The Method for Reducing the Term Vector Size for Category Classification of Text Documents

Golub T.V., Tiahunova M. Yu.
Zaporozhye National Technical University
Zaporozhye, Ukraine

***Abstract*.** The article proposes a method for reducing time necessary for subsuming a certain document in order to classify the text documents by reducing the term vector size of certain categories. According to the method, the term weight factors were calculated for each classification category to implement subsuming process at the stage of training a certain system. As a result of the analysis of the obtained data, the individual category terms, whose weight values did not exceed the experimentally determined threshold value, were excluded from the term vector of the category by equating them to zero. Those terms were not involved in the further subsuming process at the testing stage. As the input data for the experimental part, the TF-SLF reference method and its modernization CTFSLF according to those described above were proposed. Due to the application of the method proposed, the differential term vector size for each category was decreased. Despite the increase in the compile time of the term vector according to categories, which was performed only once, the calculation time used to determine whether or not a document belonged to a specific category decreased without losing the classification quality. In addition, due to the fact that the proposed method excluded the words that were used in the texts frequently, it became possible to exclude the stage of removing the stop words from the pretreatment process of the analyzed text. For the same reason, the problem of misprints and the words "stuck together" in the initial, training sample was solved.
***Keywords*:** text classification, stemming, terms vector, term weight, TF-SLF.
DOI: 10.5281/zenodo.3240216

Metoda de reducere a dimensiunii vectorului termenilor de clasificare a documentelor text pe categorii

Golub T.V., Tyagunova M.Yu.
Universitatea Tehnică Națională din Zaporoje
Zaporoje, Ucraina

Rezumat. Scopul studiului prezentat în articol a fost de a dezvolta o metodă de reducere a timpului necesar evaluării proprietăților unui document pentru anumite categorii, pentru a clasifica documentele text. Acest obiectiv este realizat prin reducerea dimensiunii vectorului termenilor anumitor categorii. Pentru a implementa procesul de determinare a proprietății unui document dintr-o anumită categorie în stadiul de pregătire a sistemului, conform metodei propuse, ponderile termenilor se calculează separat pentru fiecare categorie de clasificare. Ca rezultat al analizei datelor obținute, termenii categoriilor individuale, ale căror valori ale greutății nu depășesc valoarea pragului determinat experimental, sunt excluse din vectorul termenilor unei anumite categorii prin egalarea lor cu zero. Acești termeni nu sunt implicați în procesul ulterior de evaluare a proprietății unei anumite categorii de documente în etapa de testare. Metoda de referință pentru determinarea ponderii termenilor TF-SLF, descrisă în literatură, și modernizarea acesteia pe categorii în conformitate cu descrierea de mai sus a CTFSLF, a fost utilizată ca date inițiale pentru partea experimentală. Ca rezultat al aplicării metodei propuse, mărimea vectorului termenilor caracteristici pentru fiecare categorie s-a redus, iar în consecință, în ciuda creșterii timpului de compilare a vectorului termenilor în categorii, care este efectuată o dată, timpul de efectuare a calculelor pentru a determina dacă un document aparține unei categorii specifice fără a pierde clasificarea calității de asemenea s-a redus. De asemenea, datorită faptului că metoda propusă exclude termenii frecvent utilizați în texte, devine posibilă excluderea etapei de eliminare a cuvintelor stop din textul analizat din procesul de preprocesare a documentelor.
Cuvinte-cheie: clasificare text, derivare, vector termen, pondere termen, TF-SLF.

Метод уменьшения размера вектора термов для классификации текстовых документов по категориям

Голуб Т.В., Тягунова М.Ю.
Запорожский национальный технический университет
Запорожье, Украина

***Аннотация*.** Целью исследования, представленного в статье, была разработка метода для уменьшения времени, затрачиваемого на процесс оценки принадлежности документа отдельным категориям, с целью классификации текстовых документов. Данная цель достигается путем уменьшения размера вектора термов отдельных категорий. Для реализации процесса определения принадлежности документа

отдельной категории на этапе обучения системы, согласно предложенному методу, выполняется расчет весовых коэффициентов термов для каждой категории классификации в отдельности. В результате анализа полученных данных термы отдельных категорий, весовые значения которых не превышают экспериментально определенное пороговое значение, исключаются из вектора термов отдельной категории путем приравнивания их к нулю. Данные термы не участвуют в дальнейшем процессе оценки принадлежности документа отдельной категории на этапе тестирования. В качестве исходных данных для проведения экспериментальной части были использованы опорный метод определения весовых значений термов TF-SLF, описанный в литературе, и предложенная авторами его модернизация по категориям согласно приведенному выше описанию CTFSLF. В результате применения предложенного метода уменьшился размер вектора характерных термов для каждой категории, вследствие чего, не смотря на увеличение времени на составление вектора термов по категориям, которое выполняется один раз, уменьшилось время на выполнение расчетов для определения принадлежности документа конкретной категории без потери качества классификации. Также, в связи с тем, что предложенный метод исключает часто используемые в текстах слова, из процесса предварительной обработки документа становится возможным исключить этап удаления стоп-слов из анализируемого текста. По этой же причине решается проблема опечаток и «слипшихся» слов в исходной, обучающей выборке. Таким образом, поставленную в начале цель исследования можно считать достигнутой.
Ключевые слова: классификация текстов, стемминг, вектор термов, вес терма, TF-SLF.

ВВЕДЕНИЕ

Количество информации, представленной в текстовом виде, увеличивается беспрерывно. Текстовая информация накапливается во всех областях деятельности человека. Начиная с хранения данных на персональных компьютерах и гаджетах и заканчивая Big Data как части киберфизических систем. Она охватывает такие области, как бизнес, работу исследовательских институтов, государственных и финансовых учреждений, интенсивно использующих технологии. Текстовая информация содержит статистические данные, управляющие команды, справочную информацию, законы изменения произвольных процессов. Особенностью такой информации является отсутствие ее структурированности, что усложняет процесс анализа [1].

Аналитика текста преобразует текст в числа, а числа, в свою очередь, позволяют упорядочить данные и помогают выявлять закономерности. Чем более структурированы данные, тем лучше анализ и, в конечном итоге, тем более качественными будут решения, принятые на его основе [2, 3].

Для возможности ориентирования во всем многообразии такой информации, в частности для поиска необходимых пользователю данных появляется необходимость ее классификации [4]. Данный процесс и является объектом рассмотрения предложенного исследования.

Классификация текстов относится к одной из задач компьютерной лингвистики, которая включает в себя определение тематической принадлежности текстов, автора текста, эмоциональной окраски высказываний и другого.

Для упрощения поиска необходимой информации в массивах данных необходимо решить задачу систематизации документов. Такая задача является одной из наиболее актуальных, для решения которых требуется использование классификации текстов [5]. В связи с беспрерывным возрастанием потока данных, требующих классификации, реализация данной задачи значительно усложняется и потому является актуальной.

В литературе описано множество подходов к решению данной задачи. В [1, 6 – 8] приведены обзор и сравнение актуальных на данный момент методов в соответствии с различными этапами данного процесса. В результате их анализа можно сделать вывод, что одним из важных моментов процесса классификации документов является выделение ключевых признаков. Решению данной задачи были посвящены работы [9 – 15], в которых раскрыты различные подходы, включая статистические, частотные, латентно-семантические и другие. Однако предложенные методы рассматривают термы в рамках всей коллекции документов, что не позволяет оценить важность отдельного терма в рамках каждой категории отдельно.

Классификация текстовых документов подразумевает процесс анализа его содержания и автоматического определения документа в одну или несколько категорий [16, 17]. Категориями являются множества документов, объединенные общей тематикой. При этом множество категорий задается

экспертом, либо определяется автоматически на основании обучающей выборки. На этапе обработки документов в информационно-аналитической системе используют автоматический классификатор – программу, определяющую тематику документов и осуществляющую их отнесение к рубрикам [6].

Также является актуальной обратная задача - выбор из множества документов тех, которые принадлежат отдельной категории, определенной пользователем. В этом случае является актуальным время выполнения определения принадлежности документа категории. Данная задача и является предметом проведенного исследования. В статье предлагается метод уменьшения времени, затрачиваемого на процесс оценки принадлежности документа отдельным категориям путем уменьшения размера вектора термов отдельных категорий для классификации текстовых документов.

I. Методы исследования

A. Математическая модель процесса классификации

Приведенные в [18, 19] математические модели процесса классификации текстовых документов являются общими. В данной статье авторами предложено усовершенствование существующих вариантов процесса определения весовых коэффициентов термов как части процесса классификации с учетом требований поставленной задачи. Приведем необходимые для понимания метода определения.

Для формального описания процесса решения поставленной задачи принадлежности текстовых документов отдельной категории предположим следующее.

Пусть имеется:

$T=\{t_1,\dots t_{|A|}\}$ — множество термов (слов) документа;

$B=\{b_1,\dots b_{|B|}\}$ — множество возможных слов;

$D=\{d_1,\dots d_{|D|}\}$ — множество документов;

$C=\{c_1,\dots c_{|C|}\}$ — множество категорий;

$E=\{e_1,\dots e_{|E|}\}$ — множество термов категории.

В этом случае решение поставленной задачи можно формализовать следующим образом. Существует множество документов D, из которых нужно выбрать те, что принадлежат определенной заранее категории c_i из множества категорий C.

$$\Phi(d_j,c_i)=\begin{cases}0, & if \quad d_j \notin c_i\\ 1, & if \quad d_j \in c_i\end{cases} \qquad (1)$$

B. Создание множества данных документа

Классификация текстовых документов выполняется на основании анализа термов этих документов.

Для выполнения задачи классификации необходимо представить текст в форме модели множества. В этой форме текст рассматривается как множество слов – термов, имеющих некоторый вес. Термом является интуитивно определенное выражение формального языка, являющееся формальным именем объекта [13]. В данном случае под термом будем понимать слово, полученное в результате выполнения операции стемминга - приведения к некоему нормальному виду с усечением его окончаний и суффиксов. Одним из простейших методов стемминга слов является алгоритм Портера [20, 21] Это одна из задач этапа предварительной обработки текста.

При определении принадлежности документа какой-либо категории с учетом значимости каждого терма необходимо выполнить предварительную обработку текста.

Процесс предварительной обработки текста имеет следующие характеристики:

1. $T \in B$ — все термы документа входят в множество возможных слов;

2. $E \in B$ — все термы категории входят в множество возможных слов;

Совокупность элементов множеств T и E формируют множество B. Таким образом, множества T и E являются составляющими множества B.

3. $T_M = <n_1(a_1), n_2(a_2), \dots n_{/A/}(a_{/A/})>$ — мультимножество множества T, которое позволяет собрать вхождение элементов множества несколько раз;

Формирование мультимножества документов одной группы, то есть категории, позволяет для каждого терма определить его мощность. Этот параметр оценивает количественный показатель встречаемости данного терма (в скольких документах категории встречается данный терм хотя бы один раз).

4. $E_M = <n_1(e_1), n_2(e_2), \dots n_{/E/}(e_{/E/})>$ — мультимножество множества E, которое позволяет собрать вхождение элементов множества несколько раз.

При формировании мультимножества категорий возможно выделить термы по мощности их вхождения. Этот параметр показывает, во скольких категориях коллекции встречается рассматриваемый терм хотя бы один раз, что позволяет выделить термы, характерные для всех категорий и не являющиеся признаком принадлежности отдельной категории. Они перестают быть информативными и их можно исключить из анализируемого множества, используемого для дальнейшей классификации.

Последующая обработка текста выполняется с учетом этих характеристик.

C. Определение веса терма

Для оценки принадлежности документа категории вначале нужно определить вес каждого встречаемого терма множества B, т.е. весовые значения термов множества E для каждой категории.

Вес терма в множестве характеризует значимость, или важность, данного терма. Чем выше этот коэффициент, тем с большей вероятностью указанный терм является значимым для определения категории документа в целом. Если терм не встречается в документе, то его вес равен нулю [6].

Для определения весовых значений каждого терма множества B в рамках коллекции в целом используется параметр SLF – коэффициент, характеризующий оценку термов с учетом их вхождения в категории [13]. В отличие от множества других подходов определения весовых значений, данный метод учитывает важность каждого отдельного терма для конкретной категории.

Параметр SLF для каждого терма в рамках коллекции определяется согласно формуле:

$$TFSLF_t = TF_t(E_{/t/}) \cdot SLF_t, \qquad (2)$$

где $TF_t(E_{/t/})$ — частота терма, принадлежащего множеству B, определяется как отношение числа вхождения некоторого терма к общему количеству термов документа. Таким образом, оценивается важность терма t_i в пределах отдельного документа d_j [9].

SLF_t — логарифмированная сумма частот терма t.

$$SLF_t = log(|C|/\sum(NDF_{tc})), \qquad (3)$$

где NDF_{tc} — нормализованная частота встречаемости терма t в категории c, данная оценка является локальной для категории.

$$NDFtC = df_{tc}/N_c, \qquad (4)$$

где df_{tc} — число документов категории c, в которых встречается хотя бы раз терм t;

N_c – количество документов в категории c.

В результате будет получен вектор B_T, содержащий значение весовых коэффициентов термов множества T в рамках всей коллекции в целом. В данном случае не в полной мере учитывается значимость термов, принадлежащих отдельной категории, что снижает качественные показатели выполнения классификации текстов, принадлежащих близких по смыслу и используемых слов тематик. Для реализации этой задачи предлагается подход, представленный далее.

D. Предложенный авторами подход

Целью предложенного авторами подхода является выделение и исключение неинформативных термов для отдельной категории, т.е. в результате оставить информативные термы, характеризующие категорию и присущие ей. Данные действия ведут к сокращению вектора термов и, как следствие, объема вычислений, приводимых при поиске в общей коллекции документов, принадлежащих отдельной рассматриваемой категории. В результате сокращается время анализа каждого документа для принятия решения о его принадлежности категории.

На основании параметра SLF для каждой категории при достаточно большом объеме обучающей выборки можно выделить малоинформативные термы для отдельной категории и исключить их из анализа последовательности (прировнять весовое значение терма к нулю для отдельной категории).

Для выполнения приведенного анализа предлагается использовать предложенный авторами параметр CTFSLF. Данный параметр находится как произведение параметра SLF для коллекции в целом и параметра tf_{ic}/df_{ic} для каждой категории в отдельности. В результате получается нормированное значение относительно суммарного количества термов в коллекции, которое показывает процентную часть принадлежности терма к отдельной категории относительно всей коллекции, или взвешенный параметр веса терма, не зависящий от размера документа.

Алгоритм определения весовых значений термов вектора E_T для каждого e_i:

1. Определение коэффициента *tf/df* для каждого терма множества e_i, относящихся к категории, в рамках коллекции в целом как отношение общего количества каждого терма в рамках отдельной категории к общему количеству каждого терма в рамках коллекции в целом:

$$TF(t_i, c_j) = \frac{fr_{ij}}{\sum_i fr_{ij}}, \quad (5)$$

где $0 \le i \le |E|$, $0 \le j \le |C|$.

2. Определение значения веса каждого терма по категориям с учетом его встречаемости в категориях коллекции (множество E, содержащее значения $CTFSLF(t_i,c_j)$ каждого терма категории) как произведение коэффициента tf/df для каждого терма отдельных категорий и параметра SLF:

$$CTFSLF(t_i, c_j) = TF(t_i, c_j) * SLF_k, \quad (6)$$

где $0 \le i \le |E|$, $0 \le j \le |C|$, $0 \le k \le |B|$

Данный показатель веса терма характеризует степень его принадлежности отдельной категории.

3. Следующим этапом является определение и отсечение нехарактерных термов для каждой категории в отдельности. Для определения порогового значения с целью отсечения неинформативных термов на основании экспериментальных данных принимается величина, обратная количеству документов, принадлежащих анализируемой категории.

$$K_j = 1/|D_j| \quad (7)$$

где $0 \le j \le |C|$

Термы, весовые значения которых ниже указанного порогового значения, характеризуются высокой степенью встречаемости во всех категориях коллекции, потому их можно исключить из перечня характерных термов для отдельной коллекции как неинформативные.

4. На основании предыдущих данных принимается решение о значении весовых коэффициентов термов в рамках каждой категории. С этой целью для каждого терма каждой категории выполняется сравнение его значения веса с пороговым значением:

$$\Psi(e_i, c_j) = \begin{cases} 0, & if \quad e_i < k_i \\ CTFSLF(t_i, c_j), & if \quad e_i \ge k_i \end{cases} \quad (8)$$

где $0 \le i \le |E|$, $0 \le j \le |C|$.

Если данный параметр меньше порогового значения, эта величина веса терма приравнивается к нулю.

В результате данных преобразований уменьшается вектор анализируемых ключевых термов для отдельных категорий, что приводит к уменьшению затрачиваемого времени на выполнение анализа принадлежности документа категории на этапе функционирования предложенного метода.

II. Результаты исследования и их обсуждение

В качестве исходных категорий для проведения тестирования предложенного метода были выбраны категории классификатора УДК класса 004 «Информационные технологии. Вычислительная техника. Обработка данных», а именно:

- 004.0 «Специальные определители для вычислительной техники»,
- 004.2 «Архитектура вычислительных машин»,
- 004.4 «Программные средства»,
- 004.9 «Прикладные информационные (компьютерные) технологии. Методы, основанные на применении компьютеров».

Основной проблемой автоматического определения индекса УДК является близость характерных термов категорий между собой и сложность определения признаков их дифференциации. Потому данное направление является актуальным.

A. Описание обучающей выборки и обучения

Под обучающей выборкой авторами понимается набор исходных текстовых документов с определенными заранее категориями принадлежности, на основании которых выполняется построение опорного вектора весовых коэффициентов термов категорий, в дальнейшем используемых для классификации.

В качестве обучающей выборки для апробации разработанного метода были выбраны научные статьи по соответствующим тематикам, представленные на одном языке. Тематики (значения УДК) определялись и указывались авторами научных работ непосредственно в публикации.

Для каждой категории была подобрана обучающая выборка, представляющая собой перечень документов *D*, принадлежащих отдельным категориям *C*, для составления опорного вектора термов, множества *E*. Обучающая выборка была представлена 30 публикациями для каждой выбранной категории.

Для построения опорного вектора классификации на основании обучающей выборки основываясь на использовании параметра SLF [22] были выполнены следующие этапы:

1. предварительная обработка текстового документа для получения множества термов E: исключение вспомогательных символов и знаков пунктуации, стемминг слов;

2. определение весовых коэффициентов термов:

- подсчет частоты встречаемости термов в рамках отдельных категорий и коллекции в целом;
- подсчет коэффициента CTFSLF(ti,cj) для каждого терма коллекции.

Для реализации предложенного метода этап определения весовых коэффициентов термов усовершенствован авторами тем, что добавлены следующие подэтапы:

1. определение весовых значений термов в рамках отдельных категорий;

2. определение порогового значения для отсечения неинформативных термов в коллекциях;

3. определение отдельно для каждой категории коэффициента Ψ(ei,cj) каждого терма.

B. Описание тестовой выборки и тестирования

В качестве тестовой выборки для предложенной модели был взят перечень документов, которые не участвовали на этапе обучения.

Для подготовки тестовой выборки были выполнены следующие этапы:

1. предварительная обработка текстового документа для получения множества термов E: исключение вспомогательных символов и знаков пунктуации, стемминг слов;

2. определение весового значения термов в анализируемом документе:

- подсчет частоты встречаемости термов в рамках документа;
- определение весового значения терма как произведение результатов предыдущего пункта и коэффициента $\Psi(e_i,c_j)$ для каждой анализируемой категории;
- принятие решения о принадлежности документа к отдельной категории на основании суммарного значения весов термов для каждой категории.

C. Тестирование предложенного метода

Проведем проверку режимов работы исходного и предложенного авторами методов и сравним результаты их работы. Для этого реализуем этапы обучения и тестирования для систем, построенных на их основе.

Обучение. В результате выполнения описанных этапов обработки обучающей выборки были получены следующие результаты, приведенные в таблице 1.

В таблице 1 приведены суммарное количество слов в документах обучающей выборки и количество термов, полученных в результате обработки исходным и предложенным методами отдельно для каждой категории. А также долевая часть размера вектора термов для обоих методов относительно общего количества слов в исходных документах. Колонки части «Сравнение методов» отражают качественные показатели уменьшения вектора термов согласно предложенному методу относительно исходного. Как видно из таблицы 1, размер вектора термов относительно суммарного количества слов документов обучающей выборки для исходного метода составляет в среднем 15,05%. Для предложенного метода это значение составляет 11,75%. При этом размер вектора термов предложенного метода относительно исходного сократился в среднем на 21,53%.

Для большей наглядности на рис. 1 приведено графическое представление сравнения размера векторов термов исходного и предложенного методов.

Как видно из упомянутого рисунка, вне зависимости от количества слов в исходных документах в каждой категории наблюдается сокращение результирующего размера вектора термов, полученного при применении предложенного, метода относительно исходного.

Таблица 1 [1]

Результаты выполнения этапа обучения

Категория (Categories)	Кол-во слов всего в документах (Number of words in documents)	SLF		CTFSLF		Сравнение методов	
		Кол-во термов (Terms number)	Доля термов от коллекции (Terms part from corpus)	Кол-во термов (Terms number)	Доля термов от коллекции (Terms part from corpus)	Кол-во исключенных термов (Excluded terms number)	Доля уменьшения вектора термов (Part of terms set decreasing)
004.0	148419	22118	14,90%	18450	12,43%	3668	16,58%
004.2	111213	12510	11,25%	8978	8,07%	3532	28,23%
004.4	108077	18752	17,35%	14652	13,56%	4100	21,86%
004.9	104207	17411	16,71%	13473	12,93%	3938	22,62%
Среднее значение (Average value)	117979	17698	15,05%	13888	11,75%	3810	21,53%

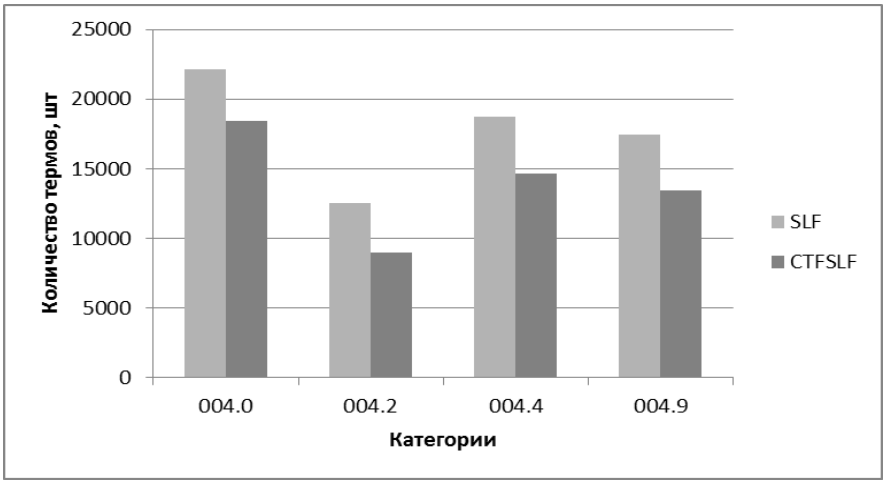

Рис. 1 Размеры векторов термов исходного и предложенного методов после этапа обучения [2].

Исключенные термы. Примерами термов, присутствующих в результирующем векторе исходного метода, и исключенных при применении предложенного метода, являются: *счита, одинаков, рассмотрен, функционир, реж, литератур, начальн, пример, реализац, течен, модел, аналогичн, длин, содержа, порог, описа, передач, мож, эт, недостаток, минимизирова, переходн* и другие.

Как видно из приведенного примера, опираясь на данные термы сложно определить принадлежность документа к определенной категории. Потому исключение данного перечня термов не влияет на качество классификации. Что и подтвердили дальнейшие исследования.

Тестирование. На этапе тестирования проводилось исследование определения наиболее вероятной категории принадлежности анализируемого документа для исходного и предложенного методов, а также замер временных затрат на выполнение соответствующих расчетов. Результаты проведенных наблюдений представлены в таблице 2.

В первой колонке указаны категории, к которым были отнесены анализируемые документы авторами публикаций. Как видно из таблицы 2, результирующее время на тестирование сократилось на 18,61% без потери качества классификации.

На рисунке 2 приведены временные затраты на процесс классификации документов, наглядно показывающие представленные в таблице 2 результаты.

[1,2]Appendix 1

Таблица 2 [3]

Результаты выполнения этапа тестирования

Категория тестового документа (Text document category)	SLF		CTFSLF		Процент уменьшения временных затрат (Percentage of spent time reducing)
	Определенная категория (Certain category)	Затраченное время, с (Spent time, s)	Определенная категория (Certain category)	Затраченное время, с (Spent time, s)	
004.056	004.0	0,03125	004.0	0,02500	-20,00%
004.274	004.2	0,01875	004.2	0,01250	-33,33%
004.4	004.4	0,02188	004.4	0,02188	-0,01%
004.93	004.9	0,02813	004.9	0,01563	-44,44%
004.94	004.0	0,02813	004.4	0,03125	+11,11%
004.93+004.4	004.9	0,02500	004.9	0,01875	-25,00%
Результирующее значение (Result values)		0,15313		0,12501	-18,61%

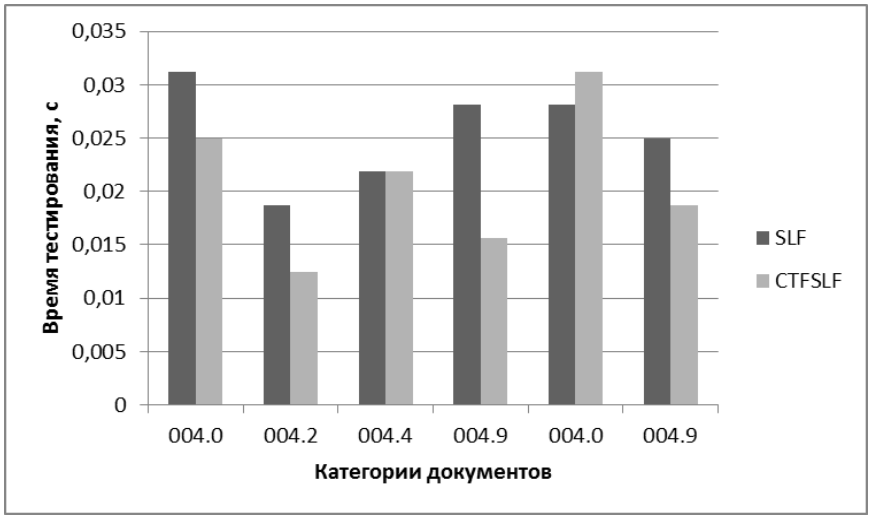

Рис. 2 Время, затраченное на классификацию документов с помощью исходного и предложенного методов [4].

Сравнение результатов. Временные затраты на обучение и тестирование исходного и предложенного методов приведены в таблице 3. Для проведения экспериментальной части предложенного метода использовалась следующая аппаратная платформа: процессор — Intel Core i5 (1,70 ГГц), ОЗУ — 8 ГБ, накопитель SSD – 120 ГБ.

Для большей наглядности на рис. 3 приведено графическое представление временных затрат на этапах обучения и тестирования для обоих анализируемых методов.

Как видно из таблицы 3 и рис. 3, время на составление вектора термов увеличилось на 2,27%. При этом затраченное время на этапе тестирования сократилось на 18,37%.

Таким образом, количество значимых термов, включенных в оценку принадлежности документа отдельной категории, в среднем сократилось на 21,53%. При этом время на анализ каждого документа в отдельности без потери качества классификации сократилось в среднем на 18,61%. Общее время на классификацию сократилось на 18,37% при увеличении затраты времени на этапе обучения на 2,27%. Следовательно, не смотря на увеличение времени на этапе обучения, данный метод показал уменьшение временных

[2,3] Appendix 1

затрат на этапе работы классификатора без потери качества и потому является перспективным.

Таблица 3 [5]

Сравнение временных затрат на этапах обучения и тестирования

Метод (Method)	Время на обучение, с. (Training time, s)	Время на тестирование, с. (Testing time, s)
Исходный метод SLF (Initial method SLF)	1,25145	0,15313
Предложенный метод CTFSLF(t_i,c_j) (Proposed method CTFSLF(t_i,c_j))	1,2798	0,12501
Процент отклонения (Deviation percentage)	+2,27%	-18,37%

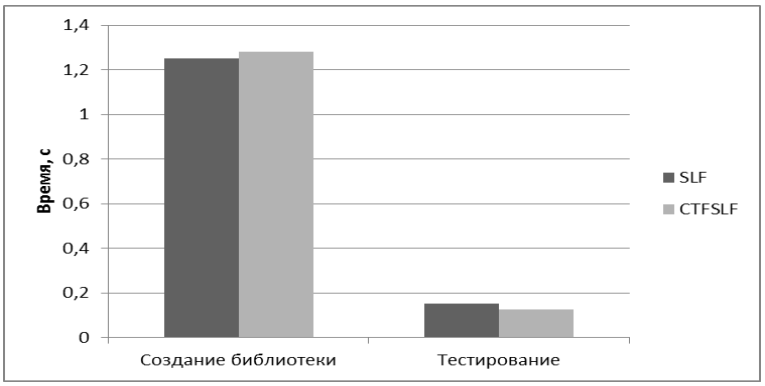

Рис. 3. Временные затраты на обучение и тестирование исходного и предложенного методов [6].

III. Выводы

В статье предложен метод уменьшения времени, затрачиваемого на процесс оценки принадлежности документа отдельным категориям с целью классификации текстовых документов путем уменьшения размера вектора термов отдельных категорий.

В результате применения предложенного метода уменьшился размер вектора характерных термов для каждой категории, в следствие чего, не смотря на увеличение времени на составление вектора термов по категориям, которое выполняется один раз, уменьшилось время на выполнение расчетов для определения принадлежности документа отдельной категории без потери качества классификации. Также, в связи с тем, что предложенный метод исключает часто используемые в текстах слова, из процесса предварительной обработки документа становится возможным исключить этап удаления стоп-слов из анализируемого текста. По этой причине же таким же образом решается проблема опечаток и «слипшихся».

Таким образом, результаты тестирования данного метода показали его перспективность, а именно в сравнении с исходным методом определения весовых значений термов, основанного на использовании метода SLF для коллекции в целом, и предложенного метода выделения весовых значений по отдельным категориям на основании экспериментальных данных были получены следующие результаты:

- увеличилось время на составление вектора термов категорий на 2,27%;
- уменьшилось время на выполнение расчетов для определения принадлежности документа конкретной категории на 18,37%;
- уменьшился размер вектора характерных слов в среднем на 21,53%;
- весовые значения термов были приравнены к 0 в рамках каждой категории в среднем на 3810 термов из 117979 исходных слов;
- определение принадлежности документа категории выполняется без потери качества (в сравнении с исходным методом);
- из процесса предварительной обработки документа становится возможным исключить этап удаления стоп-слов из анализируемого

[5,6]Appendix 1

текста;

- решается проблема опечаток и «слипшихся» слов в анализируемом тексте.

APPENDIX 1 (ПРИЛОЖЕНИЕ 1)

[1]**Table 1.** The results of the training phase.

[2]**Fig. 1.** Terms vectors sizes of the original and proposed methods after the training phase.

[3]**Table 2.** The results of the testing phase.

[4]**Fig. 2.** Time spent on the documents classification using original and proposed methods.

[5]**Table 3.** Comparison of time spent at the stages of training and testing.

[6]**Fig. 3.** Time spent on learning and testing to the original and proposed methods.

Литература (References)

[1] Thangaraj M., Sivakami M. Text classification techniques: A literature review. *Interdisciplinary Journal of Information, Knowledge, and Management*, 2018, vol. 13, pp. 117-135.

[2] Brindha S., Sukumaran S., Prabha, K. A survey on classification techniques for text mining. *Proceedings of the 3rd International Conference on Advanced Computing and Communication Systems. IEEE*. Coimbatore, Indi,. 2016 (In English) Available at: https://doi.org/10.1109/ICACCS.2016.7586371 (accessed 13.03.2019)

[3] Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey. *Frontiers of computer science in China*, 2010, vol. 4, no. 2, pp. 280–301.

[4] Korde V., Mahender N. Text classification and classifiers: a survey. *International Journal of Artificial Intelligence & Applications (IJAIA)*, 2012, Vol. 3, no. 2, pp. 85–99

[5] Pankov S. V., Shebanin S. P., Ribakov A. A., Thematic classification of text. *ROOKEE, ROMIP 2010*, Kazan', Russia, 2010, pp. 142-147

[6] Golub T. The Analysis of text documents classifiers constructing methods, *Modern problems of radio engineering, telecommunications, and computer science*, 2016, pp.742-745.

[7] Yang Y., Zhang J., Kisiel B. A scalability analysis of classifiers in text categorization. *ACM SIGIR'03*, 2003. Available at: https://dl.acm.org/citation.cfm?id=860455 (accessed 13.03.2019)

[8] Sebastiani F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 2002, vol. 34, pp. 1-47.

[9] Karpovich S.N. Mnogoznachnaya klassifikatsiya tekstovykh dokumentov s ispol'zovaniyem veroyatnostnogo tematicheskogo modelirovaniya ml-PLSI [Multi-valued text documents classification using probabilistic thematic modeling ml-PLSI]. *SPIIRAS Proceedings.* 2016. Issue 4(47), pp.92 – 104. doi: 10.15622/sp.47.5.

[10] Kuralegov I. Automatic classification of documents based on latent semantic analysis. *1st Int. Conf. Digital Libraries: Advanced Methods and Technologies, Digital Collections,* St-Petersburg, Russia, 1999, pp. 89-96. (In English)

[11] Andreev A. M. Automatic classification of text documents using the neural network algorithms and semantic analysis. *Advanced Methods and Technologies, Digital Collections,* St-Petersburg, Russia, 2003, pp. 76-86. (In English)

[12] Krasnov A., Ilatovskiy A.S., Khomonenko A.D., Arsen'yev V.N. Otsenka semanticheskoy blizosti dokumentov na osnove latentno-semanticheskogo analiza s avtomaticheskim vyborom rangovykh znacheniy [Evaluation of documents semantic proximity based on latent-semantic analysis with automatic selection of rank values]. *Trudy SPIIRAN – SPIIRAN proceedings,* 2017. no. 5(54), pp. 185-204.

[13] Rehman Abdur, Barbi H., Saeed M., Feature Extraction for Classification of Text Documents *International Conference on Communications and Information Technology (ICCIT 2012)*, Hammamet, Tunisia, 2012, pp. 234 - 239. (In English)

[14] Budanitsky A. Hirst G. Evaluating WordNet-based Measures of Lexical Semantic Relatedness *Computational Linguistics*. 2006. Vol. 32. pp. 13-47.

[15] Bondarchuk D.V. Vektornaya model' predstavleniya znaniy na osnove semanticheskoy blizosti termov [Vector model of knowledge representation based on semantic proximity of terms] *Vestnik YUUrGU. Seriya: Vychislitel'naya matematika i informatika – Bulletin of SUSU. Series: Computational Mathematics and Computer Science.* 2017. vol. 6 no. 3. pp. 73–83. doi: 10.14521/cmse170305.

[16] Tsoumakas G., Katakis I. Multi-label classification: an overview. *International Journal of Data Warehousing & Mining.* 2007. vol. 3(3). pp. 1–13.

[17] Rubin T.N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multilabel document classification. *Machine Learning.* 2012. vol. 88. no. 1–2. pp. 157–208.

[18] Erpev A.S., Automatic classification of text documents, *Mathematical Structures and Modeling.* 2010, vol. 21, pp. 65-81.

[19] Zyuz'kov V. M., Matematicheskaya logika i teoriya algoritmov [Mathematical logic and theory of algorithms], - Tomsk: El Content, 2015, 236 p. Available at: http://www.math.tsu.ru/sites/default/files/mmf2/e-resources/math%20logika%20i%20teoriya-%20algoritmov.pdf (accessed 13.03.2019)

[20] Willett P. The Porter Stemming Algorithm: Then and Now *Program: Electronic Library and Information Systems.* 2006. vol. 4, no. 4. pp. 219-

223.
[21] Golub T.V., Tyahunova M.YU. Metod steminhu ukrayinomovnykh tekstiv dlya klasyfikatsiyi dokumentiv na bazi alhorytmu Portera [The method of Ukrainian language stitemming for the classification of documents based on Porter's algorithm] *Naukovi pratsi Donets'koho natsional'noho tekhnichnoho universytetu. Seriya: Informatyka, kibernetyka – Scientific papers of the Donetsk National Technical University. Series: Informatics, Cybernetics and Computing* 2017, no. 1, pp. 59 – 63.
[22] Oliynyk YU. O., Katyushchenko D. O. Analiz metodiv vyznachennya vah oznak tekstovykh dokumentiv [Analysis of the methods of determining the text documents signs weight] *Naukovyy ohlyad – Scientific Review*, 2018, no. 3(46), pp. 112 – 123.

Сведения об авторах.

Голуб Татьяна Васильевна, ассистент кафедры Компьютерных систем и сетей. Область научных интересов: Классификация текстовых документов, анализ текстовых данных, машинное обучение.
E-mail: golub.tv6.gmail.com

Тягунова Мария Юрьевна, доцент кафедры Компьютерных систем и сетей. Область научных интересов: Интернет вещей, киберфизические системы, нейронные сети, генетические алгоритмы, машинное обучение.
E-mail: golub.tv6.gmail.com