk Polytechnic University*. 8(309). pp. 23-27.

15. Zakharova, I.G. & Pushkarev, A.N. (2012) Matematicheskoe obespechenie dinamicheskoy integrirovannoy ekspertnoy sistemy podderzhki prinyatiya resheniy v marketing [Software the dynamic integrated expert system of support of decision-making in marketing]. *Vestnik TyumGU – Tyumen State University Herald*. 4. pp. 151-155.
16. Luo, Q., Chen, E. & Xiong, H. (2011) A semantic term weighting scheme for text categorization. *Expert Systems with Applications*. 38. pp. 12708-12716. DOI: 10.1016/j.eswa.2011.04.058
17. Manning, C., Raghavan, P. & Schutze, H. (2011) *Vvedenie v informatsionnyy poisk* [Introduction to information retrieval]. Translated from English by D. Klyushin. Moscow: Williams.
18. Dunaev, V.V. (1990) Ob odnoy modeli klassifikatsii [Model of classification]. *Nauchno-tekhnicheskaya informatsiya*. 2 (3). pp. 22-27.
19. Mangalova, E.S. & Agafonov, E.D. (2013) On features selection approach for text mining problem. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 1(22). pp. 96-103. (In Russian).
20. Kachanovskiy, Yu.P. & Korotkov, E.A. (2011) Preprocessing data for training neural networks. *Fundamental'nye issledovaniya – Fundamental research*. 12-1. pp. 117-120. (In Russian).
21. McLachlan, G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. New Jersey: Wiley Interscience.
22. Database of metatextual marking of the Russian National Corpus (a collection of children's literature). 2014. (In Russian).
23. Glazkova, A.V. (2015) [Classification features informational content testing for automatic natural texts classification task]. *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem (OSTIS-2015)* [Open semantic technologies for intelligent systems (OSTIS-2015)]. Proc. of the 5th International Scientific and Engineering Conference. Minsk. pp. 541-544. (In Russian).