УДК 004.89, 004.91
DOI: 10.15827/0236-235X.030.3.478-486
Дата подачи статьи: 22.05.17
2017. Т. 30. № 3. С. 478–486

МЕТОД ЧАСТОТНО-МОРФОЛОГИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ

А.А. Осочкин, аспирант, osa585848@bk.ru;
В.В. Фомин, д.т.н., профессор, v_v_fomin@mail.ru;
А.В. Флегонтов, д.т.н., профессор, flegontoff@yandex.ru
(Российский государственный педагогический университет им. А.И. Герцена, наб. реки Мойки, 48, г. Санкт-Петербург, 191186, Россия)

Появление централизованных хранилищ данных и накопление в них информации в виде как структурированных таблиц, так и слабоструктурированных текстов стали следствием растущего внимания к методам анализа данных. Анализ данных в перспективе позволяет получать важную информацию, на основе которой можно принять верное управленческое решение или спрогнозировать дальнейшее развитие событий. Одним из важных направлений этого анализа является автоматическая классификация накопленных данных в электронном виде, упрощенная модель которой сводится к считыванию, обработке текста и присвоению документу темы из заранее заданного списка. Все чаще работы зарубежных коллег посвящаются классификации данных в области медицины для последующего прогноза развития болезни на основе статистики или постановки диагноза на основе истории болезни. Главную сложность в классификации представляют тексты на естественном языке, которые в силу лингвистических особенностей языка и поддержки частью методов классификации исключительно числовых данных трудно поддаются классификации.

В настоящей работе исследуется научная активность в сфере классификации данных на естественном языке на основе ежегодной публикации научных трудов в данной сфере, а также предлагается на рассмотрение метод классификации русскоязычных текстов, интегрирующий в себе алгоритмы частотного, морфологического и интеллектуального анализов.

Процедура классификации текстов предполагает применение частотных, морфологических показателей и регрессионных деревьев. Также в данной работе представлены результаты ряда экспериментов по идентификации метода классификации с наиболее высокой точностью. Классификация осуществлялась по функциональным, литературным и авторским стилям.

Ключевые слова: *классификация текстов, частотный анализ, морфологический анализ, деревья решений, data mining, text mining.*

В среде информационно-коммуникационных технологий и систем происходят устойчивый рост и накопление текстовой слабоструктурированной информации [1], увеличивается объем хранилищ данных (библиотек, банков данных, репозиториев и т.д.). Потребность в эффективном извлечении ценных знаний из текстовых массивов влечет за собой усложнение и появление новых методов обработки информации – интеллектуального анализа текстов (text mining), в том числе за счет применения ресурсоемких статистических алгоритмов, алгоритмов интеллектуального анализа данных (data mining) [2], семантического поиска, использования сетевых и интернет-технологий и т.д. Вследствие роста объема данных и времени их обработки из-за сложности алгоритмов растут затраты на повышение производительности вычислительной техники.

Задачей развития методов text mining является извлечение полезных знаний из информационных массивов с учетом особенностей обработки *естественного языка* (ЕЯ), в том числе классификация текстов, извлечение информации, реферирование, информационный поиск и т.д. [3, 4]. Методы text mining используются в различных программных и информационных технологиях и как отдельные приложения, библиотечные модули, и в составе инструментария интеллектуального анализа данных, систем бизнес-аналитики, корпоративного управления и т.д. Одной из ключевых задач text mining является классификация текстов на ЕЯ [5].

Рост объема массивов данных и потребность в эффективном извлечении из них ценной информации обусловливают усложнение и появление новых методов обработки информации [6], в том числе за счет применения ресурсоемких статистических алгоритмов, алгоритмов семантического поиска, нейронных и интернет-технологий и т.д. Из-за увеличения массивов информации, подлежащих обработке, и применения для анализа все более сложных и глубоких методов растут требования к вычислительным ресурсам.

В борьбе с растущей сложностью и увеличивающимися затратами на технологии обработки текстов прослеживается тенденция возврата к классическим методам частотно-морфологического анализа. Поисково-текстовые методы и алгоритмы [7] акцентируют внимание на попытке использования небольшого, минимального арсенала теоретико-лингвистических изысков и делают акцент на формальных методах статистической обработки упрощенных словоформ. Авторами статьи разработан специальный комбинированный метод классификации с целью решения вышеназванной проблемы. Он включает в себя этап извлечения из текста числового набора показателей, что позволяет применить для классификации методики data mining, тем самым расширяя спектр возможных методик, применяемых для классификации.

Ключевой особенностью комбинированного метода классификации является наличие коллек-

ции алгоритмов, позволяющей выбрать из нее наиболее эффективные для классификации анализируемого набора текстов. Коллекция алгоритмов не статична: появляющиеся новые алгоритмы классификации могут быть включены в комбинированную методику классификации текстов, что делает ее еще более универсальной.

Обзор исследований в области классификации данных на ЕЯ

Исследования в области классификации данных на ЕЯ начались в середине XX века. Динамика развития публикационной научной активности по классификации текстов проиллюстрирована на рисунке 1, где представлены результаты исследования количества научных работ, опубликованных в период с 1960 по 2008 гг. [8].

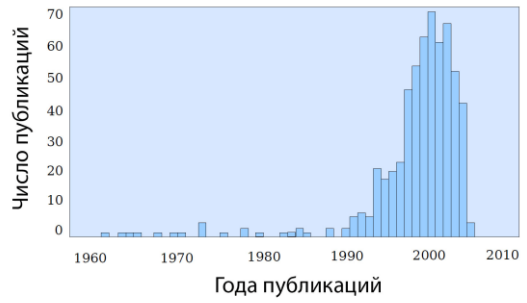

Рис. 1. Публикации в сфере text mining и обработки текстов на ЕЯ в 1960–2008 гг.

Fig. 1. Publications in text mining and natural language text processing in 1960–2008

На основе интернет-поисковика Base [9] был проведен обзор публикаций, книг и других работ в данной области в период с 2008 по 2016 гг. Результаты представлены на рисунке 2.

Для отбора работ использовались ключ «Классификация текстов» и фильтры classification и text mining.

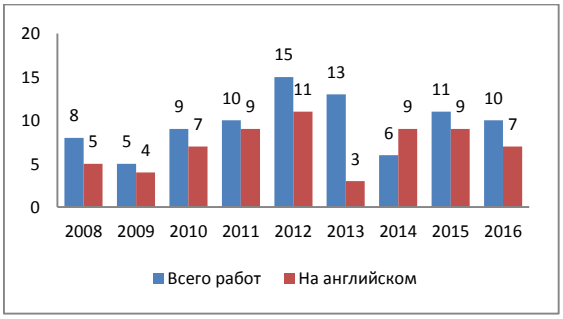

Рис. 2. Публикации в сфере text mining и обработки текстов на ЕЯ в 2008–2016 гг.

Fig. 2. Publications in text mining and natural language text processing in 2008–2016

При этом за период с 2008 по 2016 гг. учитывались не только изданные книги, но и публикации в различных известных журналах [9], находящихся в БД. На рисунке 2 в левых колонках отражено общее количество публикаций в данной сфере независимо от языка, в правых – количество публикаций только на английском языке. Средняя доля публикаций в период с 2008 по 2016 гг. на английском языке от общего числа публикаций за данный период составляет 76 %. Это говорит о том, что данное направление активно развивается англоязычными авторами и, как следствие, рассматривается классификация данных на английском языке.

Теперь сузим тему. Для этого выполним поиск работ по ключу text classification и используем два фильтра: classification и machine learning (рис. 3).

Рис. 3. Публикации в сфере text mining и обработки текстов на ЕЯ (английском) с 2008 по 2016 гг.

Fig. 3. Publications in text mining and natural language processing in English from 2008 to 2016

Графики на рисунке 3 отражают тенденцию спада интереса к данной задаче и, как следствие, количества работ в данной области с 2005 года.

В современных работах по классификации текстов авторы все чаще опираются на статистические способы классификации данных, полученных при помощи синтаксического и семантического анализа.

Классификация наборов текстов

В классификации текстовых данных при опоре на статистические методы важным этапом является процесс извлечения набора статистических показателей (рис. 4).

Набор показателей из данных на ЕЯ может быть получен при помощи различных процедур анализа, которые условно можно разделить на три вида: частотный, морфологический и морфологическо-семантический анализ.

Именно этап извлечения набора показателей является наиболее ресурсоемким и важнейшим фактором, влияющим на эффективность классификации в целом.

Поэтому для дальнейшего развития таких направлений text mining и обработки текстов на ЕЯ необходимо искать методы классификации не только высокоэффективные по точности, но и поз-

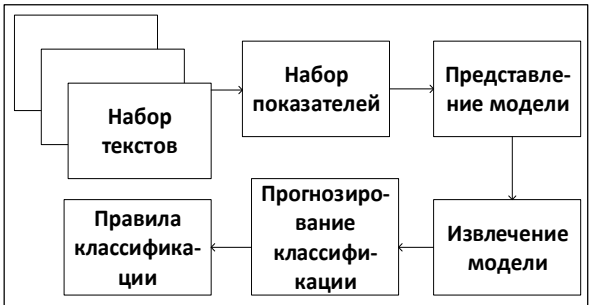

Рис. 4. Обобщенная схема классификации

Fig. 4. General classification scheme

воляющие сократить использование вычислительных ресурсов при классификации данных.

Сокращение ресурсоемкости при классификации данных на ЕЯ может быть достигнуто путем применения различных методов классификации, в том числе основанных на теории вероятности и распространенной теореме Байеса, а также упрощенных статистических мерах [10] (частота вхождения слова в текст). Эти два подхода демонстрируют проблематику баланса между ресурсоемкостью и точностью классификации.

Большой потенциал заложен в методах, основанных на деревьях решений (регрессионных деревьях). В классификации данных деревья решений используются не так часто, как другие методы. Это обусловлено вытеснением метода деревьев решений другими методиками классификации данных, особенно основанных на теории вероятности, которые в частных случаях показали более точный результат классификации.

Однако есть работы, экспериментально доказывающие эффективность метода дерева решений для классификации данных. Метод дерева решений основан на машинном обучении, поэтому для него обязательна учебная выборка. Значительным положительным эффектом деревьев решений является их способность снижать множество признаков, оставляя только значимые. Важным преимуществом деревьев решений является логический аппарат интерпретации и пояснения результата. Использование метода дерева решений позволяет классифицировать данные, с высокой точностью используя минимальный набор показателей, тем самым уменьшая ресурсоемкость классификации.

Рассмотрим гипотезу о том, что часть задач классификации текстов можно успешно осуществить с небольшими затратами вычислительных ресурсов путем манипуляции минимальным набором частотных характеристик, ограниченных формализованным набором ЕЯ и алгоритмами классификации. Решение данной задачи можно свести к применению различного рода алгоритмов морфологического, частотного анализа для получения минимального набора данных и к использованию эффективного метода классификации.

Уменьшение ресурсоемкости процесса возможно путем использования частотно-морфологического анализа, позволяющего получить минимальный набор теоретико-лингвистических показателей и статистически упрощенных словоформ, извлекаемых из текстов. При таком подходе уменьшается алгоритмическая и вычислительная нагрузка на процесс оценки и классификации текстов, но ограничивается интерпретация смыслового контекста извлекаемой информации.

Полученный в результате морфологического анализа теоретико-лингвистический набор показателей можно условно разделить на морфологические словоформы и их характеристики, а также синтаксические показатели.

Синтаксис – форма или структура выражений, предложений и программных единиц. Исторически сложилось так, что синтаксическая единица больше слова и представляет собой словосочетание или предложение. Извлечение синтаксиса из текста позволяет осуществлять синтаксический анализ текста и на его основе строить деревья синтаксического анализа, которые часто дублируются, так как предложения построены на общих синтаксических правилах ЕЯ.

Результаты ряда экспериментов [11] показали, что тексты в целом содержат не более 3–10 отличающихся друг от друга видов синтаксических деревьев. Их вариативность может быть объяснена множеством факторов [12], таких как использование различных стилей речи при написании текста (официальный, разговорный и т.д.), стилистическая особенность автора, которая выражается в нарушении грамматических правил, построения предложений и т.д. Таким образом, последовательность частей речи позволяет получить множество различных показателей, которые отображают не всегда очевидные, но характерные для какого-либо кластера характеристики.

Число последовательностей частей речи для каждого ЕЯ разное, так как число частей речи в них варьируется, но для большинства языков в среднем оно равно 10. Таким образом, количество различных цепочек частей речи может составлять 10^{10}. Большинство частей речи имеют индивидуальные характеристики, например, для существительных это падеж, одушевленное, неодушевленное и т.д. При учете индивидуальных характеристик еще больше обостряется проблема роста вычислительных ресурсов.

Комбинированный метод классификации текста

Комбинированный метод классификации текста – это комплекс применяемых алгоритмов, приемов, операций и инструментария к анализируемому набору текста для его классификации. В комбинированный метод заложены концепция

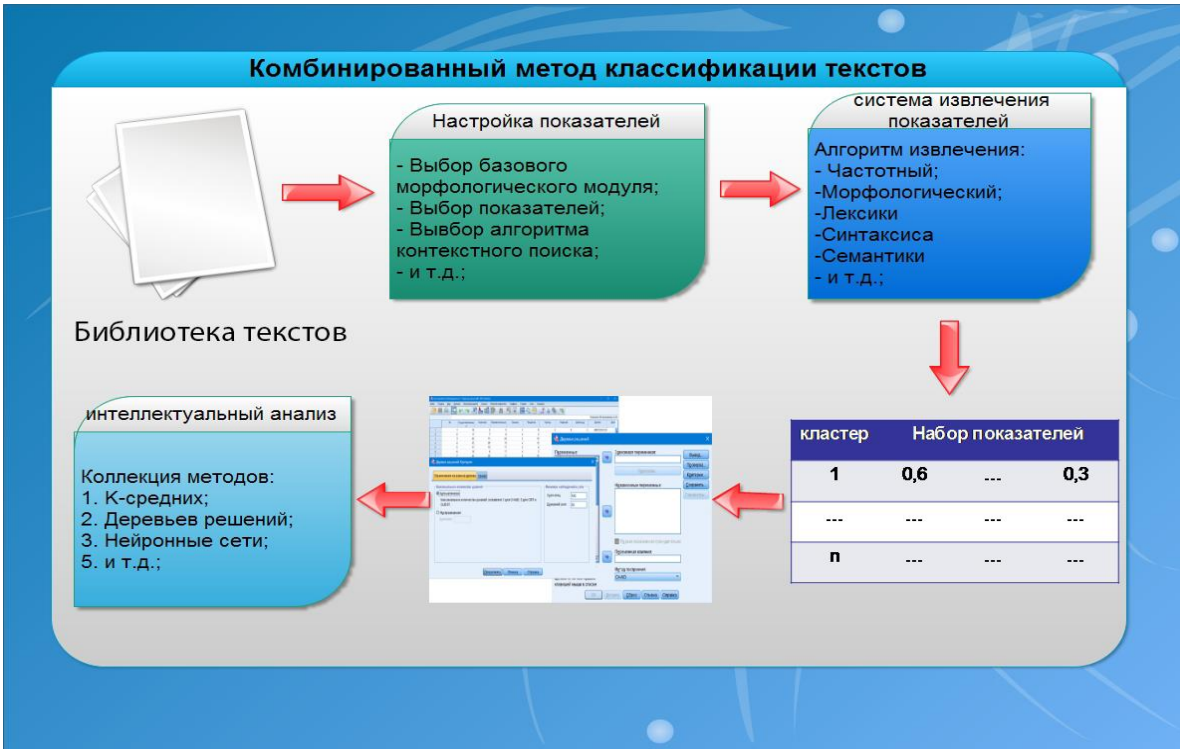

Рис. 5. Подсистемы комбинированного метода

Fig. 5. Combined method subsystems

минимальности действий для использования и замены модулей и компонентов, а также возможности расширения функций за счет привлечения сторонних программ и библиотек.

Система реализации text mining на базе комбинированного метода состоит из трех блоков (см. рис. 5).

1. Библиотека текстов – поиск, извлечение, каталогизация, хранение текстов. Предполагает развитую систему управления: навигация, добавление, удаление, просмотр, форматирование и пр.
2. Извлечение параметров – извлечение числовых параметров текста, морфологический и синтаксический анализ текста, ведение БД параметров и т.д.
3. Интеллектуальный анализ данных – обучение, кластеризация, идентификация объектов-текстов, формирование статистических данных, обучающей выборки в форматах алгоритмов интеллектуального анализа данных. Включает коллекцию алгоритмов для выбора наиболее эффективных методов классификации.

Ключевой особенностью комбинированного метода text mining является возможность выбора наиболее эффективного алгоритма data mining (из коллекции алгоритмов) для классификации анализируемого набора текстов. Коллекция алгоритмов не статична, и по мере развития исследований в нее могут быть включены новые алгоритмы классификации.

Описание программы

Для реализации комбинированного метода классификации было разработано специальное программное средство – Frequency and morphological analysis, или сокращенно FaM. Основная цель программы – извлечение при помощи частотного и морфологического анализа из анализируемых текстов информации о частоте употребления слов и их характеристиках, расчет на их основе коэффициентов для последующего анализа в интеллектуальных пакетах данных.

Для частотного и морфологического анализов используются интегрированные модули сторонних авторов, работающие только в 64-битных версиях Windows. Исходя из этого факта, анализ текста возможен только под управлением ОС Windows 64-битной версии.

Программа разрабатывается на объектно-ориентированном языке C#, среда разработки – Microsoft Visual Studio 2013 SP3. Этот выбор обусловлен отсутствием поддержки у некоторых модулей анализа text mining ОС Linux.

Результатом анализа текста является набор показателей, который содержится в БД программы, а также может быть продублирован в Excel для быстрого экспорта в интеллектуальные пакеты анализа данных.

На рисунке 6 представлен интерфейс программы версии 1.01. Цифрами на рисунке обозна-

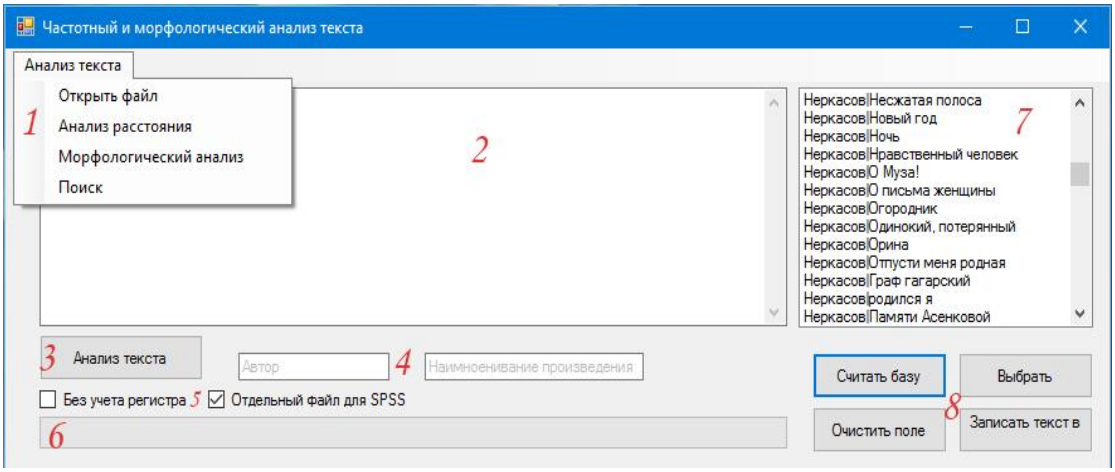

Рис. 6. Интерфейс FaM

Fig. 6. FaM interface

чено следующее: 1 – меню, позволяющие перейти к дополнительным видам анализа; 2 – основное поле работы, в которое пользователь может вводить текст вручную или импортировать из документов; 3 – кнопка для осуществления частотного и морфологического анализов; 4 – два поля для ввода автора и наименования произведения для удобства навигации по архиву, ввод необязателен; 5 – две опциональные настройки анализа: а) анализ с учетом регистра нижних и верхних букв, б) экспорт результатов анализа в специальный лист Excel, позволяющий сразу же импортировать его в SPSS; 6 – шкала прогресса, отображающая во время работы завершенность процедуры; 7 – список текстов в базе; 8 – кнопки для работы с базой (считать базу, выбрать, записать текст), очистить поле – быстрая очистка основного поля.

Реализация частотного анализа в комбинированном методе классификации текста

В комбинированном методе классификации текста наиболее сложными этапами являются частотный и морфологический анализы.

При частотном анализе была поставлена задача не только получить точные частотные данные, но и оптимизировать процесс анализа, уменьшив время, необходимое для контекстного поиска. Схема частотного анализа в комбинированном методе классификации текста изображена на рисунке 7.

Инициализация алгоритма частотного анализа начинается с процедуры считывания текста.

Базовыми в частотном анализе являются алгоритм подсчета всех символов, определения содержания иностранных слов в тексте и необходимость использования морфологических модулей для других языков.

Следующим этапом является очистка текста от спама, под которым подразумеваются скобки, кавычки, слитные с текстом знаки препинания и т.п. Данная процедура позволяет очистить слова от знаков препинания, тем самым повысив точность при поиске словоформ в морфологическом анализе. Чтобы не испортить оригинальный текст, создается дубликат, к которому применяется фильтрация от спама.

После очистки программа приступает непосредственно к частотному анализу, используя алгоритм контекстного поиска слов. Для обеспечения максимальной скорости определена процедура подключения алгоритма контекстного поиска из двух алгоритмов – Боейра–Мура или brute force. Выбор алгоритмов не случаен, он основан на результатах экспериментов по скорости поиска. Если объем текста превышает 800 символов, программа

Рис. 7. Частотный анализ в комбинированном методе

Fig. 7. A frequency response analysis in a combined method

автоматически использует алгоритм Боейра–Мура, если меньше – brute force.

Описанные этапы частотного анализа готовят текст для дальнейшего морфологического анализа.

Реализация морфологического анализа

Ключевым элементом комбинированного метода классификации текста является морфологический анализ, структура которого показана на рисунке 8.

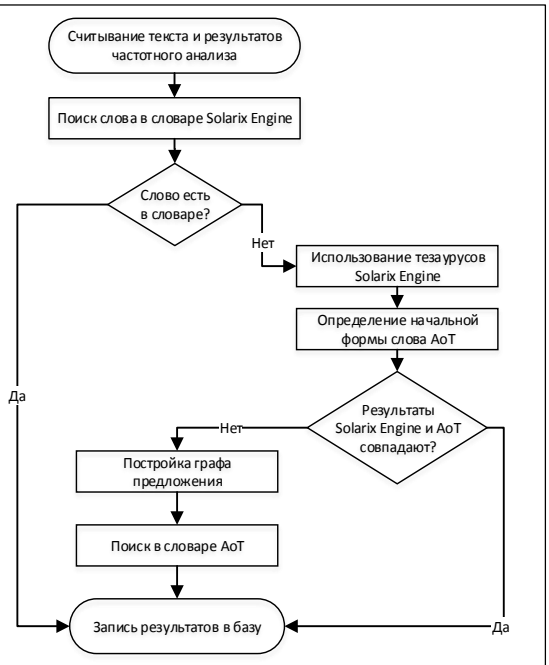

Рис. 8. Гибридный алгоритм в комбинированном методе

Fig. 8. A hybrid algorithm in a combined method

Морфологический анализ осуществляется специальным гибридным алгоритмом с использованием двух встроенных модулей – автоматической обработки текста, или AoT, и Solarix Engine. Морфологический модуль AoT [13] содержит русский морфологический словарь, включающий около 161 000 слов с различными формами, а также синтаксический и семантический анализ текста. Solarix Engine [14] – модуль морфологического анализа, включающий в себя словарь из 1 800 000 слов и 218 000 статей; тезаурус содержит информацию о возможных отношениях подчинения и ассоциативных отношениях между словами для машинного обучения.

В гибридном алгоритме осуществляется двойная проверка на часть речи, что позволяет получить более точные результаты. Если результаты двух модулей противоречат друг другу, строится граф предложения, в процессе построения идентифицируются взаимосвязи между словами и частями речи.

Таким образом, использование в морфологическом анализе двух морфологических модулей разных типов является одной из особенностей гибридного алгоритма.

Был проведен эксперимент по оценке точности морфологического анализа. Анализировались пять простых текстов длиной не более 200 слов, причем заранее было известно, какой частью речи является каждое слово. Точность анализа определялась по формуле $Qx = \frac{\sum X_i}{\sum Y_i}$, где X_i – количество слов, отнесенных морфологическим модулем к части речи i; Y_i – количество слов, которые действительно относятся к части речи i.

В таблице 1 показано, что в среднем общая точность двух модулей одинаковая. Стоит обратить внимание, что модуль AoT почти всегда ошибался при определении деепричастия, а модуль Solarix Engine – при анализе причастий и прилагательных. Однако при анализе всего предложения целиком (тезаурусом) точность повысилась на 9 %.

Таблица 1

Оценка точности модулей, %

Table 1

Module accuracy estimation, %

Текст	AoT	Solarix Engine	Solarix Engine с тезаурусами	AoT + Solarix Engine с тезаурусами
№ 1	76	76	82	86
№ 2	73	81	89	96
№ 3	71	74	79	93
№ 4	77	77	86	91
№ 5	79	68	82	88
Ср. рез.	**75**	**75**	**84**	**91**

Применение к 16 % слов, которые не смог идентифицировать Solarix Engine, метода AoT с лексимизацией (приведением слов в начальную форму) позволило повысить точность еще на 7 %.

Набор показателей

После получения в результате частотного и морфологического анализов набора числительных показателей алгоритм выполнения комбинированного метода рассчитывает на основе формул относительные показатели. Набор состоит из 76 показателей, 55 из которых пользователь может отключать или подключать в настройках для уменьшения времени, необходимого для расчета коэффициентов. Опишем базовые формулы, используемые для расчета основных показателей.

$D_i = \frac{k_i}{\sum_{j=1}^{n} k_j}$ – доля частей речи (глаголы, существительные, прилагательные, наречия, частицы, союзы и т.д.), где n – количество частей речи; k_J – количество слов в тексте, принадлежащих к j-й ча-

сти речи. Коэффициент, отображающий долю *i*-й части речи от общего числа других частей речи в тексте.

$L_i = \frac{C}{K}$ – средняя длина – усредненный коэффициент, где C – общее количество единиц информации; K – общее количество слов, предложений или абзацев. Отображает *i*-е отношение числа единиц информации (символов, слов, предложений) к количеству синтаксически более сложных образований в тексте (слов, предложений, абзацев). К таким показателям относятся среднее количество символов на слово, слов на предложение, предложений на абзац.

$C_{\text{сим}} = \frac{c}{K_{\text{сим}}}$ – доля символов, где c – количество или гласных, или согласных, или числительных и т.д.; $K_{\text{сим}}$ – общее количество символов в тексте. Коэффициент, отображающий среднее число гласных, согласных, числительных и других знаков, приходящихся на общее количество символов.

$N_{\text{пад}} = \frac{n_{\text{пад}}}{N}$ – количество существительных в падеже, где $n_{\text{пад}}$ – существительных в падеже; N – общее количество существительных. Коэффициент, отображающий число существительных в различных падежах, приходящихся на общее число существительных.

Возможность расчета большого количества коэффициентов увеличивает вероятность того, что один из показателей отображает уникальное свойство как объекта, так и класса. После завершения расчета коэффициентов они записываются в базу Excel, которая экспортируется в интеллектуальный пакет анализа данных.

Полученные результаты могут быть экспортированы в разные интеллектуальные пакеты анализа данных. В данной работе в качестве интеллектуальной программы анализа данных были выбраны три программы, имеющие множество методов классификации данных: IBM SPSS Statistic 23, Rapid miner Studio Free 7.2, SciKit-Learn 0.18rc2.

На вход интеллектуальному пакету анализа данных с алгоритмом классификации подается набор показателей, полученных в результате анализа 100 текстов (на русском языке), принадлежащих четырем функциональным стилям: литературному – 20 текстов, новостному – 20, научному – 40, официальному – 20.

С целью обучения в набор показателей добавлен условный кластер, который принимает четыре значения: литературный, новостной, научный, официальный.

Также стоит отметить общие настройки некоторых методов:

– количество конечных кластеров – 4, контрольная выборка – 40 %;

– эвклидово расстояние, используемое для измерения расстояний в методе ближайшего соседа;

– в методе дерева решений количество итераций равно 100, количество отношений между деревом-отцом и деревом-сыном – min 2.

Результаты классификации текстов при помощи различных алгоритмов в интеллектуальных пакетах данных представлены в таблице 2.

Таблица 2

Тестовая классификация

Table 2

Test classification

Программа	Метод классификации	Точность, %
Rapid miner	Метод ближайшего соседа	68
Rapid miner	Дерево решений	63
Rapid miner	Многоуровневое дерево решений	95
Rapid miner	Случайных деревьев (лучший результат)	95
SciKit-Learn	Дерево решений	88
SciKit-Learn	Метод Байеса	88
SciKit-Learn	Байес с опорными векторами	87
SciKit-Learn	Роше	50
SciKit-Learn	Метод ближайшего соседа	76
IBM SPSS	Метод ближайшего соседа	73
IBM SPSS	Дерево решений (исчерпывающий CHAID)	98
IBM SPSS	Дерево решений (метод CHAID)	94
IBM SPSS	Дерево решений (метод CRT)	83

Основным алгоритмом интеллектуального анализа для дальнейших экспериментов был выбран SPSS (дерево решений и исчерпывающий CHAID). Как видно из таблицы 2, этот алгоритм обеспечил наиболее точную классификацию текстов – 98 %, что на 3 % выше, чем показал метод деревьев решений в программе Rapid miner.

Эксперименты

С целью классификации русскоязычных текстов был проведен ряд экспериментов, результаты которых также были описаны в работах [15, 16]. В таблице 3 представлены итоги основных экспериментов. Следует отметить, что в экспериментах варьируются размеры выборки и свойства классифицируемых объектов.

Первый и второй эксперименты (проза) – классификация литературных произведений русских классиков; эксперименты направлены на определение сложной структуры стиля автора за счет структуры предложения и частей речи в нем.

Третий эксперимент – классификация художественных произведений современников по литературным жанрам.

Четвертый эксперимент – классификация текстов по функциональным стилям. Цель экспери-

Таблица 3

Результаты экспериментов

Table 3

Experimental results

Эксперимент	Автор	Контрольная выборка 20 %	Контрольная выборка 50 %	Перекрестная проверка
№1	**Произведения (6 кластеров)**			
	Д.А. Гранин	75,00 %	66,70 %	**89,40 %**
	Ф.М. Достоевский	100,00 %	94,10 %	
	А.И. Куприн	100,00 %	88,20 %	
	Л.Н. Толстой	100,00 %	78,90 %	
	А.П. Чехов	100,00 %	100,00 %	
	М.А. Шолохов	87,50 %	100,00 %	
	Общая точность	91,40 %	86,70 %	
	Количество текстов	36	90	180
№ 2	**Произведения (5 кластеров)**			
	Ф.М. Достоевский	75,00 %	73,70 %	**94,70 %**
	А.И. Куприн	88,90 %	83,30 %	
	Л.Н. Толстой	100,00 %	82,60 %	
	А.П. Чехов	100,00 %	100,00 %	
	М.А. Шолохов	100,00 %	100,00 %	
	Общая точность	91,20 %	85,90%	
	Количество текстов	30	75	150
№ 3	**Жанры литературы (4 кластера)**			
	Детектив	66,70 %	87,50 %	**69,90 %**
	Люб. роман	50,00 %	66,70 %	
	Фэнтези	80,00 %	75,00 %	
	Хоррор	100,00 %	63,60 %	
	Общая точность	77,80 %	72,20 %	
	Количество текстов	16	40	**80**
№ 4	**Функциональный стиль (4 кластера)**			
	Литературный	100,00 %	100,00 %	**85,00 %**
	Новостной	71,00 %	30,00 %	
	Научный	100,00 %	95,50 %	
	Официальный	100,00 %	92,90 %	
	Общая точность	90,00 %	84,20 %	
	Количество текстов	16	40	**80**

мента – определить возможность такой классификации данных на ЕЯ.

Результаты экспериментов подтвердили, что использование комбинированного метода позволяет с небольшой ресурсоемкостью и высокой достоверностью осуществить классификацию текстов по ряду задач семантической направленности.

Заключение

Благодаря использованию при анализе текста интегральной технологии частотного, морфологического и интеллектуального анализов исследование позволило выявить закономерности классификации слабоформализуемой информации.

Представленная процедура классификации текста ограничена полноценной обработкой текстов только на русском языке, что делает ее более узкоспециализированной. Однако использование частотно-морфологического анализа, гибкость набора показателей, скорость выполнения анализа, вариативность, применение большого числа методов классификации, а также результаты классификации набора текстов из разных функциональных стилей и их высокая достоверность распознавания позволяют говорить о возможности использования предложенной технологии как одного из эффективных инструментов для анализа естественно-языковой информации.

Литература

1. Soergel D. Organizing information: principles of data base a. retrieval systems. Orlando: Acad. press, 1985, vol. 14, 450 p.
2. Weiss S., Indurkhya N. Predictive data mining: a practical guide. SF, Morgan Kaufmann, 1998, 228 p.
3. Bird S., Klein E., Loper E. Natural language processing with Python. Sebastopol: O'Reilly Media, 2015, 479 p.
4. Islam M.Z., Rahm A., Mehler R. Text readability classification of textbooks of a low-resource language. Proc. 26th Pacific Asia Conf. on Language, Information, and Computation, 2012, pp. 545–553.
5. Manning D., Schutze H. Foundations of statistical natural language processing. Cambridge: MIT Press, 2000, 680 p.
6. Dumais S., Chen H. Hierarchical classification of web content. Proc. 23rd Annual Intern/ ACM SIGIR Conf. on Research and Development in Information Retrieval, 2002, pp. 256–263.
7. Aggarwal C. Data mining: the textbook. NY, Springer, 2015, 734 p.
8. Medlock B.W. Investigating classification for natural language processing tasks. Univ. Cambridge Publ., 2008, 138 p.
9. Internet papers base "Base". URL: https://www.base-search.net (дата обращения: 07.05.2017).
10. Weiss M.S. et al. Maximizing Text-mining performance.

Jour. Intelligent Information Retrieval, 1999, vol. 14, pp. 63–69.

11. Hellwig O. Improving the morphological analysis of classical Sanskrit. Düsseldorf Univ. URL: http://aclweb.org/anthology/W/W16/W16-3715.pdf (дата обращения: 07.05.2017).

12. Caropreso M.F., Matwin S., Sebastiani F. Learner-independent evaluation of the usefulness of statistical phrases for automated text categorization: in Publ. Text databases and document management: theory and practice. Virginia Commonwealth Univ. Publ., 2001, pp. 78–102.

13. Official website of the program automatic text processing «AoT». Chart. Russian morphological dictionary. URL: http://aot.ru/index.html (дата обращения: 07.05.2017).

14. Official website of the program Solarix Engine, Chart. Computer Russian grammar. URL: http://www.solarix.ru/index-ru.shtml (дата обращения: 07.05.2017).

15. Фомин В.В., Осочкин А.А. Классификация текстов на основе частотного и морфологического анализов с применением алгоритмов Data-mining // Информатизация образования и науки. 2016. Вып. 3. С. 137–152.

16. Фомин В.В., Фомина И.К., Осочкин А.А. Классификация текстов на основе частотного и морфологического анализов с применением алгоритмов дата-мининг // Актуальне вопросы и перспективы развития математических и естественных наук. 2016. № 3. С. 64–69.

Software & Systems
DOI: 10.15827/0236-235X.030.3.478-486

Received 22.05.17
2017, vol. 30, no. 3, pp. 478–486

METHOD OF FREQUENCY-MORPHOLOGICAL CLASSIFICATION OF TEXTS

A.A. Osochkin [1], *Postgraduate Student, osa585848@bk.ru*
V.V. Fomin [1], *Dr.Sc. (Engineering), Professor, v_v_fomin@mail.ru*
A.V. Flegontov [1], *Dr.Sc. (Engineering), Professor, flegontoff@yandex.ru*
[1] *Herzen State Pedagogical University of Russia, Reki Moyki Quay 48, St. Petersburg, 191186, Russian Federation*

Abstract. Appearing of centralized data storages and information accumulation as structured tables or semistructured texts is a result of growing attention to data analysis techniques. The analysis of similar data in the long term allows obtaining important information, which can become a basis for making right management decisions or predicting further development of events in many fields. One of the important directions of such analysis is automatic classification of collected data in electronic form. Its simplified model is reduced to reading, text processing and assigning a topic to a document from a given list of topics. Foreign papers more and more often are devoted to medical data classification for further disease progression forecast on the basis of statistics or diagnosis based on medical history. The main difficulty in classification are natural language texts. They are difficult to classify due to linguistic features of language and support by a part of classification methods of exclusively numerical data.

The paper studies scientific activity in the field of NLP based on the annual publication of scientific papers in this field. It also offers the method of Russian-language texts classification that integrates the algorithms of frequency, morphological and intellectual analysis. The paper presents the results of some experiments on the identification method of classification with high classification accuracy. The classification was carried out according to functional, literary, and authorial styles.

Keywords: text classification, frequency analysis, morphological analysis, trees of decisions, data mining, text mining.

References

1. Soergel D. *Organizing information: principles of data base and retrieval systems*. Orlando, Acad. press, 1985, vol. 14, 450 p.

2. Weiss S., Indurkhya N. *Predictive data mining: a practical guide*. SF, Morgan Kaufmann Publ., 1998, 228 p.

3. Bird S., Klein E., Loper E. *Natural language processing with Python*. Sebastopol, O'Reilly Media Publ., 2015, 479 p.

4. Islam M.Z., Rahm A., Mehler R. Text readability classification of textbooks of a low-resource language. *Proc. 26th Pacific Asia Conf. Language, Information, and Computation*. 2012, pp. 545–553.

5. Manning D., Schutze H. *Foundations of Statistical Natural Language Processing*. Cambridge, MIT Press, 2000, 680 p.

6. Dumais S., Chen H. Hierarchical classification of web content. *Proc. 23rd Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval*. 2002, pp. 256–263.

7. Aggarwal C. *Data mining: the textbook*. NY, Springer Publ., 2015, 734 p.

8. Medlock B.W. *Investigating classification for natural language processing tasks*. Univ. Cambridge Publ., 2008, 138 p.

9. *Internet papers base "Base"*. Available at: https://www.base-search.net (accessed May 7, 2017).

10. Weiss M.S., Apte C., Damerau F.J., Johnson D.E., Oles F.J., Goetz T., Hampp T. Maximizing Text-mining performance. *Jour. Intelligent Information Retrieval*. 1999, vol. 14, pp. 63–69.

11. Hellwig O. *Improving the morphological analysis of classical Sanskrit*. Düsseldorf Univ. Available at: http://aclweb.org/anthology/W/W16/W16-3715.pdf (accessed May 7, 2017).

12. Caropreso M.F., Matwin S., Sebastiani F. Learner-independent evaluation of the usefulness of statistical phrases for automated text categorization: in Publ. *Text databases and document management: theory and practice*. Virginia Commonwealth Univ. Publ., 2001, pp. 78–102.

13. *Official website of the program automatic text processing "AoT"*. Chart. Russian morphological dictionary. Available at: http://aot.ru/index.html (accessed May 7, 2017).

14. *Solarix Engine*. Chart. Computer Russian grammar. Available at: http://www.solarix.ru/index-ru.shtml (accessed May 7, 2017).

15. Fomin V.V., Osochkin A.A. Text classification based on frequency and morphological analysis using data-mining algorithms. *Informatizatsiya obrazovaniya i nauki* [Informatization of Education and Science]. 2016, iss. 3, pp. 137–152 (in Russ.).

16. Fomin V.V., Fomina I.K., Osochkin A.A. Texts classification based on a frequency and morphological analysis with date-mining algorithms. *Aktualnye voprosy i perspektivy razvitiya matematicheskikh i estestvennykh nauk* [Actual Issues and Prospects for the Development of Mathematical and Natural Sciences]. 2016, no. 3, pp. 64–69 (in Russ.).