УДК 004.912

**И.В. Глушаускайте[1], Т.Н. Заболотняя[2]**
[1]НТУУ «КПИ», г. Киев, Украина, irina_glushauskaite@gmail.com
[2]НТУУ «КПИ», г. Киев, Украина, tatiana104@yandex.ru

# КОМБИНИРОВАННЫЙ МЕТОД АВТОМАТИЗИРОВАННОГО ОПРЕДЕЛЕНИЯ АВТОРСТВА ТЕКСТОВ

В данной статье предложен метод определения авторства текстов, который сочетает в себе классификацию и кластеризацию документов. Предложен алгоритм реализации метода, который может быть реализован на ЭВМ. Дана оценка эффективности предложенного метода. Обосновано улучшение численных характеристик качества классификации благодаря использованию предложенного комбинированного метода.

АВТОМАТИЗИРОВАННОЕ ОПРЕДЕЛЕНИЕ АВТОРСТВА, КЛАССИФИКАЦИЯ, КЛАСТЕРИЗАЦИЯ, КЛАСТЕРНЫЙ АНАЛИЗ, КЛАССИФИКАЦИЯ С УЧИТЕЛЕМ, КАЧЕСТВО КЛАССИФИКАЦИИ.

## Введение

Определение авторства текста уже много лет является актуальной проблемой как в теоретической лингвистике, так и в области решения прикладных задач. Проведение анализа и сравнения стилей широко применяется в филологических дисциплинах, изучающих особенности организации текстов, в психологии и теории искусственного интеллекта, в судебной практике и криминалистике. Также реализация алгоритмов определения и сравнения стилей текстов широко используется при решении задач информационного поиска, распространенных в связи с увеличением объемов информации, в том числе электронной документации, и ростом сети Интернет.

Сегодняшняя актуальность решения данной проблемы обусловлена повсеместным переходом от рукописного текста к печатному (электронному). Если раньше в выявлении плагиата, определении автора литературного произведения или анонимного текста могла применяться почерковедческая экспертиза, то для печатного текста это невозможно. Автоматизированное определение авторства позволяет также избавиться от двух важных недостатков определения авторства экспертом: существенных временных затрат и возможной необъективности точки зрения эксперта.

На данный момент существует множество средств автоматического определения авторства текста, однако в большей их части под определением авторства понимается выбор автора некоторого текста из коллекции известных программе авторов [3, 6]. Также широко применимы средства, которые могут разделить коллекцию текстов на несколько коллекций меньшего объема, содержащих похожие тексты, но не обязательно соответствующих определенным авторам [4, 11].

Выбор автора из списка авторов, известных программе, позволяет достаточно точно определять авторство в случае тестирования программы на текстах известных ей авторов, однако результаты работы программы на текстах неизвестных авторов могут быть некорректными. Очевидным решением является снабжение программы очень большой коллекцией текстов разных авторов в качестве входных данных, однако это существенно замедляет работу программы и все равно не исключает возможности появления во входных данных программы текста автора, который программе не известен.

В качестве примера некорректного определения авторства текста можно привести ситуацию, в которой тексты неизвестных программе авторов распределяются по категориям уже существующих авторов. Следует отметить, что в случае позднего обнаружения подобной ошибки затраты на ее устранение могут оказаться критическими. Для иллюстрации приведем систему для компьютерной разметки большого и постоянно пополняющегося текстового корпуса, например – корпуса книг в большой библиотеке, которая автоматически определяет авторов поступающих в нее текстов. Для корректного поиска по корпусу важным критерием разметки является авторство каждого текста. Если данные обрабатываются неправильно, то тексты, авторы которых не были известны системе в начале ее работы, будут помечены как принадлежащие уже наличествующим авторам. Это сделает невозможным эффективный поиск в корпусе с использованием созданной разметки. Когда ошибка обнаружится, потребуется перезапускать систему и заново производить разметку корпуса, предварительно расширив множество известных системе обучающих текстов.

## 1. Постановка задачи

Для повышения вероятности корректной обработки текстов необходимо минимизировать возможность отнесения текста нового автора к категории уже существующего автора, а также ввести механизмы для обработки схожих между собой текстов, не принадлежащих известным программе авторам.

Таким образом, **целью** данной работы стало повышение точности автоматизированного определения авторства текстов соответствующими

программными средствами за счет разработки нового метода обработки текстовых данных, которым предусматривается добавление текстов к категориям известных программе авторов и автоматическое выделение категорий для текстов авторов, неизвестных программе.

В данном исследовании не берутся во внимание такие характеристики текста, как его смысловая и коммуникативная структура, а также принадлежность текста к какому-либо литературному стилю. То, как эти и другие подобные характеристики влияют на качество автоматического анализа текста, описано в [8]. В данной же работе используются формальные показатели принадлежности текста тому или иному автору, прежде всего – набор употребляемых конкретным автором слов. Также рассматривается применение комбинированных признаков, выделение составных термов и стоп-слов, методов уменьшения пространства признаков и поиска наиболее информативных признаков.

В соответствии с поставленной целью **задачами исследования** являются:

– изучение методов обработки текстовых данных (в том числе методов классификации и кластеризации), применяемых для определения авторства последних;

– разработка нового метода определения авторства текстов для повышения эффективности работы соответствующего программного обеспечения по показателю корректности отнесения заданного текста к работам того или иного автора;

– формулировка алгоритма, реализующего предложенный метод;

– анализ результатов исследования эффективности нового метода определения авторства текстов.

## 2. Обзор методов определения авторства текстов

На данный момент существует множество разнообразных подходов к решению задачи определения авторства текстов. Большинство таких подходов базируется на использовании методов классификации [1, 3, 4] или кластеризации [6, 7] текстовых данных. Рассмотрим особенности применения этих методов к задаче определения авторства текстов, их преимущества и недостатки, различия, некоторые детали реализации. Цель данного этапа – определить, могут ли средства классификации и кластеризации быть применены к поставленной задаче, и найти специфические особенности применения.

### 2.1. Применение методов классификации к задаче определения авторства

Классификация документов, или классификация с учителем, — одна из задач информационного поиска, заключающаяся в отнесении документа к одной из нескольких заранее заданных категорий на основании содержания документа [5].

Автоматическая классификация может осуществляться на основе правил (созданных вручную) или с помощью обучения алгоритма по примерам.

**Недостатки** применения методов классификации для задачи определения авторства:

1. Первый недостаток всех методов классификации кроется в самом определении: количество и параметры категорий, на которые делятся тексты (в случае определения авторства – количество и список авторов), задаются заранее, и определяется принадлежность текстов только к этим определенным категориям. Таким образом, для любого, даже незначительного изменения списка авторов, необходимо выполнить существенный объем работы. В случае обучения по примерам требуется заново выполнить обучение классификатора, что требует значительных ресурсов (как вычислительных, так и временных). Для методов на основе правил требуется заново сформулировать правила, что требует работы экспертов. В случае сильных изменений категорий может требоваться критически длительное время на обучение классификатора либо работу экспертов.

2. Невозможность автоматического определения новых категорий. В результате выполнения классификации новый текст будет отнесен к заданным заранее $n$ категориям. Классификатор может оперировать вероятностями отнесения к категории, а значит, текст будет отнесен к $n$ категориям с определенными вероятностями $p_i$, где $i = 1..n$ – индекс соответствующей категории; $n$, как и $p_i$, может быть равно 1. В случае, когда вероятность принадлежности текста к каждой категории ниже некоторого порогового значения, текст может быть не отнесен ни к одной категории ($n$ может быть равно 0) – это зависит от реализации алгоритма. Но методы классификации не предусматривают возможности добавления новой категории в случае наличия большой коллекции похожих текстов во входных данных. Новые тексты могут быть отнесены к одной из существующих категорий, могут – к разным категориям, могут остаться некатегоризированными.

**Преимущества** применения методов классификации для задачи определения авторства. Безусловным плюсом применения методов классификации является стабильно высокое значение численных оценок качества классификации – к примеру, точности и $F$ -меры, при низком значении ошибки классификации [8, 9]. Точность выше в случае использования классификации на основе задаваемых вручную правил, однако правила никогда не являются универсальными, и для добавления новых авторов правила требуется переписать. В связи с этим средства классификации на основе правил широко применяются в автоматической сортировке/фильтрации новостных сообщений (где категории практически неизменны), однако для задачи определения авторства (набор авторов каждый раз разный) эти средства малопригодны.

Наибольшая эффективность методов классификации при условии, что наборы категорий могут отличаться, достигается за счет обучения алгоритма по примерам. В этом случае для корректного определения авторства необходимо, чтобы на вход алгоритма подавалась коллекция документов, для которых авторы уже известны. При обучении по примерам классификатору достаточно получить коллекцию документов и значений автора для этих категорий в качестве входных данных.

**Численные характеристики качества классификации.** Рассмотрим поднятый нами вопрос о численных оценках качества классификации. Приведем основные параметры, по которым производится оценка. Терминология приведена в соответствии с работой В.Г. Васильева и М.П. Кривенко «Методы автоматизированной обработки текстов» [9].

Пусть $C^0 = (c_{ij}{}^0)_{k\times n}$ – матрица эталонной классификации (может быть получена путем экспертной классификации объектов) размера $k \times n$ объектов из множества $X$ по классам из множества $W$, где

$$c_{ij}{}^0 = \begin{cases} 1, x_j \in w_i \\ 0, x_j \notin w_i \end{cases} \quad (1)$$

$C = (c_{ij})_{k\times n}$ – матрица оцениваемой классификации (например, получена путем автоматической классификации объектов) размера $k \times n$, где

$$c_{ij} = \begin{cases} 1, x_j \in w_i \\ 0, x_j \notin w_i \end{cases} \quad (2)$$

Для оценки качества классификации обычно производится вычисление различных мер, характеризующих степень близости оцениваемой классификации объектов $C$ к эталонной классификации $C^0$.

При обработке текстовых данных наибольшее распространение получил подход, при котором сначала вводятся показатели качества классификации по отношению к отдельным классам, а затем на их основе уже строятся обобщенные показатели для всей совокупности классов.

Пусть зафиксирован некоторый класс

$$w_i, i = 1,...,k .$$

Результаты совместной классификации объектов $x_i,...,x_n$ по отношению к данному классу, задаваемые с помощью матриц $C^0$ и $C$, можно представить в виде табл. 1.

**Таблица 1**

| | $T_i$ | $F_i$ |
|---|---|---|
| $T_i^0$ | $n_{TT_i}$ | $n_{TF_i}$ |
| $F_i^0$ | $n_{FT_i}$ | $n_{FF_i}$ |

В табл. 1 используются следующие сокращения: $T_i$ – множество объектов, отнесенных к классу $w_i$ в классификации $C$; $F_i$ – множество объектов, не отнесенных к классу $w_i$ в классификации $C$; $T_i^0$ и $F_i^0$ – множества объектов, которые отнесены и не отнесены к классу $w_i$ в классификации $C^0$, соответственно:

$$n_{TT_i} = \left|T_i^0 \cap T_i\right|, n_{TF_i} = \left|T_i^0 \cap F_i\right|, n_{FT_i} = \left|F_i^0 \cap T_i\right|,$$
$$n_{FF_i} = \left|F_i^0 \cap F_i\right| \quad (3)$$

Наибольшее распространение получили следующие показатели:

*Точность классификации* (от англ. Precision)

$$P_i = \frac{n_{TT_i}}{n_{TT_i} + n_{FT_i}} \quad (4)$$

представляет собой процент объектов, правильно отнесенных к классу $w_i$ в классификации $C$, по отношению к общему количеству объектов, отнесенных к классу $w_i$ в классификации $C$.

*Полнота классификации* (от англ. Recall)

$$R_i = \frac{n_{TT_i}}{n_{TT_i} + n_{TF_i}} \quad (5)$$

представляет собой процент объектов, правильно отнесенных к классу $w_i$ в классификации $C$, по отношению к общему количеству объектов, отнесенных к классу $w_i$ в классификации $C^0$.

Точность и полнота классификации характеризуют различные стороны оценки качества и их нельзя использовать независимо. Например, если все объекты отнести к классу $w_i$, то в этой ситуации полнота будет равна 1, а точность классификации очень низкой. Наоборот, если к классу $w_i$ отнести только один правильный объект, то точность будет равна 1, а полнота будет близкой к 0.

Для возможности сравнения качества работы различных систем друг с другом удобно использовать один показатель, а не несколько. По этой причине при оценке качества классификации часто используется комбинированный показатель, который называется $F$-мера.

Он определяется следующим образом:

$$F_i(\lambda) = \left[\lambda \frac{1}{P_i} + (1-\lambda)\frac{1}{R_i}\right]^{-1}, \quad (6)$$

где $\lambda \in [0,1]$. Заметим, что при $\lambda = 0$ $F_i(0) = R_i$, при $\lambda = 1$ $F_i(1) = P_i$, при остальных значениях $\lambda \in (0,1)$ показатель $F_i(\lambda)$ является комбинацией точности и полноты. Обычно на практике используется значение $\lambda = 0{,}5$, в данном случае для $F$-меры используется обозначение $F_i \equiv F_i(0{,}5)$.

Помимо точности, полноты и $F$-меры также используется такой обобщенный показатель как ошибка классификации. Он определяется следующим образом:

$$E = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n}\left|c_{ij}{}^0 - c_{ij}\right|. \quad (7)$$

**2.2. Применение механизмов кластеризации для определения авторства**.

Кластеризация документов, или классификация без учителя, — одна из задач информационного

поиска, целью которой является автоматическое выявление групп семантически похожих документов среди заданного фиксированного множества документов [5]. Следует отметить, что группы формируются только на основе попарной схожести описаний документов, и никакие характеристики этих групп не задаются заранее, в отличие от классификации документов, где категории задаются до начала выполнения алгоритма.

Средства кластеризации позволяют выделить подмножества похожих текстов из общего множества обучающих текстов. Количество подмножеств может быть как фиксированным, так и вычисляемым в зависимости от близости текстов. Безусловным плюсом применения средств кластеризации является то, что количество кластеров может меняться по ходу добавления в систему новых текстов. При пересчете кластеров на основании множества имеющихся в системе текстов новые кластеры могут добавляться к уже существующим, а те, в свою очередь, могут делиться на части или объединяться.

Кластеризация может быть применена для визуализации авторского стиля и схожести стилей авторов, для определения в творчестве автора группы текстов, несхожих с остальными текстами автора [11, 12].

Кластеризация применяется и для определения авторства, но в этом случае речь идет не об определении автора текстов из группы заданных авторов, а о выделении групп похожих текстов, которые могут принадлежать одному автору. Кластеризация в чистом виде не очень широко используется в определении авторства из-за невозможности явного (предварительного) задания категорий, как в случае классификации с учителем.

Кластерный анализ также может быть применен для определения принадлежности текста к уже существующим (явно заданным либо определенным классификатором) категориям: категории трактуются как кластеры, и новый документ добавляется в один из этих кластеров.

Однако какие преимущества дает нам применение кластерного анализа, если по определению сам алгоритм классификации занимается определением категорий новоприбывших документов, к тому же применение алгоритма кластеризации само по себе не учитывает задание конкретных авторов для текстов? Ответ на этот вопрос будет рассмотрен далее.

### 3. Комбинированный метод определения авторства текста

Просуммировав вышесказанное, можем сделать выводы о том, что:

1. В большинстве случаев для автоматического определения авторства используются методы классификации.

2. Недостатки применения классификации – работа классификаторов только на заранее заданных категориях, невозможность автоматического задания новых категорий. Преимущество – высокое значение численных оценок качества.

3. Кластеризация применима для определения групп схожих по стилю текстов.

Проанализировав использование методов классификации и кластеризации, можем сделать вывод, что возможным путем добиться корректной обработки текстов, авторы которых заранее не заданы программе, является автоматическое введение новых категорий для авторов, неизвестных программе, с возможностью группировать похожие входные данные и причислять их к новой категории. Опишем, каким образом это реализуется и что позволило нам сделать такие выводы.

Из результатов анализа применения средств классификации видим, что для определения автора текстов эффективными и широко применимыми являются средства классификации, также известные как средства классификации с учителем. Однако методы классификации дают качественные и корректные результаты только в случае, если входные данные принадлежат к заранее заданным категориям, на которых обучался алгоритм. В случае определения авторства невозможным является снабжение алгоритма входными данными для обучения, которые включают тексты всех возможных авторов. В таком случае классификатор, получив на вход тексты неизвестных ему авторов, либо припишет эти тексты одному из существующих авторов, что является ошибкой и может повлечь ошибки в дальнейшем, либо же оставит текст некатегоризированным.

В случае, если классификатор оставляет часть текстов некатегоризированными, логичным является введение в его состав средств, производящих анализ этих текстов и их объединение в новые категории в случае выявления их схожести. В этом случае применимы средства кластеризации (классификации без учителя), которые дают возможность анализировать некатегоризированные тексты и выделять из них кластеры в случае схожести этих текстов.

Таким образом, для достижения поставленной цели данного исследования предлагается создать комбинированный метод определения авторства текста, в котором недостатки использованных методов классификации будут нивелированы путем введения элементов кластеризации.

Предлагаемый метод подразумевает выполнение следующих этапов:

1. Подготовительные этапы:

a. Выбор критериев классификации.

b. Выбор способа представления входных данных.

c. Подготовка обучающего множества текстов, для которых известны и заданы автора.

d. Выбор метода классификации и его модификация: тексты могут быть не причислены ни к одной категории.

e. Обучение классификатора.

f. Выбор или разработка метода выделения кластеров среди группы текстов.

2. Основные этапы:

a. Проведение классификации группы текстов по категориям известных программе авторов.

b. Применение алгоритма кластеризации к текстам, не причисленным к известным авторам, и формирование новых категорий на основе результатов работы алгоритма кластеризации.

### 4. Обобщенный алгоритм реализации метода

В предыдущем пункте статьи был представлен новый метод определения авторства текстов. Сформулируем теперь обобщенный алгоритм определения авторства текста, который реализует предложенный метод.

Подготовительные этапы – действия, которые необходимо выполнить, для того чтобы алгоритм мог конкретно определять авторов текстов коллекции.

Основные этапы – это те действия, которые необходимо выполнить над текстом или коллекцией текстов для определения авторов.

#### 4.1. Подготовительные этапы

Первый подготовительный этап **– выбор критериев классификации.** Критерии классификации – те свойства анализируемых документов, которые будут учитываться при анализе [2, 8, 10]. В качестве критериев классификации обычно принимается наличие в документах определенных термов. Однако в литературе существует много других подходов к определению критериев классификации [2]. Также стоит отметить, что для задачи определения авторства критерии классификации могут значительно отличаться от тех критериев, которые используются в общей задаче классификации.

Некоторые подходы к выбору критериев классификации:

1. Критериями классификации могут быть термы документов.

2. Для уменьшения размера пространства критериев и для повышения точности классификации можно пользоваться составными термами. Ими могут быть устойчивые словосочетания. Тогда имеет смысл пользоваться словарями для выявления таких словосочетаний. Также важными индикаторами авторского стиля являются свойственные конкретным авторам словосочетания. Однако создание обобщенного алгоритма выявления авторских словосочетаний представляется нетривиальной задачей.

3. Для уменьшения пространства термов имеет смысл не включать в число термов чересчур часто используемые слова, а также стоп-слова. Примером часто используемых слов для русского языка являются слова «быть», «делать», «красивый»; наиболее часто используемые стоп-слова – предлоги, союзы, местоимения.

4. Именно для задачи определения автора естественно рассматривать лишь те характеристики текстов, значения которых, переходя от объекта к объекту, выявляют наибольшую изменчивость. То есть, помимо выделения стоп-слов рационально выполнить выделение слов, которые могут встречаться не очень часто, но частота их использования характеризует стиль определенного автора.

5. Вместо исходных, непосредственно измеренных, признаков целесообразно рассматривать меньшее количество новых признаков, которые являются производными от исходных. Создание комбинированных признаков может существенно уменьшить количество критериев классификации, что, в свою очередь, может сделать классификацию как более быстрой из-за небольшого количества критериев, так и более точной, потому что в созданных комбинированных признаках будут учитываться наиболее значимые признаки. На практике эффективные алгоритмы выделения комбинированных признаков недостаточно изучены, потому широко применяются эвристические методы.

6. Если стандартно в алгоритмах классификации имеет значение набор слов каждого текста и его соотношение с набором слов всех текстов, то для определения авторства существует много не менее важных критериев. Например, важными являются знаки препинания и частота их использования, порядок следования слов и предложений, длина предложения.

Выделению критериев классификации посвящено множество научных работ [2, 8, 10]. Наша задача – выбрать такие критерии, которые позволят классифицировать тексты с точностью, не ниже заданной, однако их введение не повлечет значительного усложнения алгоритма.

В качестве критериев классификации предлагается использовать термы документа. Также необходимо провести работу по уменьшению пространства термов – удаление стоп-слов и термов, частота использования которых в тексте выше некоторого экспериментально подобранного порогового значения. Кроме того, предлагается использование словаря, содержащего устойчивые словосочетания, с целью уменьшения пространства термов и увеличения точности классификации за счет использования составных термов. Таким образом, формируется $m$ критериев классификации. Например, $i$-й критерий при $i = 1..m$ означает наличие в тексте некоторого простого либо составного терма.

Выбор критериев классификации обусловлен, прежде всего, требованиями к быстродействию: весь алгоритм определения авторства должен

выполняться быстро, что позволит оперативно произвести необходимую корректировку параметров. Усложнение критериев классификации замедлит процесс их формирования, но может увеличить точность результирующей классификации; уменьшение пространства термов способно ускорить процесс классификации, но может повлечь потерю точности.

Второй подготовительный этап алгоритма – **представление входных данных.** Этап подразумевает выполнение всех действий, которые необходимо выполнить над текстами естественного языка для преобразования их во входные данные алгоритма классификации. Входные данные должны отображать соответствие текстов критериям, о которых речь шла ранее.

Традиционно входными данными алгоритмов классификации является либо вектор термов документа, либо множество термов (так называемый "мешок слов", bag of words) [8, 9]. Однако, как уже было сказано, может быть проведена работа по уменьшению пространства термов, по выделению многословных термов, по удалению стоп-слов и т.д., результатом которой будет выделение критериев классификации [2, 8, 9].

В рамках предлагаемого алгоритма остановимся на простейшем и наиболее широко применяемом случае: представление документа – вектор документа в пространстве выбранных на предыдущем этапе критериев классификации, значения элементов вектора нормированы. Например, находятся в диапазоне , 0 – критерий в документе отсутствует, 1 – критерий присутствует либо частота его применения максимальна.

Для каждого текста его векторным представлением является вектор из $m$ элементов, $i$ -й элемент вектора означает наличие (частоту употребления) или отсутствие в тексте $i$ -й характеристики, или же соответствие $i$ -му критерию.

Третий подготовительный этап – **подготовка обучающего множества текстов.** Для построения классификатора, обучающегося на примерах, в первую очередь необходимо обеспечить наличие коллекции текстов для обучения классификатора. Коллекция должна быть достаточно обширной, и для всех текстов этой коллекции должны быть заданы авторы. Это позволяет достичь двух целей: дает алгоритму возможность собрать сведения про стиль каждого автора, а также дает возможность автоматического тестирования классификатора.

Четвертый подготовительный этап – **выбор метода классификации**. В классификации и кластеризации для определения схожести документов применяются некоторые метрики. Для примера рассмотрим простейшую метрику: близость между текстами – это расстояние между векторными представлениями этих текстов в пространстве выбранных критериев классификации. Чем меньше это значение, тем более близкими друг к другу мы считаем тексты, а для близких текстов можно говорить об одинаковом авторе. Еще раз подчеркнем, что перед нами стоит задача выделения именно таких характеристик текстов, которые были бы существенными для определения авторства.

В действующих системах классификации для определения схожести документов применяется ряд стандартных методов: например, Байесовский метод классификации (Naive Bayes classifier), метод опорных векторов (Support Vector Machines), деревьев принятия решений (Decision Trees), $k$-ближайших соседей (k-Nearest Neighbors) и т.д.

Для реализации этого этапа в алгоритме предлагается использование метода опорных векторов [9], поскольку этот метод имеет большую точность и легко реализуется программно. Метод работает с векторными представлениями текстов, однако определение близости текстов происходит сложнее, чем в примере, описанном выше.

Пятый подготовительный этап – **обучение классификатора.** Алгоритмом обучения классификатора предусматривается деление всего множества входных данных (заранее отобранных текстов с заданными авторами) на обучающее и проверочное множества текстов. Алгоритм обучается с помощью части наличествующих у нас текстов – обучающего множества. В качестве входных данных алгоритм получает векторные представления текстов и информацию об их принадлежности авторам; на основании этих данных формируются правила отнесения документов к определенному автору. Далее алгоритм тестируется (экспертом либо же программно) на проверочном множестве текстов. Отслеживается, правильно ли тексты проверочного множества были отнесены к категориям, в случае большого числа случаев неправильного отнесения алгоритм классификации может быть пересмотрен. Например, могут быть изменены некоторые коэффициенты.

После выполнения этих действий классификатор способен получать на вход новые тексты (в соответствующем представлении) и относить их к какой-либо из заранее заданных рубрик с определенной вероятностью.

Ясно, что в задаче определения авторства алгоритм не должен относить документ более чем к одной категории; введение возможности соавторства чрезмерно усложнит алгоритм. Однако нетрудно ввести в алгоритм модификацию: если документ не относится к числу документов обучающей выборки, и программа относит его к каждой категории с вероятностью менее какого-то порогового значения, документу присваивается "неопределенная" категория. Суть этой модификации в том, что каждый документ теперь не обязан быть отнесен к какой-то категории. Добавляются в существующие категории те документы, которые

явно (то есть, с достаточно большой вероятностью) принадлежат к этим категориям. Те же документы, принадлежность которых к какой-либо категории спорна, остаются в "неопределенной" категории. Они могут быть категоризированы на следующем этапе работы алгоритма. Пороговое значение следует подбирать экспериментально, оно будет зависеть от выбранных критериев классификации и схожести. Для определения принадлежности документа к категории используются критерии схожести, определенные на подготовительном этапе алгоритма.

И, наконец, последний, шестой подготовительный этап – **выбор метода выделения кластеров**. Существует множество алгоритмов кластеризации текстовых данных [8, 9]. На вход алгоритма поступает коллекция текстов, результатом работы алгоритма является эта коллекция с выделенными кластерами. В общем случае алгоритм кластеризации выполняет разделение некоторого множества документов на заранее не заданные категории, анализируя содержимое документов. Количество категорий может быть как фиксированным, так и определяемым алгоритмом. Подбирается метод кластеризации, работающий с нефиксированным количеством кластера, с возможностью задания минимального размера кластера.

### 4.2. Основные этапы алгоритма определения авторства текста

**Проведение классификации группы текстов по категориям известных программе авторов.** После проведения подготовительных этапов обученный классификатор может определять автора определенного текста. На этом этапе на вход алгоритма подается группа текстов, на выходе части текста присвоены категории известных программе авторов, часть текстов остается не категоризированной. Из не категоризированных текстов могут быть выделены новые категории в ходе выполнения следующего этапа алгоритма.

**Применение алгоритма кластеризации к текстам, не причисленным к известным авторам,** и формирование новых категорий.

Для обработки документов, которым не была присвоена категория в процессе классификации, используется механизм кластеризации.

В нашем случае алгоритм определяет схожесть документов по критериям схожести, подобным тем, которыми мы воспользовались при классификации (можно использовать как тот же набор критериев схожести, так и некоторый другой). Кластеризация выполняется после классификации, ее цель – выделить новые категории из документов, которые не могли быть отнесены к существующим категориям из-за недостаточной близости к ним. Нет нужды выполнять кластеризацию после классификации (успешной или нет) каждого документа. Кластеризация может выполняться после подачи $n$ документов на вход классификатора, после определенного промежутка времени, при наличии определенного количества документов с "не определенной" категорией.

На вход алгоритма кластеризации поступают документы с "не определенной" категорией. Алгоритм анализирует близость документов друг к другу с точки зрения критериев схожести, и создает новые категории в случае нахождения групп близких документов (документов, схожесть которых выше некоторого порогового значения).

Два основных вопроса в данном случае – критерии схожести документов: набор используемых признаков, размер создаваемых категорий. Например, может случиться так, что один текст одного писателя будет очень похож на один текст другого. В случае если каждый из этих текстов будет определен как "не похожий" на остальные тексты писателя, можно ли говорить о том, что эти тексты написаны одним человеком? Очевидным кажется решение ввести минимальный размер создаваемой категории: например, можно говорить о появлении нового автора, если для не менее чем $k$ (например, при $k = 10, 20, 100$) текстов, которые классификатор пометил как тексты неопределенной категории, их попарная близость выше заданного ранее порогового значения. В таком случае этим документам присвоится новая категория, и на следующем шаге алгоритм классификации будет учитывать также эту категорию при категоризации новых документов, поступающих в алгоритм.

Для настройки поведения алгоритма выполняется тестирование. **Тестирование** алгоритма включает наблюдение эксперта за ходом выполнения алгоритма и корректировку параметров алгоритма в случае необходимости. Как при работе алгоритма, так и при тестировании эксперт может подтвердить, что новая категория была создана корректно, и назвать ее именем автора этих текстов. В случае некорректного определения новых категорий следует изменить пороговое значение, критерии схожести документов или признаки, по которым мы производим классификацию. Тестирование подразумевает выполнение алгоритма на некоторых входных данных и корректировку параметров алгоритма – критериев для классификации и кластеризации, размерности пространства критериев, размера создаваемых рубрик, метрик схожести. Также в рамках тестирования оценивается корректность работы алгоритма по некоторым заранее заданным критериям.

## 5. Исследование эффективности комбинированного метода определения авторства текстов

Алгоритм позволяет повысить значение параметров качества классификации, например точности, полноты, $F$-меры, и снизить значение

ошибки классификации по сравнению с применением стандартных алгоритмов к задаче определения авторства. Рассмотрим значение численных характеристик качества классификации для предложенного алгоритма.

В табл. 1 были рассмотрены характеристики для оценки качества алгоритмов классификации. Эти характеристики обозначены как $n_{TT_i}$, $n_{TF_i}$, $n_{FT_i}$, $n_{FF_i}$. В дальнейшем будем считать, что это оценки для алгоритма определения авторства, использующего средства классификации без добавления средств кластеризации.

На данном этапе то, какой алгоритм классификации мы рассматриваем, не принципиально, поскольку мы ставим цель показать улучшение значений оценки качества вследствие введения кластеризации в рассмотренный алгоритм. Будем считать, что алгоритм классификации классифицирует каждый текст из входных текстов, и присваивает ему одно значение автора. Это верно для преимущественного большинства случаев применения алгоритмов классификации для определения авторства [1, 3, 4]

Введем аналогичные рассмотренным ранее обозначения для созданного нами алгоритма: $n_{TT_i}^*$, $n_{TF_i}^*$, $n_{FT_i}^*$, $n_{FF_i}^*$, а также производные от них $P_i^*$, $R_i^*$, $F_i^*$, $E_i^*$. Параметры $P_i^*$, $R_i^*$, $F_i(\lambda)^*$ характеризуют категорию, для характеристики всего классификатора в простейшем случае используется среднее арифметическое параметров категорий.

Рассмотрим значения этих параметров и изменение численных характеристик качества классификации:

1. Для категорий текстов, авторы которых известны программе в начале ее работы.

Введение средств кластеризации позволяет уменьшить число текстов, некорректно отнесенных к категории какого-то автора. В случае использования только средств классификации в категорию, соответствующую определенному автору, попадут и тексты, действительно принадлежащие автору, и часть текстов, которые принадлежат не известным программе авторам и были ошибочно классифицированы. Используя символьные обозначения,

$$n_{FT_i} > n_{FT_i}^* . \qquad (8)$$

Это изменение улучшает значение точности и *F*-меры. Подставив в формулы (4) и (6) нововведенные обозначения и беря во внимание свойство (8), имеем:

$$P_i = \frac{n_{TT_i}}{n_{TT_i} + n_{FT_i}} < \frac{n_{TT_i}^*}{n_{TT_i}^* + n_{FT_i}^*} = P_i^* ; \qquad (9)$$

$$F_i \equiv F_i(0,5) = \left[ \frac{1}{2 * P_i} + \frac{1}{2 * R_i} \right]^{-1} < \left[ \frac{1}{2 * P_i^*} + \frac{1}{2 * R_i^*} \right]^{-1}$$

$$F_i^* \equiv F_i(0,5)^* . \qquad (10)$$

Видим, что введение методов кластеризации позволило увеличить значение точности и $F$-меры для категорий, автора которых уже известны программе.

2. Для категорий текстов, авторы которых не известны программе в начале ее работы.

Без применения средств кластеризации $n_{TT_i}$ для всех категорий авторов, заранее не известных программе, было равно 0. Текст не мог быть отнесен к корректной категории, поскольку категория не существовала в начале работы программы и не могла быть создана позднее.

Представленный алгоритм выделяет категории для новых авторов, и часть категорий выделяется корректно. То есть, некоторое число текстов, присвоенных этой категории, в действительности принадлежат одному автору, и новосозданная категория соответствует этому автору. Для этих категорий $n_{TT_i}$ станет больше 0. В таком случае значение оценок классификации, имеющих $n_{TT_i}$ в числителе, возрастет; значение точности, полноты и *F*-меры станет ненулевым. Для некорректно введенных категорий $n_{TT_i}$ останется равным 0, что не повредит результирующим оценкам качества классификации.

Подставим в формулы (4), (5), (6) характеристики алгоритма классификации и созданного нами алгоритма:

$$P_i = \frac{n_{TT_i}}{n_{TT_i} + n_{FT_i}} = \frac{0}{n_{FT_i}} < \frac{n_{TT_i}^*}{n_{TT_i}^* + n_{FT_i}^*} = P_i^* ; \qquad (11)$$

$$P_i = \frac{n_{TT_i}}{n_{TT_i} + n_{TF_i}} = \frac{0}{n_{TF_i}} < \frac{n_{TT_i}^*}{n_{TT_i}^* + n_{TF_i}^*} = R_i^* ; \qquad (12)$$

$$F_i \equiv F_i(0,5) = \left[ \frac{1}{2 * P_i} + \frac{1}{2 * R_i} \right]^{-1} = 0 ; \qquad (13)$$

$$F_i^* \equiv F_i(0,5)^* = \left[ \frac{1}{2 * P_i^*} + \frac{1}{2 * R_i^*} \right]^{-1} . \qquad (14)$$

Сознательно избегаем деления на 0, потому считаем, что $F_i = 0$, а не $\infty$.

Видим, что введение методов кластеризации позволило увеличить ранее равные 0 значения точности, полноты и $F$-меры для корректно выделенных средствами кластеризации категорий.

Просуммировав вышесказанное, видим, что численные оценки улучшаются, поскольку минимизируется вероятность отнесения текста неизвестного автора к некорректной категории, соответствующей уже известному программе автору. Кроме того, с большой вероятностью средства кластеризации корректно создадут категорию и внесут туда текст, даже если он принадлежит автору, заранее не известному программе. Для стандартных же средств классификации такая вероятность равна нулю, и если автор не известен программе, то его

текст точно не будет отнесен к корректной категории. Из этого делаем вывод, что введение средств кластеризации позволяет улучшить значение численных характеристик качества классификации.

**Выводы**

В данной статье предложен метод определения авторства текстов, который сочетает в себе классификацию и кластеризацию документов. Данный метод эффективен в случае его использования в системе для установления факта авторства текстов с возможностью определения автора из заданных и автоматического добавления категорий для новых авторов. Первоначально тексты, представляющие входные данные, рассортированы по заданным категориям, но в дальнейшем при добавлении новых текстов средства кластеризации позволяют объединять добавленные тексты в новые категории. Нововведенные категории соответствуют группам текстов авторов, которых не было в первоначальном списке авторов. Этот подход рассмотрен подробней, также рассмотрены методы представления данных и выбора информативных признаков документа, которые позволяют точнее задавать новые категории. Проведено исследование эффективности созданного метода, которое показало, что метод позволяет достичь улучшения численных оценок качества классификации.

Предложен алгоритм реализации метода определения авторства с использованием средств классификации и кластеризации, который может быть реализован на ЭВМ.

Дальнейшего рассмотрения требуют отдельные параметры классификации. Также планируется создать программную реализацию алгоритма для корректировки параметров классификации и проведения тщательного тестирования.

**Список литературы. 1.** *Агеев, М.С.* Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов [Текст] : дис. ... канд. физ.-мат. наук : 05.13.11 / М.С. Агеев. — М., 2004. — 136 с. **2.** Автоматическое определение авторства [Электронный ресурс] : / Л. М. Пивоварова // Режим доступа: http://www.slideshare.net/lmp/09-9770049 - 20.01.2012 г. – Загл. с экрана. **3.** *Романов, А.С.* Методика и программный комплекс для идентификации автора неизвестного текста [Текст] : дис. ... канд. техн. наук : 05.13.18 / А.С. Романов. — Томск, 2009. — 149 с. **4.** *Шевелев, О.Г.* Разработка и исследование алгоритмов сравнения стилей текстовых произведений [Текст] : дис. ... канд. техн. наук : 05.13.18 / О.Г. Шевелев. — Томск, 2006. — 176 с. **5.** *Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze.* An Introduction to Information Retrieval Draft. Online edition. – Cambridge University Press, 2009. – 544 p. **6.** *Сидоров, Ю.В.* Математическая и информационная поддержка методов обработки литературных текстов на основе формально-грамматических параметров [Текст] : дис. ... канд. техн. наук : 05.13.18 / Ю.В. Сидоров Юрий Владимирович. — Петрозаводск, 2002. — 127 с. **7.** *Грушников, А.В.* Аутентификация произведений живописи по цифровым изображениям [Текст] / А.В. Грушников // Всероссийский журнал научных публикаций. – 2011. – Июнь. – С. 5-7. **8.** Автоматическая обработка текстов на естественном языке и компьютерная лингвистика [Текст] : учеб. пособие / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ [и др.]. — М.: МИЭМ, 2011. — 272 с. **9.** *Васильев, В.Г.* Методы автоматизированной обработки текстов [Текст ] / В. Г. Васильев, М. П. Кривенко. – М. :ИПИ РАН, 2008. – 305 с. **10.** *Стадник, А В.* Использование искусственных нейронных сетей и вейвлет-анализа для повышения эффективности в задачах распознавания и классификации [Текст] : дис. ... канд. техн. наук : 05.13.18 / А.В. Стадник. — Иваново, 2004. — 88 с. **11.** *Шмулевич, М.М.* Метод автоматической кластеризации текстов, основанный на извлечении из текстов имен объектов и последующем построении графов совместной встречаемости ключевых термов [Текст] : дис. ... канд. физ.-мат. наук : 05.13.17 / М.М. Шмулевич. — М., 2009. — 120 с. **12.** Обзор методов кластеризации текстовой информации [Электронный ресурс] : / К. М. Кириченко, М. Б. Герасимов // Режим доступа http://www.dialog-21.ru/Archive/2001/volume2 /2_26.htm - 20.01.2012 г. – Загл. с экрана.

***Поступила в редколлегию 18.11.2011***

УДК 004.912

**Комбінований метод автоматичного визначення авторства текстів** / І.В. Глушаускайте, Т.М. Заболотня // Біоніка інтелекту : наук.-техн. журнал. – 2012. – № 1 (78). – С. 102-110.

У даній роботі запропоновано метод автоматизованого визначення авторства текстів, який поєднує в собі застосування засобів класифікації та кластеризації. Розроблений метод дає можливість визначення тих авторів текстів, які не були знайдені після проведення класифікації. Подано узагальнений алгоритм реалізації методу, який може бути реалізований програмно. Дана оцінка ефективності запропонованого методу.

Табл. 1. Бібліогр .: 12 найм.

UDC 004. 912

**The combined method of automatic text authorship determination** /I.Glushauskaite, T. Zabolotnyaya// Bionics of Intelligense: Sci. Mag. – 2012. – № 1 (78). –P. 102-110.

In this work a method of automated text authorship determination which combines implementations of both the means of classification and clustering is proposed. The developed method enables attribution of the texts that couldn't be attributed to some authors during classification. A generalized algorithm for implementation of the method, which can be implemented programmatically, is presented. An estimation of the efficiency of the proposed method is given.

Tab. 01. Ref.: 12 items.