АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 004.85.021:[025.4.025:82 – 1/–9]

Н. Н. Буйлова

Классификация текстов по жанрам с помощью алгоритмов машинного обучения*

Рассмотрена проблема классификации документов по жанрам, выделены основные характеристики текста, используемые для распознавания его жанра, и описаны наиболее применяемые алгоритмы машинного обучения. Приведенные методы служат для классификации научных, технических, публицистических и художественных текстов.

Ключевые слова: *классификация текстов, определение жанра, машинное обучение*

ВВЕДЕНИЕ

Классификация документов – одна из основных подзадач информационного поиска. Из неструктурированных данных пользователь стремится получить документы, релевантные его запросу. Часто поисковый запрос включает не только ключевые слова, но требования к функциональному стилю документа, удобочитаемости текста, гендерной принадлежности автора, его возрасту и т. д. В этом случае документ должен иметь набор метатегов, описывающих его характерные особенности. Для крупных хранилищ текстов, таких как корпуса и библиотеки, определение тегов является ключевым требованием метаразметки.

Одним из способов классифицирования является определение функциональных стилей текстов. В свою очередь, функциональные стили делятся на жанры [1]. Согласно определению М.М. Бахтина, жанр – это «устойчивый тематически, композиционно и стилистически тип высказывания» [2, с. 255]. В отличие от литературной формы (обладающей четкими критериями) понятие жанра текста с трудом поддается формализации: при том, что параметры, описывающие конкретный жанр, вывести не удается, экспертная оценка человека, как и машинный классификатор, пользуясь неявными признаками, в большинстве случаев верно определяют жанр текста. Выявление скрытых признаков подробно рассматривается в статье [3] на примере коллекции медицинских текстов.

Автоматическое определение функциональных стилей текста – одно из приоритетных направлений классификации документов, однако в настоящий момент эта задача в значительной степени решена – в том числе в связи с особо актуальной для Всемирной паутины проблемой фильтрации спама, например [4]. Однако жанровое многообразие (особенно для художественной литературы) все еще недостаточно изученная область, что, по всей вероятности, связано с отсутствием четкого набора формальных признаков жанра; более того, анализ жанровой принадлежности затрудняется тем, что каждая пара жанров различается уникальными параметрами – от длины текста в символах до структуры предложения.

Несмотря на значительную сложность, задача автоматической классификации текстов по жанрам является привлекательной сферой исследований как в прикладном (вышеупомянутая фильтрация спама, индексация документов при поиске в корпусе или библиотеке, «родительский контроль» Интернета и т. д.), так и в теоретическом плане, например, четкое определение стилистических особенностей того или иного жанра может использоваться в обучении студентов словесности.

В прикладной области особое значение приобретает автоматическое определение жанра для структурирования библиотек и баз данных – постоянно пополняющиеся коллекции научных публикаций и художественной литературы растут со скоростью, которая не позволяет классифицировать тексты вручную, а объемы подобных корпусов на сегодняшний день таковы, что и поиск без привлечения алгоритмов не принесет плодов. Стоит отметить, что эта задача существует для корпусов на различных языках, а для большей части европейских и азиатских языков уже

* Исследование выполнено на основании гранта Российского научного фонда (проект № 16-18-02071 «Пограничный русский: оценка сложности восприятия русского текста в теоретическом, экспериментальном и статистическом аспектах»).

есть корпуса текстов, сравнимые с первым репрезентативным корпусом Брауна (т. е. содержащие не менее одного миллиона словоупотреблений) [5-7].

Актуальность рассматриваемой нами проблемы настолько высока, что уже имеются обзоры методов классификации текстов [8-10]. Однако эти обзоры сосредотачивались либо на технической стороне реализации алгоритмов, либо на рассмотрении общих подходов к классификации текста. Согласно нашим сведениям, не было предпринято попытки рассмотреть классификацию по жанрам в отдельности. Наш обзор посвящен методам определения жанра в научной, технической, публицистической и художественной литературе.

РАННИЕ ПУБЛИКАЦИИ, ПОСВЯЩЕННЫЕ ПРОБЛЕМЕ КЛАССИФИКАЦИИ ЖАНРА

Первые прикладные работы по распознаванию жанра текста появились в середине 1990-х гг. [11-12], в их основе лежали теоретические работы Д. Бибера [13]. Проведенные на Корпусе текстов университета Брауна эксперименты применяли методы дискриминантного анализа [12] и логистической регрессии, в том числе с использованием нейросети [11]. Базовым описанием текстов были частеречные характеристики текста, а также различные меры удобочитаемости. На их основе текстам присваивалось три типа признаков – сложность текста, наличие нарратива и жанр. Несмотря на высокое качество исходной разметки, доля правильного распознавания жанра не поднималась выше 83%. Более того, размеры этих выборок были недостаточны для обсуждения качества алгоритма.

Представляется необходимым упомянуть работу [14], в которой описанная комбинация алгоритмов позволила значительно повысить качество снятия неоднозначности. В дальнейшем этот подход лег в основу современных методов «мешка слов и деревьев решений», получивших широкое развитие в середине 2000-х гг. из-за стремительного разрастания Интернета и необходимости классифицировать все большее количество текстов. Тогда же появляются способы классификации, основанные на частеречной разметке [15]. На тот момент использование частеречной разметки позволяло распознавать жанр текста с высокой (96,9%) точностью в случае качественных исходных данных (газетная статья), в противном случае точность падала до 85,7% (сообщения форумов). Другим крупным ответвлением классификации документов стала классификация с использованием HTML-разметки, позволяющей комбинировать квантитативные методы описания самого текста с нетекстовыми элементами разметки гипертекста [16, 17]. Кроме того, следует упомянуть использование синтаксически размеченных корпусов (*treebank*), позволяющих проводить анализ дискурсивных связей [18].

Сегодня существует ряд способов машинной классификации текстов по жанрам, которые используют классические методы машинного обучения. Далее мы рассмотрим основные из них: наивный байесовский классификатор, деревья решений, случайный лес, метод опорных векторов.

ХАРАКТЕРИСТИКИ ТЕКСТА, ИСПОЛЬЗУЕМЫЕ ПРИ КЛАССИФИКАЦИИ

Функционирование классификаторов любой природы предполагает предобработку текстов для получения машиночитаемых данных. На сегодняшний день создано много способов описания отдельных репрезентативных признаков документа. Эту задачу можно описать как представление текстов в виде векторов, атрибуты которых делятся на два типа – частотные (каждое значение в векторе d соответствуют количеству вхождений признаков в документ d) и бинарные (каждое значение в векторе – бинарное и отражает факт присутствия признака в документе) [19].

Наиболее простым способом представления текста являются униграммы, также называемые «мешком слов» (*bag of words*), которые представляют собой набор слов документа без каких-либо связей между ними. Несмотря на свою простоту, модель имеет ряд недостатков: так, не учитываются грамматические связи между словами, порядок токенов и вероятность совместной встречаемости слов [20, 21].

Метод «мешка слов» обладает большим адаптационным потенциалом: например, в работе [22] помимо классического «мешка слов» использовалось расширение этого понятия, а именно подстроки из каждого предложения (если в стандартном методе «мешка слов» документ разбивается на токены от пробела до пробела, то в этой работе выделялись последовательности из нескольких слов, идущих в предложении подряд).

Помимо униграмм, применяются также би- и триграммы, которые могут представлять собой как букво- так и словосочетания из двух или трех элементов (эти комплексы являются частным случаем n-грамм, однако сочетания с n больше трех встречаются реже из-за больших объемов корпусов, необходимых для сбора достаточно репрезентативной выборки). В работе [23] критерий кластеризации, основанный на близости между двухбуквенными распределениями текстов, позволяет правильно идентифицировать автора с ошибкой не более 5%, а жанр – с ошибкой не более 15%.

Кроме простого «мешка слов» возможно использование процессированного списка слов, характеризующих документ. Такая метрика называется TF-IDF (TF – *term frequency*, IDF – *inverse document frequency*), и она оценивает важность слова в определенном документе относительно других текстов коллекции [24, 25]. Синтаксическая аннотация текста – еще одна важная характеристика, используемая для описания стиля [26] и, таким образом, имеющая потенциал в качестве признака при классификации жанров. Синтаксис, понятый как способ связи слов в предложении и предложений в тексте, характеризует строение фраз и иных структурных единиц, что позволяет выделить важные черты документа и жанра в целом [27].

Еще одним интересным типом признаков для классификации являются дискурсивные связи, а именно – способы объединения текста в единое целое при помощи вспомогательных лексических еди-

ниц. В работе [18] использовались эксплицитные (союзы и прочие служебные слова) и имплицитные (подразумеваемые, но не выраженные явно) дискурсивные связи.

Параметры удобочитаемости (*readability*) также могут быть использованы в качестве параметров классификации. В их состав входят такие атрибуты, как длина предложения, длина текста в символах, количество единиц определенных частей речи и т.д. Такой формат описания дает краткую характеристику документа [28].

Современные документы, существующие в сети Интернет, обладают еще одним классифицирующим параметром, а именно – HTML-разметкой. Особый способ структурной организации текста позволяет использовать метаинформацию о документе для определения жанра [24, 25].

Более подробно проблема выделения наиболее информативных признаков для классификации текстов описана в статье [3]. Все перечисленные характеристики текста используются в качестве информации для классификаторов.

КЛАССИФИКАТОРЫ ЖАНРА ТЕКСТА

Наивный байесовский классификатор

Метод наивного байесовского классификатора основывается на предположении, что слова независимы друг от друга (появление в тексте слова *A* не влияет на появление в тексте слова *B)*. В этом случае можно вычислить вероятность существования какого-либо списка слов (текста) при условии, что текст относится к определенному классу документов и заданы априорные вероятности появления каждого класса. Максимальная из вычисленных вероятность будет соответствовать классу документа [29].

Классификация более чем 9000 текстов, относящихся к семи различным жанрам – теле- и радионовостям, рекламе, репортажам и т.д. была проведена в работе [30]. Для этого использовались частотности слов, лингвистические параметры текста (временные формы глаголов, синтаксическая сложность), а также комбинация обоих вариантов. Использование наивного баейсовского классификатора при учете частотности слов позволило получить 76,7% точности распознавания, тогда как применение лингвистических параметров в отдельности снизило точность до 33,9%. Это хорошо отражает такие недостатки наивного баейсовского классификатора, как низкая восприимчивость к грамматической информации, в том числе, этот метод не учитывает вероятность появления в документе слов одного семантического поля («убийство» и «преступник» с большей вероятностью окажутся в одном тексте, нежели «любовь» и «пистолет») и тот факт, что вероятность встретить слово в разных местах текста также различается.

Тем не менее, модификация наивного байесовского классификатора, такая как линейный многоклассовый классификатор с отбором признаков, показывает высокие результаты распознавания при классификации научных текстов по отраслям знания [31].

Деревья решений

Дерево решений представляет собой граф или модель решений, учитывающий их возможные исходы и их вероятности. В применении к классификации документов алгоритм дерева решений начинается с выбора разделяющего слова, затем коллекция делится на две части и процедура выполняется заново до тех пор пока все документы коллекции не будут рассортированы. В листьях разрешающего дерева размещаются значения целевой функции, в прочих узлах — условия перехода, определяющие направление движения вдоль ребер дерева. Для классификации каждого примера алгоритму необходимо пройти все дерево от корня до одного из листов и тем самым получить значение целевой функции [32].

Такой подход реализован в статье [33], авторы которой предполагают, что пространство текста частотно (т.е. образовано частотами появления в тексте наборов признаков, к которым относятся служебные слова, биграммы, буквосочетания и т.д.). Полученные данные классификации предполагается использовать для определения метапараметров текста – жанра, автора, стиля и т.д.

Случайный лес

Дальнейшим развитием метода дерева решений является алгоритм «случайный лес» (*random forest*) – ансамблевый метод машинного обучения, который использует совокупность решающих деревьев, построенных независимо друг от друга. Финальная классификация документов проводится с помощью «голосования», т. е. итоговым классом объекта объявляется тот класс, который был решением большинства деревьев [34]. Известной сложностью применения алгоритма «случайный лес» является значительное число решающих деревьев, требующееся для большинства задач, что предъявляет высокие требования к объему памяти.

Оценка качества классификации текстов с помощью алгоритма «случайный лес» была проведена в работе [35]. Классификация материалов, представленных в сети Интернет, имеет практическую ценность как для оценки их содержания, так и для поиска и извлечения специализированной информации (например, научных публикаций по заданной теме). В этом исследовании выполнялась классификация неструктурированной информации по наличию в ней тем, связанных с противозаконной деятельностью. Примененный метод «случайный лес» показал высокую точность при классификации изучаемых данных, что особенно хорошо проявилось при использовании сбалансированных положительных и отрицательных выборок при обучении. Таким образом следует отметить, что алгоритм «случайный лес» склонен к переобучению при неравновесных выборках, что заметно сказывается на точности классификации.

В настоящее время алгоритм «случайный лес» широко используется при решении самых разнообразных задач классификации жанров на корпусах разных языков, примером чего может служить клас-

сификации по жанрам корпуса турецких газет, где с помощью этого алгоритма жанр распознавался правильно в диапазоне от 88% до 93% в зависимости от имеющихся наборов признаков [36].

Метод опорных векторов

Одним из мощных алгоритмов обучения с учителем является метод опорных векторов (*Support Vector Machine* – SVM). Классификация с помощью этого метода происходит благодаря поиску оптимальной разделяющей гиперплоскости в пространстве векторов высокой размерности [37].

Работа [30], посвященная классификации новостных текстов, наравне с наивным байесовским классификатором использует подход SVM, который показывает хороший результат распознавания (82%) даже с простыми признаками (частотности слов), а комбинация частотностей слов с лингвистическими параметрами только улучшает результаты работы классификатора.

ЗАКЛЮЧЕНИЕ

Определение жанра текста – это одна из необходимых задач, решаемых при организации электронных библиотек научных публикаций или художественной литературы. С середины 1990-х гг. до настоящего времени как теоретические основы, так и практическое применение классификации текстов по жанрам неуклонно расширяются. Нами были рассмотрены способы описания текста такими метриками, как «мешок слов», би- и триграммы, TF-IDF, дискурсивные связи, удобочитаемость и HTML-разметка. Применяемые как по отдельности, так и в комбинациях, эти характеристики служат надежной базой для работы машинных классификаторов. В задаче распознавания жанра текста используются алгоритмы различной природы и сложности – от наивного байесовского классификатора и деревьев решений, до методов «случайного леса» и SVM.

СПИСОК ЛИТЕРАТУРЫ

1. Голубева И.Б. Стилистика русского языка. – 2-е изд., испр. – М.: Рольф, 1999. – 448 с.
2. Бахтин М. М. Эстетика словесного творчества. – М.: Искусство, 1986. – 556 с.
3. Мангалова Е.С., Агафонов Е.Д. О проблеме выделения информативных признаков в задаче классификации текстовых документов // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. – 2013. – №1.
4. Скляренко Н.С. Обзор алгоритмов машинного обучения, решающих задачу обнаружения спама // Новые информационные технологии в автоматизированных системах. – 2017. – №20. – С. 26-31.
5. Ehsani R., Muzaffer E.A., Gülsen E. at all. Disambiguating Main POS tags for Turkish // ROCLING. – 2012. – №20. – P. 1121-1128.
6. Stamatatos E., Fakotakis N., Kokkinakis G. Automatic text categorization in terms of genre and author // Computational linguistics. – 2000. – Vol. 26, № 4. – P. 56-63.
7. Al-Harbi S., Almuhareb A., Al-Thubaity A., Khorsheed M., Al-Rajeh A. Automatic Arabic text classification // JADT'08. – 2008. – P. 77–83.
8. Епрев А.С. Автоматическая классификация текстовых документов // Математические структуры и моделирование. – 2010. – №1. – С. 72-76.
9. Agarwal B., Mittal N. Text Classification Using Machine Learning Methods-A Survey // Proceedings of the Second International Conference on Soft Computing for Problem Solving. – 2012. – Vol. 236. – С. 89-95.
10. Sebastiani F. Machine learning in automated text categorization // ACM Comput. – 2002. – Surv. 34(1). – P. 1–47.
11. Kessler B., Nunberg G., Schutze H. Automatic Detection of Text Genre // Computing Research Repository. – 1997. – Vol. 29, № 1. – P. 1224-1229.
12. Karlgren J., Cutting D. Recognizing text genres with simple metrics using discriminant analysis // Proceedings of Coling. – 1994. – Vol. 4, № 3. – P. 12-19.
13. Biber D. The multidimensional approach to linguistic analyses of genre variation: An overview of methodology and finding // Computers in the Humanities. – 1992. – № 3 – С. 58-64.
14. Schütze H. Automatic word sense discrimination // Computational Linguistics. – 1998. – Vol. 24, №1. – С. 29-36.
15. Giesbrecht E., Evert S. Part-of-speech tagging – a solved task? An evaluation of POS taggers for the Web as corpus // Proceedings of the 5th Web as Corpus Workshop (WAC5). – NY: NYPublish, 2009.
16. Rehm G. Towards Automatic Web Genre Identification // Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02). – 2002. – Vol. 4.
17. Boese E.S., Howe A.E. Effects of web document evolution on genre classification // Proceedings of the 14th ACM international conference on Information and knowledge management. – 2005. – Vol.6. – P. 89-95.
18. Webber B. Genre distinctions for Discourse in the Penn TreeBank // Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. – 2009. – P. 674–682.
19. Maas A., Daly R., Pham P. and all. Learning word vectors for sentiment analysis // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. - 2011.
20. Wallach H. Topic modeling: beyond bagofwords // Proceedings of the 23rd International Conference on Machine Learning. – 2006. – P. 977-984.
21. McCallum A., Wangand X., Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval // Proceedings of the Seventh IEEE International Conference on Data Mining. – 2007. – P. 697-702.
22. Radošević D., Dobša J., Mladenić D. at all. Genre Document Classification Using Flexible Length Phrases // Information and Intelligent Systems. – 2006. – Vol. 6 – P . 66-75.

23. Борисов Л.А., Орлов Ю.Н., Осминин К.П. Идентификация автора текста по распределению частот буквосочетаний // Прикладная информатика. – 2013. – Т. 26, № 2. – С. 95-108.
24. Lee Y.-B., Myaeng S.H. Text genre classification with genre-revealing and subjectrevealing features // Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. – 2002. – P. 145–150.
25. Snyman D.P., Van Huyssteen G.B., Daelemans W. Automatic Genre Classification for Resource Scarce Languages // Proceedings of the 22nd Annual Symposium of the Pattern Recognition Association of South Africa. – 2011. – P. 132–137.
26. Biber D. Dimensions of Register Variation: A Cross-Linguistic Comparison. – Cambridge: Cambridge University Press, 1995. – 428 p.
27. Stamatatos E., Fakotakis N., Kokkinakis G. Automatic text categorization in terms of genre and author // Computational linguistics. – 2000. – Vol. 26, № 4. – P. 471–495.
28. Falkenjack J., Santini M., Jonsson A. An Exploratory Study on Genre Classification using Readability Features // The Sixth Swedish Language Technology Conference. – 2016. – Vol. 6. – P. 72-78.
29. Li Y.H., Jain A.K. Classification of text documents // Computer Journal. – 1998. – № 41(8). – P. 537-46.
30. Dewdney N., Van Ess-Dykema C., MacMillan R. The form is the substance: classification of genres in text // Proceedings of the workshop on Human Language Technology and Knowledge Management. – 2001. – Vol. 7. – P. 56-65.
31. Сухарева А.В., Царьков С.В. Классификация научных текстов по отраслям знаний // Машинное обучение и анализ данных. – 2014. – Т. 1, № 8. – С. 22-25.
32. Quinlan J.R. Simplifying decision trees // International Journal of Man-Machine Studies. – 1987. – Vol.7 – P. 152-160.
33. Кубарев А.И., Кукушкина О.В., Поддубный В.В. и др. Построение таблиц стилей текстовых произведений с использованием алгоритмов классификации на основе деревьев решений // Вестн. Томск. гос. ун-та. Сер. Управление, вычислительная техника и информатика. – 2012. – № 4. – С. 79–88.
34. Kam T. Random Decision Forests // Proceedings of the 3rd International Conference on Document Analysis and Recognition. – 1995. – P. 278–282.
35. Веретенников И., Карташев Е., Царегородцев А. Оценка качества классификации текстовых материалов с использованием алгоритма машинного обучения «Случайный лес» // Известия АлтГУ. – 2017. – №4 (96). – С. 25-36.
36. Amasyali F. M., Banu D. Automatic Turkish Text Categorization in Terms of Author, Genre and Gender. – Springer-Verlag Berlin Heidelberg, 2006.
37. Cortes C., Vapnik V. Support vector networks // Machine Learning. – 1995. – Vol. 20. – P. 237–297.

Материал поступил в редакцию 16.03.18.

Сведения об авторах

БУЙЛОВА Надежда Николаевна – преподаватель Школы лингвистики Факультета гуманитарных наук Национального исследовательского университета – Высшая Школа Экономики, Москва
e-mail: nbujlova@hse.ru