УДК 004.415.2+ 81.322.2 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait.2020.2/2924
Поступила в редакцию 25.03.2020
Подписана в печать 15.06.2020

АНАЛИЗ ИНФОРМАЦИОННЫХ КРИТЕРИЕВ ОТБОРА ЗНАЧИМЫХ ПРИЗНАКОВ В МЕТОДАХ TEXT MINING

© 2020 А. Л. Калабин✉, Е. И. Корнеева

Тверской государственный технический университет
наб. Афанасия Никитина, 22, 170026 Тверь, Российская Федерация

Аннотация. В работе проведена количественная и качественная оценка методов отбора признаков документов на основе теории информации. Целью исследования являлась проверка применения ряда критериев для редуцирования множества терминов в коллекции текстов, к которой впоследствии будут применены методы классификации с учителем и без учителя. Входные данные программной реализации были разделены по схожести тематик и, в зависимости от эксперимента, включали наборы из 45 документов трех категорий технических текстов в различных концентрациях. Для расчета критериев использовалась программная система анализа текстовых данных TextStageProcessor, расположенная как проект с открытым исходным кодом. В разделе оценки работоспособности критериев введены две величины. Первая определяет относительное количество документов, которые принадлежат категории и содержат термин. Вторая равна относительному количеству документов, принадлежащих категории и не содержащих термин. Построены графики зависимости упомянутых величин от критериев. Рассмотрены ограничения для указанных параметров. Полученные результаты для критериев *MI*, *CHI*, *IG* не монотонны, что свидетельствует о возможной неработоспособности этих критериев для входной коллекции и необходимости дальнейших исследований. Для второй части эксперимента проведена предварительная обработка текстов, включающая удаление стоп-слов, нормализацию термов и приведение их к нижнему регистру. Качественный вид графиков зависимостей критериев *TFD*, *DF* и *TF·IDF* от ранга слова в коллекции свидетельствует о том, что с их помощью можно сократить множество входных значимых термов для классификации без потери качества для исследования.
Ключевые слова: анализ текстовых данных, методы отбора значимых признаков, частота повторения термина, коллекция документов, оценка критериев.

ВВЕДЕНИЕ

Категория, к которой принадлежит документ, в задаче Text Mining может быть определена по набору терминов (признаков) в тексте документа. Выборка входных данных включает множество так называемых «шумовых» признаков, которые обладают слабой классификационной способностью. В частности, с избытком терминов во входном текстовом наборе связаны высокие вычислительные затраты при получении мер близости документов и низкое качество результата при отнесении документа к некоторой категории. Отсюда возникает необходимость создания редуцированного множества терминов для классификации T', которое будет включать наиболее информативные в некотором смысле признаки из входного множества T, так что $|T'| \ll |T|$ [1–5]. Для этого предлагаются критерии оценки на основе теории информации: взаимная информация (MI), информационная выгода (IG), мера хи-квадрат (CHI), мера документной частоты (DF), а также меры, основанные на оценке частоты повторения терминов в документе ($TF \cdot IDF$, TFD). Однако применимость указанных критериев для от-

✉ Калабин Александр Леонидович
e-mail: akalabin@yandex.ru

сечения «шумовых» признаков требует проверки, что и является целью данного исследования — экспериментальная качественная и количественная оценка работоспособности методов отбора признаков документов в Text Mining.

Для расчета критериев использовалась программная система анализа текстовых данных, разработанная на кафедре Программного обеспечения ТвГТУ [6] и расположенная как проект с открытым исходным кодом в веб-сервисе по адресу: https://github.com/mhyhre/TextStageProcessor.

1. МАТЕРИАЛЫ И МЕТОДЫ

1.1. Оценка работоспособности методов MI, IG, CHI

Наиболее распространенными методами выбора терминов для классификации [1–5] являются меры MI, IG, CHI, DF. Вышеперечисленные критерии сравниваются в работе [3], в которой сообщается, что IG и CHI являются наиболее эффективными в выборе признаков. Рассмотрим подход к выделению значимых термов с помощью расчета критериев отбора признаков документов (табл. 1), предлагаемый в [3].

Поясним использованные в табл. 1 обозначения для расчета критериев (табл. 2).

Пусть дана коллекция документов $D = \{d_1, d_2, \ldots, d_{|D|}\}$, $i = \overline{1, |D|}$, которая описана с помощью набора проиндексированных терминов $T = \{t_1, t_2, \ldots, t_{|T|}\}$. Тогда размерность коллекции $|D| = N$, где N — общее количество документов, а $T = \{t_k \in T : DF(t_k) > t\}$, где $DF(t_k)$ — это количество документов, в которых встречается термин t_k.

Существует множество тематических категорий $C = \{c_1, c_2, \ldots, c_{|C|}\}$, $j = \overline{1, |C|}$, которым принадлежат документы коллекции. Для оценки работоспособности критериев определим $|C| = 2$, т. е. рассмотрим для простоты только две категории документов. Введем величины:

$$K_1 = A/\Omega \quad (1)$$

Таблица 1. Общие обозначения для расчета критериев отбора значимых признаков [Table 1. Generic notations for the calculation of selection criteria of relevant features]

Обозначение	Расшифровка
Ω	обучающее множество документов
c	категория документа
t	термин в словаре документа
T	набора проиндексированных терминов коллекции документов
A	количество документов обучающего множества, которое принадлежит категории c и содержит термин t
B	количество документов обучающего множества, которое не принадлежит категории c и содержит термин t
C	количество документов обучающего множества, которые принадлежат категории c и не содержат термин t
D	количество документов обучающего множества, которые не принадлежат категории c и не содержат термин t

$$K_2 = C/\Omega \quad (2)$$

K_1 равную относительному количеству документов, которые принадлежат категории c_1 и содержат термин $t(A)$ (1), и K_2 равную относительному количеству документов, которые принадлежат категории c_1 и не содержат термин $t(C)$ (2). Для указанных параметров (табл. 2) выполняются следующие ограничения (3–6).

$$\Omega = A + B + C + D \quad (3)$$

$$\Omega < N \quad (4)$$

$$|C_1| = A + B \quad (5)$$

$$|C_2| = C + D \quad (6)$$

В ограничении (4) за N принимается общее количество документов.

Таким образом, термин t и отношение K_1 прямо пропорционально зависят друг от дру-

Таблица 2. Критерии отбора значимых признаков
[Table 2. Selection criteria of relevant features]

Критерий	Формула
MI (взаимной информации)	$MI(t_k, c_j) = \log_2 \frac{A \cdot \lvert\Omega\rvert}{(A+C)\cdot(A+B)}$
IG (информационной выгоды)	$IG(t_k, c_j) = \frac{A}{\lvert\Omega\rvert} \cdot \log_2 \frac{A \cdot \lvert\Omega\rvert}{(A+C)\cdot(A+B)} \cdot \frac{C}{\lvert\Omega\rvert} \cdot \log_2 \frac{C \cdot \lvert\Omega\rvert}{(A+C)\cdot(C+D)} \times$ $\times \frac{B}{\lvert\Omega\rvert} \cdot \log_2 \frac{B \cdot \lvert\Omega\rvert}{(B+D)\cdot(A+B)} \cdot \frac{D}{\lvert\Omega\rvert} \cdot \log_2 \frac{D \cdot \lvert\Omega\rvert}{(D+C)\cdot(D+B)}$
CHI (ХИ-критерий)	$CHI(t_k, c_j) = \frac{\lvert\Omega\rvert \cdot (A \cdot D - C \cdot B)^2}{(A+C)\cdot(B+D)\cdot(A+B)\cdot(C+D)}$
DF (документной частоты)	$T = \{t_k \in T : DF(t_k) > t\}$

га — чем больше K_1, тем больше признак t характеризует категорию c. Вычислим критерии по формулам (табл. 1) и представим результаты на графиках зависимости критериев MI, IG, CHI, DF от отношения K_1 на рис. 1а–1г. При вычислении значение Ω было принято как 70 % от общего числа документов.

Аналогично сделаем и для K_2 — предположим, что чем больше параметр K_2, тем меньше признак t характеризует категорию c. Результат для IG представим на рис. 2. График IG(K1) и остальные три зависимости зеркален относительно зависимости относительно K_1 т. к. $K_1 + K_1 = \Omega$.

1.2. Оценка работоспособности методов DF, IDF, TF·IDF

Мера инверсной документной частоты IDF [3, 7–12] рассматривает количество повторений τ определенного термина в коллекции документов после предварительной обработки и в зависимости от того, в скольких текстах встречается слово, понижает или повышает его значимость. IDF вычисляется как логарифм отношения числа всех документов ($\lvert D\rvert$) к числу документов, которые содержат некоторое слово ($\lvert\{d_i \in D \mid t \in d_i\}\rvert$, при $n_t \neq 0$) (7). Если термин встречается во многих документах набора, то критерий будет близок к 0, если во всех документах — равен 0.

$$idf(t, D) = \log \frac{\lvert D\rvert}{\lvert\{d_i \in D \mid t \in d_i\}\rvert} \qquad (7)$$

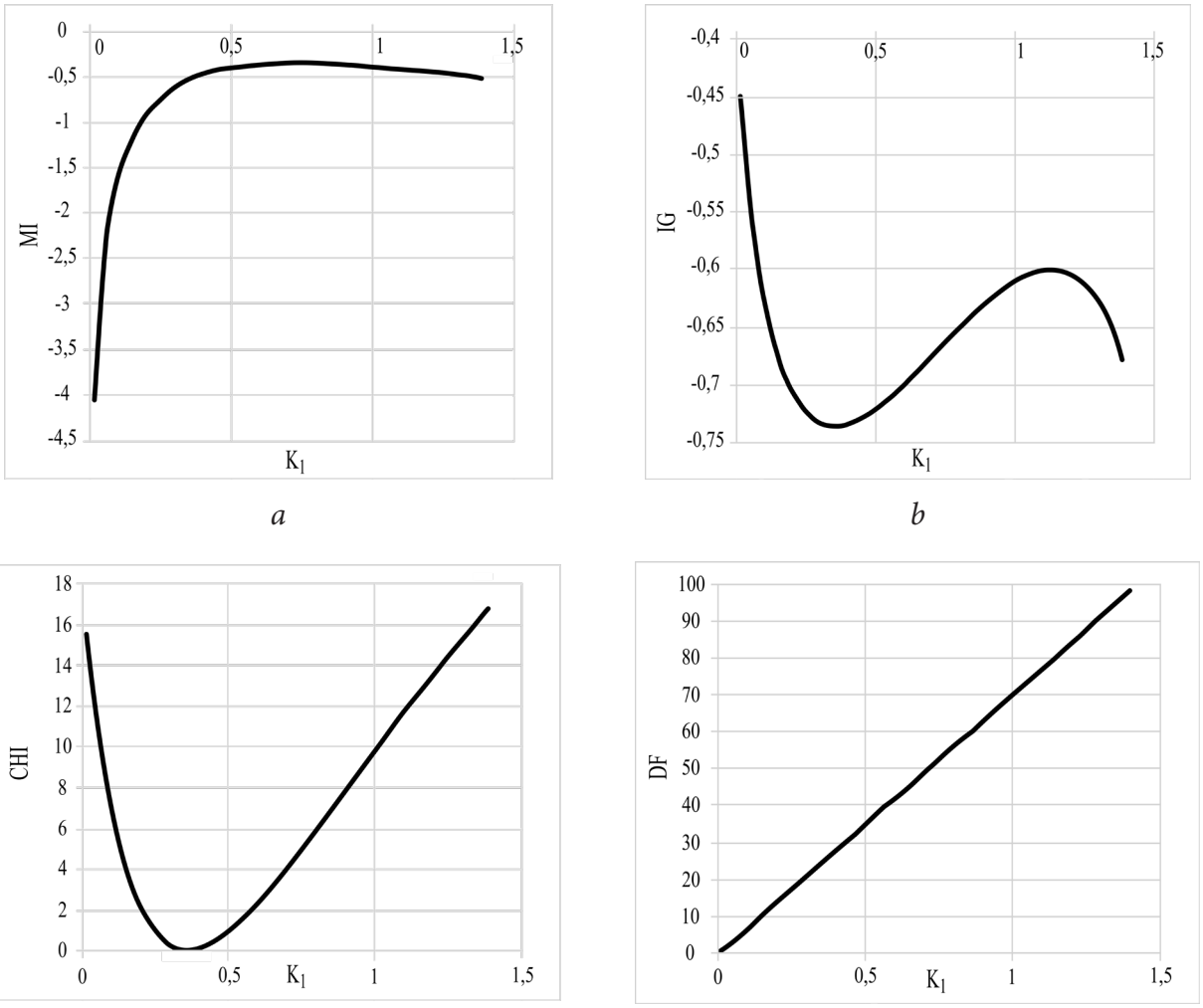
Чтобы сбалансировать показатели частотности повторения терминов для маленьких и больших текстов входной выборки, применяется параметр *TF* или отношение частоты повторения слова в документе (n_t) к общему числу слов в документе ($\Sigma_k n_k$).

$$tf(t, d) = \frac{n_t}{\Sigma_k n_k} \qquad (8)$$

Критерий TF·IDF является методикой оценки степени важности признака для определенного набора документов или его части (категории), произведением коэффициентов TF и IDF (9). В частности, TF·IDF используется как функция определения веса термина для построения векторной модели некоторого документа [3].

$$tf_{idf}(t, d, D) = tf(t, d) \cdot idf(t, D) \qquad (9)$$

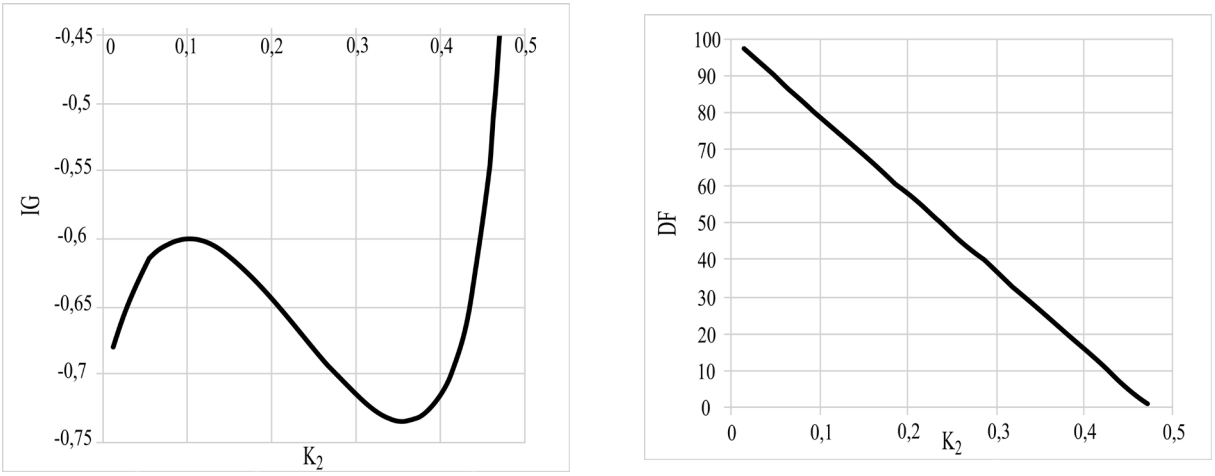
Приведем результаты, полученные расчетным методом. Тестовая выборка данных для исследования состоит из 45 текстов на русском языке предметной области «Составляющие системного блока компьютера», подобранных с сайтов http://www.ferra.ru/, http://www.ixbt.com/. Коллекция разбита на три категории, в соответствии с тематикой статьи — накопитель на жестких магнитных дисках (HDD), твердотельный накопитель



a *b*

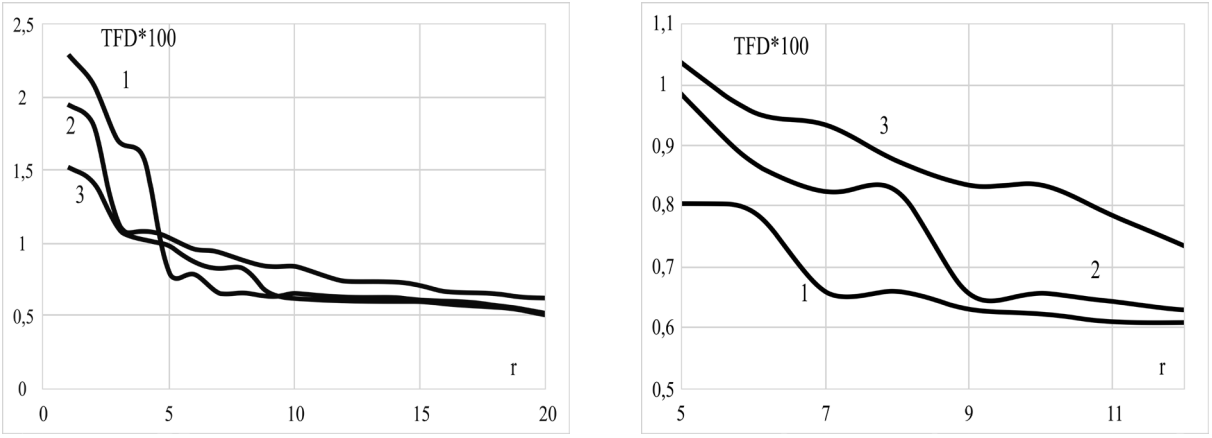
c *d*

Рис. 1. Графики зависимости отношения K_1 от критериев: a — MI, b — IG, c — CHI, d — DF
[Fig. 1. Graphs of the dependence of the ratio K_1 on the criteria: a — MI, b — IG, c — CHI, d — DF]



a *b*

Рис. 2. Графики зависимости отношения K_2 от критерия: a — IG, b — DF
[Fig. 2. Graphs of the dependence of the ratio K_2 on the criteria: a — IG, b — DF]



a *b*

Рис. 3. Зависимости TFD (частота термина в документе) меры от ранга значимых слов для тематики: 1 — HDD; 2 — SSD; 3 — CPU
[Fig. 3. Dependences of the TFD (the term frequency in the document) measure on the rank of the relevant words for the topic: 1 — HDD; 2 — SSD; 3 — CPU]

(SSD), центральное процессорное устройство (CPU). Проанализировано две коллекции – различные варианты сочетания категорий документов. В первом случае тестировались схожие между собой документы категорий (Case01) — HDD&SSD; т. к. у них одинаковые функции и характеристики. Во втором случае документы выбраны с большим различием категорий (Case02) — CPU&HDD, у которых функции, характеристики и устройство существенно отличны.

Прежде чем рассчитать указанные выше критерии, необходимо предварительно обработать тексты предметной области, используя библиотеку pyMorphy2. Препроцессинг проводится в несколько этапов и заключается в удалении неинформативных слов и повышения строгости текстов следующими методами: 1) исключение из выборки стоп-слов, загружаемых из текстового файла; 2) приведение к базовой (нормальной) форме с использованием стемминга и разбиения предложений на N-граммы; 3) проведение конвертации всех термов к нижнему регистру. Перечисленные этапы применены для эффективной очистки текстов. Приведем численные значения размера входных коллекций данных до и после предварительной обработки в табл. 3.

Для сокращения числа терминов множества T при сохранении наиболее информативных в некотором смысле терминов предлагается мера TFD, которая измеряется отношением частоты повторения слова в документах одной категории (n_k) к общему числу слов во всех документах этой категории (10).

$$tfd(t,d) = \frac{n_k}{\sum_k n_k} \qquad (10)$$

Предполагается, что указанная в (10) мера будет иметь удовлетворительную классификационную способность для определения от-

Таблица 3. Размер коллекции термов до предварительной обработки и после
[Table 3. The size of the collection of terms before and after the preprocessing]

Эксперимент	Рассматриваемые категории	Количество документов	Количество термов без предобработки текстов	Количество термов после предобработки текстов
Case01	HDD и SSD	30	8201	3314
Case02	CPU и HDD	30	9700	3855

дельных категорий. Зависимость TFD от ранга r имеет качественный вид для всех трех категорий документов HDD, SSD и CPU (рис. 3а) и хорошо аппроксимируется степенной функцией $f(r) = ar^{-b}$, т. к. коэффициент достоверности аппроксимации для всех трех функций находится в диапазоне от $0,94$ до $0,97$ (табл. 4). Этот вид зависимости приближенно соответствует закону Зипфа [13].

Значения коэффициентов $TFD(r)$ для категорий HDD, SSD и CPU (рис. 3б) весьма существенно отличаются для термов с величиной ранга r от 5 до 11 и эта разница составляет 20–30%, что говорит о возможности использовать меру для классификации категорий.

Для указанной тестовой выборки программной системой рассчитаны критерии $TF \cdot IDF(r)$ и $DF(r)$. Приведем на рис. 4а и рис. 4б графики зависимостей указанных критериев от ранга значимых слов r для первых 40 терминов коллекции.

2. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Предварительная гипотеза о том, что для значимого в коллекции документов категории термина график отношения K_1 должен расти и график отношения K_2 убывать не подтверждается для критериев *MI*, *IG* и *CHI*. Эти зависимости не монотонны и имеют экстремумы, что свидетельствует об их возможной неработоспособности и необходимости дальнейших исследований. Критерий документной частоты *DF* верно отражает прямо пропорциональную зависимость термина t от отношения K_1 (рис. 1г) и от отношения K_2 (рис. 2), что очевидно, согласно его определению.

Зависимость $DF(r)$ имеет одинаковый качественный вид для всех категорий документов и хорошо аппроксимируется линейной функцией, т. к. коэффициент достоверности аппроксимации для всех четырех функций находится в диапазоне от $0,97$ до $0,99$. Критерий *DF* может быть использован

Таблица 4. Аппроксимирующие функциональные зависимости для $TFD(r)$ меры
[Table 4. Approximation functional relationships for the $TFD(r)$ measure]

	CPU	HDD	SSD
$f(r)$	$2,27x^{-0,49}$	$2,87x^{-0,51}$	$2,34x^{-0,48}$
R^2	0,94	0,99	0,97

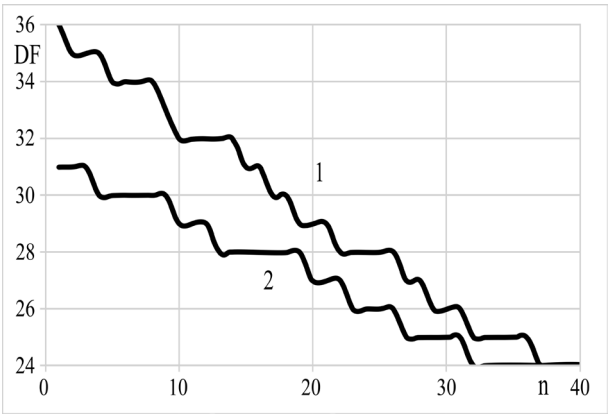


Рис. 4а. Зависимость DF меры от ранга значимых слов для случаев:
1 — HDD&SSD; 2 — HDD&CPU
[Fig. 4a. Dependence of the DF measure on the rank of the relevant words for the cases:
1 — HDD & SSD; 2 — HDD & CPU]

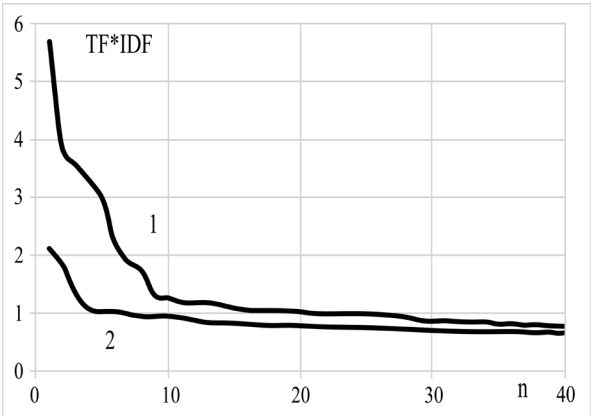


Рис. 4b. Зависимость TF·IDF меры от ранга значимых слов для случаев:
1 — HDD&SSD; 2 — HDD&CPU
[Fig. 4b. Dependence of the TF·IDF measure on the rank of the relevant words for the cases:
1 — HDD & SSD; 2 — HDD & CPU]

как вспомогательный, для отсечения незначимых слов. Отличие для двух случаев Case01 и Case02 максимально при $n<10$, при этом среднее относительное отклонение не менее 12 %, что свидетельствует о возможности определения отличия категории друг от друга по этому критерию.

График зависимости для меры $TF \cdot IDF(r)$ имеет качественный вид для всех категорий документов и хорошо аппроксимируется степенной функцией $f(r)=ar^{-b}$, т. к. коэффициент достоверности аппроксимации во всех экспериментах превышает значение $0,95$. Таким образом, критерий $TF \cdot IDF$ аналогично критерию DF дает возможность выделить особенности категорий, т. к. отличие параметра для термов при $n<10$ максимально и среднее относительное отклонение не менее 10 %.

ЗАКЛЮЧЕНИЕ

Проведена экспериментальная качественная и количественная оценка работоспособности методов отбора признаков документов в Text Mining.

Полученные результаты для критериев зависимости MI, CHI, IG от доли документов, которые принадлежат определенной категории и содержат термин, характеризующий эту категорию, не монотонны, что свидетельствует о возможной неработоспособности этих критериев и необходимости дальнейших исследований.

Критерий документной частоты DF является мерой при отнесении документа к некоторой категории. Зависимость $DF(r)$ от ранга имеет одинаковый качественный вид для всех категорий документов и хорошо аппроксимируется линейной функцией и работоспособен для отсечения незначимых слов.

График зависимости для меры $TF * IDF(r)$ имеет одинаковый качественный вид для всех категорий документов и хорошо аппроксимируется степенной функцией $f(r)=ar^{-b}$ и дает возможность выделить особенности категорий.

Полученные результаты свидетельствуют о работоспособности критериев TFD, DF и $TF \cdot IDF$ для сокращения числа терминов множества.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Yang, Y.* A Comparative Study on Feature Selection in Text Categorization / Y. Yang, J. O. Pedersen // Proceedings of the 14th International Conference on Machine Learning (Nashville, 8–12 July 1997). – Nashville, 1997. – P. 412–420.

2. *Meng, J.* A Two-stage feature selection method for text categorization / J. Meng, H. Lin, Y. Yu // Computers and Mathematics with Applications (October 2011) – V. 62, iss. 7, 2011. – P. 2793–2800. – DOI: 10.1016/j.camwa.2011.07.045.

3. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е. И. Большакова [и др.]. – Москва : МИЭМ, 2011. – 272 с.

4. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Е. И. Большакова [и др.]. – Москва : НИУ ВШЭ. – 2017. – 268 с.

5. *Барсегян, А. А.* Анализ данных и процессов: учеб. пособие / А. А. Барсегян [и др.]. – 3-е изд., перераб. и доп. – Санкт-Петербург : БХВ-Петербург, 2009. – 512 с.

6. *Калабин, А. Л.* Программная система для анализа текстов. / А. Л. Калабин, А. В Туляков // Математические методы в технике и технологиях: сб. тр. междунар. науч. конф.: в 12-ти т. – Т. 8 / под общ. ред. А. А. Большакова. – Санкт-Петербург: Изд-во Политехн. ун-та. – 2018. – С. 55–58.

7. *Нгуен, М. Т.* Тестирование методов машинного обучения в задаче классификации http запросов с применением технологии TF-IDF / М. Т. Нгуен // Вестник Воронеж. гос. ун-та. Сер.: Системный анализ и информационные технологии. – 2019. – № 4. – С. 119–131.

8. *Kim, S.* Research paper classification systems based on TF-IDF and LDA schemes. /

S. Kim, J. Gil. – Human-centric Computing and Information Sciences 9, 30 (2019). – 2019. – DOI: 10.1186/s13673-019-0192-7.

9. *Havrlant, L.* A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation) / L. Havrlant, V. Kreinovich // International Journal of General Systems, vol. 46, no. 1. – 2017. – P. 27–36. – DOI: 10.1080/03081079.2017.1291635.

10. *Asir, D.* Literature Review on Feature Selection Methods for High-Dimensional Data / D. Asir, S. Appavu, E. Jebamalar // International Journal of Computer Applications, V. 136, No 1. – 2016. – P. 9–17. – DOI: 10.5120/ijca2016908317.

11. Feature selection in machine learning: A new perspective / J. Cai [etc.]. // Neurocomputing, vol. 300 (26 July 2018). – 2018. – P. 70–79. – DOI: 10.1016/j.neucom.2017.11.077.

12. *Mikhaylov, D. V.* An approach based on tf-idf metrics to extract the knowledge and relevant linguistic means on subject-oriented text sets / D. V. Mikhaylov, A. P. Kozlov, G. M. Emelyanov // Computer Optics, vol. 39, iss. 3. – 2015. – P. 429–435. – DOI: 10.18287/0134-2452-2015-39-3-429-435.

13. *Ландэ, Д. В.* Поиск знаний в Internet. Профессиональная работа / Д. В. Ландэ; пер. с англ. – Москва: Изд. дом «Вильямс», 2005. – 272 с.

Калабин Александр Леонидович – д-р. физ.-мат. наук, проф., заведующий кафедрой Программного обеспечения ФГБОУ ВО «Тверской государственный технический университет».
Email: akalabin@yandex.ru
ORCID iD: https://orcid.org/0000-0002-4112-4996

Корнеева Елена Игоревна – аспирант 4-го года обучения кафедры Программного обеспечения ФГБОУ ВО «Тверской государственный технический университет».
E-mail: yelena.korneeva@yandex.ru
ORCID iD: https://orcid.org/0000-0002-9793-9713

DOI: https://doi.org/10.17308/sait.2020.2/2924 ISSN 1995-5499
Received 25.03.2020
Accepted 15.06.2020

ANALYSIS OF INFORMATION CRITERIA OF RELEVANT FEATURE SELECTION IN TEXT MINING METHODS

© 2020 A. L. Kalabin✉, Y. I. Korneeva

Tver State Technical University
22, Afanasya Nikitina Embankment, 170026 Tver', Russian Federation

Annotation. In this paper, a quantitative and qualitative assessment of document feature selection methods based on information theory was conducted. The aim of the research was to verify the application of a number of criteria for reduction of a multitude of terms in a collection of texts, to which supervised and unsupervised classification methods would be subsequently applied. The input data for the implemented software was divided by the similarity of topics and, depending on the experiment, included sets of 45 documents of three categories of technical texts in various concentrations. The TextStageProcessor software system for text mining, an open source code project, was used to calculate the criteria. Two values were introduced in the section of criteria performance evaluation. The first determined the relative number of documents which belonged to the category and contained a specified term. The second one was equivalent to the relative number of documents which belonged to the category and did not contain the specified

✉ Kalabin Alexander L.
e-mail: akalabin@yandex.ru

term. Graphs for the dependence of the above-mentioned values on the criteria were constructed. Limitations for the specified parameters were considered. The results obtained for MI, CHI, and IG criteria are not monotonous, which indicates the possible inoperability of these criteria for the input collection of documents and the need for further research. The texts were preprocessed for the second part of the experiment, which included the removal of stop words, normalising the terms, and making them lowercase. The quality view of the graphs of the dependence of TFD, DF, and TF·IDF criteria on the word rank in the collection shows that they can be used to reduce the multitude of relevant input terms for the classification with no loss in quality of the research.
Keywords: text mining, feature selection methods, term frequency, collection of documents, criteria evaluation.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Yang Y. A, Pedersen J. O.* Comparative Study on Feature Selection in Text Categorization. Proceedings of the 14th International Conference on Machine Learning, 8–12 July 1997, Nashville, Tennessee, USA. 1997. P. 412–420.

2. *Meng J., Lin H., Yu Y.* A Two-stage feature selection method for text categorization. Computers and Mathematics with Applications, October 2011. 2011. V. 62, iss. 7. P. 2793–2800. DOI: 10.1016/j.camwa.2011.07.045.

3. *Bol'shakova E. I., Klyshinskij E. S., Lande D. V., Noskov A. A., Peskova O. V., Yagunova E. V.* Natural Language Processing and Computational Linguistics. Moscow, Russia. Moscow, MIEM publ. 2011. p. 272. (In Russian)

4. *Bol'shakova E. I., Voroncov K. V., Lukashevich N. V., Sapin A. S.* Natural Language Processing and Data Mining. Moscow, Russia. Moscow, HSE publ. 2017. p. 268. (In Russian)

5. *Barsegyan A. A., Kupriyanov M. S., Holod I. I.* Data Mining and process analysis. Saint-Petersburg, Russia. BHV-Peterburg publ. 2009. p. 512. (In Russian)

6. *Kalabin A. L., Tulyakov A. V.* Text Mining computer software system. Proceedings of the 31st International Conference Matematicheskie metody v tekhnike i tekhnologiyah: Bol'shakov A. A. (ed.). Vol. 8. Saint-Petersburg, Russia. Izdatel'stvo politekhnicheskogo universiteta publ. 2018. P. 55–58. (In Russian)

7. *Nguyen M. T.* Machine Learning methods testing within http requests classification problem with the use of TF-IDF algorithm. Proceedings of Voronezh State University. Series: Systems analysis and information technologies. 2019. (4). P. 119–131. (In Russian)

8. *Kim S. Gil L.* Research paper classification systems based on TF-IDF and LDA schemes. Human-centric Computing and Information Sciences 9, 30 (2019). DOI: 10.1186/s13673-019-0192-7.

9. *Havrlant L., Kreinovich V.* A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). International Journal of General Systems, 2017. V. 46 (1). P. 27–36. DOI: 10.1080/03081079.2017.1291635.

10. *Asir D., Appavu S., Jebamalar E.* Literature Review on Feature Selection Methods for High-Dimensional Data. International Journal of Computer Applications, 2016. V. 136 (1). P. 9–17. DOI: 10.5120/ijca2016908317.

11. *Cai J., Luo J., Wang S., Yang S.* Feature selection in machine learning: A new perspective. Neurocomputing, vol. 300 (26 July 2018). 2018. P. 70–79. DOI: 10.1016/j.neucom.2017.11.077.

12. *Mikhaylov D. V., Kozlov A. P., Emelyanov G. M.* An approach based on tf-idf metrics to extract the knowledge and relevant linguistic means on subject-oriented text sets. Computer Optics, 2015. V. 39, iss. 3. P. 429–435. DOI: 10.18287/0134-2452-2015-39-3-429-435.

13. *Lande D. V. (ed.)* Web Knowledge Retrieval. Specialized work. Moscow, Russia. Moscow, Izdatel'skij dom "Vil'yams" publ., 2005. 272 p. (In Russian)

Kalabin Alexander L. – DSc in Physics and Mathematics, Professor, Head of Software Department of Tver State Technical University.
E-mail: akalabin@yandex.ru
ORCID iD: https://orcid.org/0000-0002-4112-4996

Korneeva Yelena I. – 4th year postgraduate student, Software Department of Tver State Technical University.
E-mail: yelena.korneeva@yandex.ru
ORCID iD: https://orcid.org/0000-0002-9793-9713