УДК 004.912

В.А. ЯЦКО
(Абакан)

Y-МЕТОД КЛАССИФИКАЦИИ ТЕКСТОВ*

Рассматриваются основные особенности автоматической классификации текстовых документов. Описываются процедуры нового метода, основанного на вычислении отклонений распределения стоп-слов от коэффициента Ципфа: распознавание стоп-слов и составление ранжированных списков; вычисление отклонений частотностей терминов от коэффициента Ципфа; вычисление индексов текстов на основе среднего квадратичного отклонения; определение степени близости текстов. Разработаны показатели эффективности классификации: дискриминирующей силы, симилирующей силы и обобщённый показатель. Тестирование метода показало его эффективность при решении задачи жанровой классификации текстов.

Ключевые слова: *автоматическая классификация текстовых документов, методы и алгоритмы, распределение Ципфа, показатели эффективности, дискриминирующая сила, жанровая классификация, степень близости текстов.*

VIATCHESLAV YATSKO
(Abakan)

Y-METHOD OF TEXT CLASSIFICATION

The article deals with the specific features of the automatic text classification. There are described the procedures of a new classification method based on the calculation of the deviations of stop-words distribution from Zipfian score: the recognition of stop-words and the creation of ranked lists; the calculation of deviations of terms frequencies from Zipfian score; the calculation of documents indices basing on standard deviation; finding the degree of document similarity. The author introduces the indicators of classification efficiency, such as the discriminative power, the similarative power and the generalized index. The method was tested and proved to be efficient for the solution of the genre classification task.

Key words: *automatic text document classification, methods and algorithms, Zipf distribution, efficiency indices, discriminative power, genre classification, degree of text documents similarity.*

Классификация текстовых документов – интенсивно развивающееся направление информационных технологий, которое имеет непосредственное значение для адекватного функционирования информационно-поисковых систем, электронных библиотек, систем анализа мнений пользователей, программ фильтрации спама, систем распознавания плагиата, авторской атрибуции текстов [6]. В информационно-поисковых системах, выполняется тематическая категоризация баз данных и запросов пользователей, позволяющая существенно сокращать время поиска, т. к. распределение терминов запроса анализируется не по всем документам, а только по тем, которые включены в определённый тематический раздел базы данных (например, *спорт, путешествия, развлечения*). Жанровая классификация критически важна для электронных библиотек, поскольку читателей интересуют произведения, относящиеся к определённому жанру (например, *детектив, фантастика, любовный роман*). Анализ мнений пользователей основывается на классификации оценок и текстов на положительные, отрицательные, нейтральные. Фильтрация спама и распознавание плагиата предусматривают бинарную классификацию на «спам» и «не спам», «плагиат» и «не плагиат». Авторская атрибуция предполагает определение авторства, т. е. соотнесение анализируемого текста с классом, который включает работы данного автора.

Широкое распространение классификационных технологий делает актуальным разработку новых методов автоматической классификации текстовых документов. Цель настоящей статьи – описать особенности применения предлагаемого нами Y-метода, основы которого были изложены в предыдущей работе [1].

* Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта № 20-07-00124.

Автоматическая классификация выполняется программой-классификатором, выдающей пользователю имя класса, с которым соотносится загружаемый им входной текст. С целью определения класса текста программа выполняет анализ распределения единиц текста и определяет степень близости между входным текстом и эталонным текстом (корпусом текстов), который содержит документы, наиболее типичные для данного класса [5]. Разработка классификатора предусматривает следующие основные этапы. 1) Идентификация единиц текста (терминов) в процессе его предварительной обработки. В качестве терминов могут выступать отдельные слова, основы слов, n-грамы, а также их лексико-грамматические характеристики, например, категории частей речи. 2) Анализ распределения терминов, предусматривающее их взвешивание. В результате взвешивания терминам присваиваются числовые коэффициенты, отражающие их дискриминативную силу – способность идентифицировать данный класс текстовых документов [9]. Такими коэффициентами могут быть частотности терминов, либо нормализованные величины. На основе коэффициентов отдельных терминов вычисляется классификационный индекс всего текстового документа. 3) Сопоставление индексов документов и определение степени их близости. Полученные классификационные индексы текстов сопоставляются с индексом эталонного текста/корпуса, или некоторым пороговым уровнем. Если индекс входного текста превышает заданный пороговый уровень, то он соотносится с классом, представленным эталонным текстом. 4) Тестирование эффективности классификатора. Тестирование проводится на текстах, класс которых уже известен. Правильное определение класса свидетельствует об эффективности классификатора.

Применение Y-метода включает следующие процедуры.

1) Распознавание стоп-слов. К ним относятся артикли, предлоги, частицы, местоимения, союзы и союзные слова, которые часто встречаются в текстах различных типов и жанров. Взятые вне контекста их использования, эти слова не выражают значения. Анализ распределения стоп-слов составляет основу Y-метода. С целью их распознавания мы применяем дополненный список К. Фокса, который включает 426 единиц [8]. Словоформы из этого списка сопоставляются с каждым из анализируемых текстов, и стоп-слова находятся по точному совпадению.

2) Получение ранжированных списков стоп-слов для каждого из текстов. Под ранжированным списком нами понимается список, отсортированный по нисходящей по частотностям и по восходящей по рангам. Слово с первым рангом имеет самую высокую частотность. В большинстве текстов на английском языке таким словом будет определённый артикль *the.*

3) Нахождение итеративного порогового уровня. Под этим уровнем понимается первый повтор частотности. В результате составляется сокращённый ранжированный список, в который включаются стоп-слова с рангами и частотностями, занимающие в списке позиции до первых слов с повторяющимися частотностями. Таким образом, в окончательный список входят только наиболее частотные слова с уникальными значениями.

4) Вычисление коэффициента Ципфа для каждого стоп-слова по формуле:

$$Z(w_{ij}) = F(w1_j)/R(w_{ij}), \quad (1)$$

где $F(w1_j)$ – частотность первого по рангу слова в некотором *j*-м тексте, а *R* – номер ранга слова. Коэффициент Ципфа представляет собой идеальную величину, которая задаётся законом Ципфа [4].

5) Вычисление отклонения от коэффициента Ципфа для каждого стоп-слова по формуле:

$$DevZ\ (w_{ij}) = |Z(w_{ij}) - F(w_{ij})|, \quad (2)$$

Таким образом, создаётся четыре числовых ряда: ряд *R* с рангами слов; ряд *F*, содержащий реальные частотности слов; ряд Z с коэффициентами Ципфа; ряд DevZ, включающий разницы по модулю между двумя показателями.

Как правило, реальные частотности слов больше коэффициента Ципфа. В основе Y-метода лежит предположение о том, что отклонение частотностей стоп-слов от распределения Ципфа специфично

для каждого текста, и у текстов, относящихся к одному классу, разница отклонений будет меньшей, чем у текстов, относящихся к разным классам.

6) Получение классификационных индексов. Для последнего числового ряда DevZ, в котором указываются разницы между реальными частотностями и коэффициентами Ципфа, находится среднее квадратичное отклонение σ по формуле:

$$\sigma(\mathrm{DevZ_j}) = \sqrt{Var\,(DevZ_j)}, \quad (3)$$

где *Var* – дисперсия, а *DevZ* – указанный числовой ряд. Полученное отклонение есть индекс текста, на основе которого вычисляются расстояния между входными и эталонным текстом.

7) Процедура классификации. Классификация проводится на основе вычисления расстояний между входными и эталонным текстом по формулам:

$$\mathrm{Dis(T1, R)} = |\,\sigma(\mathrm{DevZ}_{T1}) - \sigma(\mathrm{DevZ}_R)\,| \quad (4)$$

$$\mathrm{Dis(T2, R)} = |\,\sigma(\mathrm{DevZ}_{T2}) - \sigma(\mathrm{DevZ}_R)\,|, \quad (5)$$

где T1 и T2 – входные тексты, а R – эталонный. Меньшее расстояние указывает на то, что данный входной относится к классу, представленному эталонным текстом.

8) Вычисление разницы расстояний в процентах и нахождение коэффициента дискриминирующей силы: от большей величины отнимается меньшая и получившаяся разница делится на большую величину. Так находится на сколько расстояние между одним входным текстом и эталонным меньше, чем расстояние между другим входным текстом и эталонным в процентах. Например, если величина Dis(T1, R) < Dis(T2, R) (т. е. T1 относится к тому же классу, что и R), то дискриминирующая сила будет равна:

$$\mathrm{DP=(Dis(T2, R) - Dis(T1, R)) / (Dis(T2, R) * 100}\ . \quad (6)$$

Под дискриминирующей силой нами понимается коэффициент, указывающий на разницу между текстом, относящимся к данному классу, представленному эталонным текстом, и текстом, не относящимся к этому классу.

9) Нахождение коэффициента симилирующей силы в процентах по формуле:

$$SP = \frac{\sigma(\mathrm{DevZ}_{T1})}{\sigma(\mathrm{DevZ}_R)} * 100, \quad (7)$$

где в числителе меньшая величина. Коэффициент симилирующей силы определяется по соотношению индексов эталонного текста и ближайшего к нему входного текста, обозначенного в формуле как *T1*. Под ближайшим текстом понимается текст с коэффициентом σ(DevZ), наиболее близким к коэффициенту эталонного текста. Таким образом, под симилирующей силой понимается степень близости (англ. *similarity*) между эталонным текстом и соответствующим входным текстом.

10) Нахождение общего коэффициента эффективности (Q) предлагаемого метода по формуле:

$$Q = \frac{DP+SP}{2}$$

Стандартным материалом для тестирования методов автоматической классификации считается корпус *Reuters [7],* содержащий небольшие по размеру новостные тексты финансово-экономической тематики, распределённые по таким категориям, как *trade, grain, gold, jobs*. Для тестирования мы отобрали 453 файла из папки *Test* корпуса (файлы 20738–21576). Эти файлы были объединены в один, который стал первым входным файлом (*T*1). Остальные 2566 файлов составили эталонный файл (*Ref*). В качестве второго входного текста (*T*2) был произвольно выбран художественный текст *The Hound of the Baskervilles*, который по размеру примерно соответствовал первому входному тексту *T*1. Текст (известное произведение Конан Дойла) был загружен с сайта Gutenberg [4], на котором размещаются вычитанные и отредактированные произведения с истекшим авторским правом. Три файла

были выровнены по размеру методом выравнивания по нижнему пределу, который предусматривает удаление части больших по размеру текстов (в нашем случае – *T*1 и *Ref*), с тем чтобы количество токенов в этих текстах стало примерно таким же, как и в меньшем по размеру тексте (в нашем случае – *T*2). В каждом файле были найдены стоп-слова с помощью конкорданса AntConc [2]. Получившиеся списки были сокращены с учётом итеративного порогового уровня. Далее, в соответствии с описанной выше методикой, были вычислены коэффициенты Ципфа, отклонения от распределения Ципфа, и получены индексы текстов на основе среднего квадратичного отклонения. В табл. 1 представлены статистические данные выравненных текстов.

Таблица 1

Статистические данные текстов

Текст	Количество токенов	Количество уникальных слов	Количество стоп-слов	Количество стоп-слов с уникальными значениями
*T*1 (Reuters)	60084	5023	342	28
*T*2 (The Hound)	59933	5536	368	43
Ref (Reuters)	60010	6206	358	41

Мы исходили из предположения о том, что степень близости между текстами *T*1 и *Ref* должна быть существенно больше, чем между текстами *T*2 и *Ref*, поскольку первые два текста относятся к одному жанру и характеризуются сходной тематикой, в то время как *T*2 – художественный текст, написанный в начале 20 века. В терминах нашего метода расстояние между *T*1 и *Ref* должно быть существенно меньше, чем расстояние между *T*2 и *Ref*. В табл. 2 приводятся данные, подтверждающие наше предположение. Расстояние между *T*1 и *Ref* на 59,59% меньше, чем расстояние между *T*2 и *Ref* (числа округлены до сотых).

Таблица 2

Результаты классификации текстов

Тексты	σ-индекс	Расстояние		DP	SP	Q
*T*1	177,52	Dis (*T*1, *Ref*)	Dis (*T*2, *Ref*)	59,59%	92,92%	76,26%
*T*2	196,05	12,56	31,09			
Ref	164,96					

В настоящей работе описаны основные процедуры разработанного нами метода автоматической классификации текстов, который мы называем Y-методом. Можно выделить следующие особенности предлагаемого метода. В отличие от существующих методов в нашем методе применяется только один параметр: отклонения частотности стоп-слов от распределения Ципфа, сопряжённого со средним квадратичным отклонением. Это позволило существенно упростить процесс вычислений, а использование стоп-слов – повысить его быстродействие, причём итеративный пороговый уровень позволил сократить количество обрабатываемых терминов до нескольких десятков. Эффективность методов автоматической обработки данных подразумевает не только высокое быстродействие, но и качество результата. Y-метод продемонстрировал адекватные показатели дискриминативной и симилирятивной силы для решения задачи жанровой классификации газетных текстов. Ранее [1] нами была установлена его эффективность для решения задачи авторской атрибуции художественных текстов.

Вместе с тем, следует обратить внимание на проблемы, с которыми сталкивается применение предлагаемого метода. Во-первых, выравнивание по нижнему пределу (англ. undersampling), которое применяется в медицине, социологии, биологии и других областях. В компьютерной лингвистике оно до сих пор не применялось, и его влияние на структуру текста требует дополнительных исследований. Во-вторых, зависимость результатов анализа от объёма текстов. Закон Ципфа выполняется на текстах больших объёмов, в частности установлено, что распределение слов в Брауновском корпусе (объём – 1 миллион токенов) соответствует распределению Ципфа. Для тестирования нами брались также достаточно объёмные тексты, содержащие десятки тысяч токенов, и не вполне понятно, насколько эффективным будет применение предлагаемого метода для текстов меньших по размеру. Решение этих проблем – задача последующих исследований.

Литература

1. Яцко В.А. Метод автоматической классификации текстов, основанный на законе Ципфа // Научно-техническая информация. Сер. 2. Информационные процессы и системы. 2015. № 5. С. 19–24.

2. Anthony L. AntConc 3.5.8. – 2019. [Электронный ресурс]. URL: https://www.laurenceanthony.net/software/antconc/ (дата обращения: 10.06.2021).

3. Corral A., Serra I. The brevity law as a scaling law, and a possible origin of Zipf's law for word frequencies // Entropy. 2020. Vol. 22. No. 2. [Электронный ресурс]. URL: https://www.mdpi.com/1099-4300/22/2/224/htm (дата обращения: 10.06.2021).

4. Free eBooks – Project Gutenberg. 2020. [Электронный ресурс]. URL: https://www.gutenberg.org/ (дата обращения: 10.06.2021).

5. Kowsari D. et al. Text classification algorithms: A survey // Information. 2019. Vol. 10. No. 4. [Электронный ресурс]. URL: https://www.mdpi.com/2078-2489/10/4/150/htm (дата обращения: 10.06.2021).

6. Nidhi, Gupta V. Recent trends in text classification techniques // International journal of computer applications. 2011. Vol. 35. No. 6. P. 45–51. [Электронный ресурс]. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.736.3034&rep=rep1&type=pdf (дата обращения: 10.06.2021).

7. Reuters-21578 benchmark corpus. – 2017. [Электронный ресурс]. URL: https://www.kaggle.com/nltkdata/reuters (дата обращения: 10.06.2021).

8. Yatsko V.A. TF*IDF ranker. – 2021. [Электронный ресурс]. URL: http://yatsko.zohosites.com/tf-idf-ranker1.html (дата обращения: 10.06.2021).

9. Zong W. et al. A discriminative and semantic feature selection method for text categorization // International journal of production economics. 2015. Vol. 165. P. 215–222. [Электронный ресурс]. URL: https://www.sciencedirect.com/science/article/abs/pii/S0925527314004290 (дата обращения: 10.06.2021).