Вестник Тюменского государственного университета.
2014. № 7. Физико-математические науки. Информатика. 205-211

© А.В. ГЛАЗКОВА, И.Г. ЗАХАРОВА
Тюменский государственный университет
anya_kr@aol.com, izaharova@yandex.ru

УДК 004.912

ПОДХОД К МОДЕЛИРОВАНИЮ ЗАДАЧИ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ
(НА ПРИМЕРЕ ИХ ОТНЕСЕНИЯ К ОПРЕДЕЛЕННОЙ ВОЗРАСТНОЙ АУДИТОРИИ)

APPROACH TO MODELING OF AUTOMATIC TEXT CLASSIFICATION PROBLEM
(CASE STUDY OF THE AUDIENCE AGE PREDICTION)

АННОТАЦИЯ. В статье рассматривается задача автоматической классификации текстов на примере их отнесения к определенной возрастной аудитории. В работе приводятся несколько возможных путей формализации данной задачи, обсуждаются их преимущества и недостатки. Предлагается подход к математическому моделированию предметной области, подразумевающий представление категории как множества классификационных признаков и их критических значений, а текста соответственно — как множества признаков и значений признаков. В таком случае классификация множества текстов по некоторому признаку может быть представлена как отображение множества текстов во множество допустимых значений этого признака. В заключительной части работы обосновывается возможность использования нейросетевых технологий в качестве средства компьютерной реализации алгоритмов классификации и приводится краткий обзор работ, посвященных вопросам применения нейронных сетей для автоматической классификации текстов. Подход, предложенный авторами, реализован с использованием нейросетевых технологий в виде прототипа программного комплекса.

SUMMARY. The article considers the problem of automatic text classification as a case study of the audience age prediction from the text. The paper describes some possible ways to formalize the problem and discusses their advantages and disadvantages. It is proposed an approach to mathematical modeling of the domain, which implies the representation of a category as a set of classification features and their critical values and a text as a set of text features and their values. In such a case, the classification by a feature can be represented as a mapping of the set of texts in the set of permissible values for this feature. In the final part of the paper the possibility of using neural network technology as a tool for computer implementation of classification algorithms is proved and a brief review of the literature on the application of neural networks for automatic text classification is provided. The approach suggested by the authors is implemented using neural network technology in the form of a prototype software system.

КЛЮЧЕВЫЕ СЛОВА. Математическое моделирование, классификация документов, извлечение информации.

KEYWORDS. Mathematical modeling, document classification, information extraction.

Одной из актуальных задач обработки неструктурированной текстовой информации является автоматическая классификация документов. Связанные с этим вопросы освещаются в ряде научных работ, где в зависимости от основания классификации предлагаются различные подходы к моделированию процедур систематизации текстов [1-3].

В статье рассматривается задача классификации текстов на примере их отнесения к определенной возрастной аудитории. Данная задача является слабоформализуемой за счет многозначности, разнообразия и сложности естественного языка. В этом контексте важной представляется проблема моделирования автоматической классификации текстов с учетом последующей компьютерной реализации разработанных математических моделей.

В общем виде задача классификации текстов состоит в следующем. Имеется текст T и множество категорий $K = \{K_1, K_2, ..., K_n\}$, с которыми он должен быть сопоставлен. Задача сводится к тому, чтобы выбрать категорию, к которой относится текст T:

$$T \sim K_i,\ K_i \in K.$$

При начальной формализации не были учтены некоторые особенности предметной области. Например, при рассмотрении задачи классификации в общем виде считалось, что категории $K_1, K_2, ..., K_n$ — возрастные группы адресатов, являются независимыми, и, следовательно, отнесение текста к категории K_i означало, что он не может быть причислен к прочим категориям из множества K. В действительности же очевидно, что текст, адресованный некоторой возрастной аудитории, может предназначаться и другим возрастным группам. Например, отношение текста к некой категории подразумевает также то, что он будет понятен читателям старших возрастов. Учитывая эту особенность, отношения между категориями можно представить в виде $K_1 \subset K_2 \subset ... \subset K_n$, тогда:

$$T \sim K_i \Rightarrow T \sim K_j,\ j = \overline{i,n}.$$

Обозначенный подход к моделированию предметной области позволяет принять во внимание то, что текст из категории K_i принадлежит в то же время $K_i, K_{i+1}, ..., K_n$.

В качестве недостатка предложенного пути формализации можно отметить, что речь в предыдущем примере идет преимущественно не об адресованности текста определенной аудитории, а о его понятности представителям той или иной возрастной группы. Так, в рамках своей коммуникативной деятельности автор составляет текст, имея установку на максимально полное доведение до адресата. Речь должна быть ориентирована на слушателя, и естественным следствием такой установки является намерение автора использовать такие содержание и структуру, которые в своей совокупности были бы адекватны пониманию «идеального» реципиента, которому предназначен текст [4]. В нашем же примере особый интерес вызывает то, что содержание и структура текста, адресованного читателям самого младшего возраста, хотя и будут понятны другим категориям реципиентов, могут не соответствовать уровню коммуникативного развития адресатов, относящихся к другим категориям.

Учитывая описанную особенность, можно сформулировать задачу классификации следующим образом. Пусть дан текст T и множество категорий $K = \{K_1, K_2, ..., K_n\}$. Необходимо найти подмножество K_I — категории, которым может принадлежать текст:

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

где $i = j_1, j_2, ..., j_m$ и $1 \le i \le n$.

Данный подход к формализации 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 $j = \overline{i,n}$.

Отталкиваясь от формальной постановки задачи, можно представить категорию K_i в виде:

$$K_i = \{q_j^K, V_j^i\}, j = \overline{1,L},$$

где q_j^K — классификационный признак, V_j^i — критическое значение j-го признака из категории K_i, L — общее число классификационных признаков.

Таким образом, категория однозначно определяется набором поставленных в соответствие классификационным признакам критических значений. При этом критическое значение V_j^i может задаваться как в виде интервальной оценки:

$$V_j^i \in (l_j^i, r_j^i),\ V_j^i \in (l_j^i, \infty),\ V_j^i \in (-\infty, r_j^i),$$

где l_j^i, r_j^i — определяют критический интервал; так и в форме точного значения:

$$V_j^i = v_j^i \Leftrightarrow V_j^i = (v_j^i - \varepsilon_j^i, v_j^i + \varepsilon_j^i), \varepsilon_j^i = 0 \cdot$$

Для каждого признака q_j^K имеем отображение множества текстов во множество допустимых значений признака Q_f:

$$q_j^K \equiv f_j : \Im \to Q_f \cdot$$

В зависимости от множества Q_f (и соответствующей шкалы измерений) признаки могут быть отнесены к следующим типам [5]:

1) бинарный признак: $Q_f = \{0,1\}$ (например, наличие/отсутствие специальной лексики в тексте, иллюстраций и т.п.);

2) номинальный признак: Q_f — конечное множество (структурный тип текста — проза или поэзия; литературная форма — рассказ, повесть, роман и т.д.);

3) порядковый признак: Q_f — конечное упорядоченное множество (период создания, уровень образования аудитории);

4) количественный признак: $Q_f \in R$ (число сложных синтаксических конструкций, число предложений).

В общем случае при определении значения того или иного признака представляется не вполне корректным отталкиваться от предположения о детерминированности этих значений, поскольку они определяются на основе данных случайной выборки. В таком случае истинность значения признака q_j^T для

текста T определяется с вероятностью 1 - α, где α — уровень значимости, a — нижнее значение интервала, b — верхнее значение:

$$P(a \le q_j^T \le b) = 1 - \alpha .$$

Отображение текста T в его признаковое описание допустимо записать в виде:

$$\varphi : T \to F_T ,$$

где признаковое описание, которое в контексте данной задачи возможно отождествлять с самим текстом, может быть представлено в виде вектора F_T:

$$F_T = (f_1(T), f_2(T), ..., f_L(T))$$

В качестве примера рассмотрим текст T, определяемый набором значений признаков следующим образом: $F_T = (0, "проза", "min\_A", 0{,}25)$. Для классификации используются категории K_1, K_2, K_3, K_4, K_5, критические значения признаков для которых приведены в табл. 1.

Таблица 1

Наборы критических значений признаков для категорий K_1, K_2, K_3, K_4, K_5

	q_1^K	q_2^K	q_3^K	q_4^K
K_1	**0**	**проза**	**Тип 1**	**0-0,5**
K_2	**0**	**проза**	**Тип 1**	0,5-1
K_3	1	поэзия	Тип 2	**0-0,5**
K_4	1	поэзия	Тип 2	0,5-1
K_5	Остальные значения			

Сопоставляя значения признаков текста T с критическими значениями признаков категорий, получаем, что T относится к категории K_1 в случае, если классификация подразумевает совпадение значений по всем признакам q_j^K. В противном случае, а также в ситуации, когда q_4^K имеет меньшую значимость в сравнении с другими признаками, текст T может относиться одновременно к категориям K_1 и K_2. Для оценки важности признаков можно использовать методику оценивания весовых коэффициентов значимости критериев, описанную в [6].

Если признаковые описания двух текстов совпадают, будем называть эти тексты принадлежащими к одному таксономическому виду [7]:

$$\varphi(T_i) = \varphi(T_j) \Rightarrow T_i \cong T_j .$$

Это отношение является отношением эквивалентности, поскольку для него выполнены следующие условия:

1) рефлексивность: $T_i \cong T_i$;

2) симметричность: $T_i \cong T_j \Rightarrow T_j \cong T_i$;

3) транзитивность: $T_i \cong T_j, T_j \cong T_k \Rightarrow T_i \cong T_k$.

Следовательно, множество текстов можно разбить на непересекающиеся классы эквивалентности и построить фактор-множество по отношению эквивалентности ($\cong$).

Компьютерная реализация алгоритмов классификации с использованием отображения $\mathfrak{I}$ множества текстов во множество допустимых значений признака подразумевает разработку комплексной программы с большим количеством входов и выходов, то есть позволяет применять для осуществления классификации нейронную сеть, где f_j следует интерпретировать как функцию активации нейрона, а весовые коэффициенты, служащие для оценивания степени влияния каждого критерия на вероятность отнесения объекта к той или иной категории, — как межнейронные связи [8]. Предпочтительность использования нейронной сети определяется также ее способностью к обучению и обобщению накопленных знаний. Разработке методов классификации текстов на основе нейронных сетей посвящен ряд работ российских и зарубежных ученых. Так, об удобстве использования нейросетевых технологий для проведения иерархической классификации документов говорится в [9]. В [10] проводится сравнение алгоритмов классификации с помощью деревьев решений и нейронных сетей прямого распространения на примере задачи классификации текстов по авторским стилям; в [11], [12] обсуждается применение нейронных сетей с обратным распространением для рубрикации текстов, представленных в виде векторов, составленных из значимых терминов и их числовых характеристик; в [13] описывается организация нейронной сети для решения задач классификации текстов по автору или тематическим категориям.

Подход к моделированию, предложенный в данной работе, был реализован в виде прототипа программного комплекса на примере автоматической классификации текстов по их возрастной аудитории. При разработке и тестировании использовалась база Национального корпуса русского языка [14].

СПИСОК ЛИТЕРАТУРЫ

1. Thakkar, K., Shrawankar, U. Test Model for Text Categorization and Text Summarization // International Journal on Computer Science and Engineering. 2013. № 3. Pp. 1539-1545.

2. Zhang, M., Zhou, Z. Multi-Label Neural Networks with Applications to Functional Genomics and Text Categorization // IEEE Transactions on Knowledge and Data Engineering. 2006. №18 (10). Pp. 1338-1351.

3. Борисова Н.Ф., Кочуева З.А., Шаронова Н.В., Хайрова Н.Ф. Моделирование процедур систематизации и классификации информационных объектов методом компараторной идентификации // Вестник Херсонского национального технического университета. 2012. № 1. С. 91-95.

4. Каменская О.Л. Текст и коммуникация. М.: Высшая школа, 1990. С. 78.

5. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.

6. Захарова И.Г., Пушкарев А.Н. Математическое обеспечение динамической интегрированной экспертной системы поддержки принятия решений в маркетинге // Вестник Тюменского государственного университета. 2012. № 4. Серия «Физико-математические науки. Информатика». С. 151-155.

7. Дунаев В.В. Об одной модели классификации // Научно-техническая информация. Сер. 2. 1990. № 3. С. 22-27.

8. Джонс М.Т. Программирование искусственного интеллекта в приложениях. М., 2013. 312 с.

9. Ruiz, M., Srinivasan, P. Hierarchical Text Categorization Using Neural Networks // Information Retrieval. 2002. № 5 (1). С. 87-118.

10. Шевелев О.Г., Петраков А.В. Классификация текстов с помощью деревьев решений и нейронных сетей прямого распространения // Вестник Томского государственного университета. 2006. № 290. С. 300-307.

11. Jo, T. NTC (Neural Text Categorizer): Neural Network for Text Categorization // International Journal of Information Studies. 2010. № 2(2). С. 83-96.

12. Ramasundaram, S., Victor, S. Text Categorization by Backpropagation Network // International Journal of Computer Applications. 2010. № 8(6). Pp. 1-5.

13. Кошкин Д.Е. Кластеризация текстов с помощью нейронных сетей и временная оценка работы алгоритма // Философские проблемы информационных технологий и киберпространства. 2012. № 1. С. 72-78.

14. Национальный корпус русского языка. 2003-2014. URL: ruscorpora.ru (дата обращения: 30.04.2014).

REFERENCES

1. Thakkar, K., Shrawankar, U. Test Model for Text Categorization and Text Summarization. *International Journal on Computer Science and Engineering*. 2013. № 3. Pp. 1539-1545.

2. Zhang, M., Zhou, Z. Multi-Label Neural Networks with Applications to Functional Genomics and Text Categorization. *IEEE Transactions on Knowledge and Data Engineering*. 2006. №18 (10). Pp. 1338-1351.

3. Borisova, N.F., Kochueva, Z.A., Sharonova, N.V., Hajrova, N.F. Modeling of systematization and classification procedures for data objects by identifying comparator. *Vestnik Hersonskogo nacional'nogo tehnicheskogo universiteta — Kherson National Technical University Herald*. 2012. № 1. Pp . 91-95. (in Russian).

4. Kamenskaja, O.L. *Tekst i kommunikacija* [Text and communication]. Moscow, 1990. 78 p. (in Russian).

5. Ajvazjan, S.A., Buhshtaber, V.M., Enjukov, I.S., Meshalkin, L.D. *Prikladnaja statistika: klassifikacija i snizhenie razmernosti* [Applied statistics: classification and reduction of dimension]. Moscow, 1989. 607 p. (in Russian).

6. Zaharova, I.G., Pushkarev, A.N. Software for the dynamic integrated expert support system of decision-making in marketing. *Vestnik Tjumenskogo gosudarstvennogo universiteta — Tyumen State University Herald*. 2012. № 4. Pp. 151-155. (in Russian).

7. Dunaev, V.V. On a model of classification. *Nauchno-tehnicheskaja informacija. Ser. 2 — Scientific and technological information. Series 2*. 1990. № 3. Pp. 22-27. (in Russian).

8. Jones, M.T. *Programmirovanie iskusstvennogo intellekta v prilozhenijah* [Artificial intelligence application programming] / Transl. fr. Eng. by A.I. Osipov. Moscow, 2013. 312 p. (in Russian).

9. Ruiz, M., Srinivasan, P. Hierarchical Text Categorization Using Neural Networks. *Information Retrieval*. 2002. № 5 (1). С. 87-118.

10. Shevelev, O.G., Petrakov, A.V. Text classification using decision trees and neural backpropagation networks. *Vestnik Tomskogo gosudarstvennogo universiteta — Tomsk State University Herald*. 2006. № 290. Pp. 300-307. (in Russian).

11. Jo, T. NTC (Neural Text Categorizer): Neural Network for Text Categorization. *International Journal of Information Studies*. 2010. № 2(2). С. 83-96.

12. Ramasundaram, S., Victor, S. Text Categorization by Backpropagation Network. *International Journal of Computer Applications*. 2010. № 8(6). Pp. 1-5.

13. Koshkin, D.E. Texts clusterization using neural networks and temporal evaluation of the algorithm. *Filosofskie problemy informacionnyh tehnologij i kiberprostranstva — Philosophical Problems of Information Technology and Cyberspace*. 2012. № 1. Pp. 72-78. (in Russian).

14. Russian National Corpus. 2003-2014. URL: http: ruscorpora.ru. (date accessed: 30. 04.2014). (in Russian).

Авторы публикации

Глазкова Анна Валерьевна — ассистент кафедры программного обеспечения Института математики и компьютерных наук Тюменского государственного университета

Захарова Ирина Гелиевна — заведующая кафедрой программного обеспечения Института математики и компьютерных наук Тюменского государственного университета, доктор педагогических наук, профессор

Authors of the publication

Anna V. Glazkova — Assistant, Department of Software, Institute of Mathematics and Computer Sciences, Tyumen State University

Irina G. Zakharova — Dr. Sci. (Pedag.), Professor, Head of the Department of Software, Institute of Mathematics and Computer Sciences, Tyumen State University