П

УДК 004.89

В. Н. Павлыш, С. А. Зори, Е. И. Бурлаева
Государственное образовательное учреждение высшего профессионального образования «Донецкий национальный технический университет», г. Донецк
83001, г. Донецк, ул. Артёма, 58

ЗАДАЧА КЛАССИФИКАЦИИ ИНФОРМАЦИИ ПРИ ФОРМИРОВАНИИ БАЗ ДАННЫХ В КОМПЬЮТЕРНЫХ ОБУЧАЮЩИХ СИСТЕМАХ

V. N. Pavlysh, S. A. Zori,E. I. Burlaeva
State Educational Institution of Higher Education «Donetsk national technical University», Donetsk city
83001, Donetsk, Artema str., 58

THE PROBLEM OF INFORMATION CLASSIFICATIONWHEN DATA BASE FORMING INCOMPUTERTEACHING SYSTEMS

В. М. Павлиш, С. А. Зорі, К. І. Бурлаєва
Державна освітня установа вищої професійної освіти «Донецький національний технічний університет», м. Донецьк
83001, м. Донецьк, вул. Артема, 58

ЗАДАЧА КЛАСИФІКАЦІЇ ІНФОРМАЦІЇ ПРИ ФОРМУВАННІ БАЗ ДАНИХ У КОМП'ЮТЕРНИХ НАВЧАЛЬНИХ СИСТЕМАХ

В статье рассматривается задача анализа методов и выбора принципов классификации текстовой информации для формирования баз данных при конструировании компьютерных обучающих систем. Обоснована эффективность комбинированного подхода к решению задачи с учётом преимуществ и недостатков существующих методов при работе со специфическими группами информации.
Ключевые слова: информация, текст, метод, система, эффективность, принцип.

In the article the task of methods analysis and principles choice for text information classification when data base forming during computer teaching system creation is considered. The effectiveness of combinational principle for task solution noting preferences and defects of existing methods during deal with specific information groups is provided.
Key words: information, text, method, system, effectiveness. principle.

У статті розглядається задача аналізу методів та вибору принципів класифікації текстової інформації для формування баз даних при конструюванні комп'ютерних навчальних систем. Обґрунтовано ефективність комбінованого підходу до розв'язання задачі з урахуванням переваг та недоліків існуючих методів при роботі зі специфічними групами інформації.
Ключові слова: інформація, текст, метод, система, ефективність, принцип.

Введение

Процесс информатизации образования и связанное с этим использование возможностей средств новых информационных технологий в процессе обучения приводит не только к изменению организационных форм, но и к возникновению новых методов обучения.

Математизация и информатизация предметных областей, интеллектуализация учебной деятельности, общие интеграционные тенденции процесса познания окружающей информационной, экологической, социальной среды, поддерживаемые использованием компьютерных технологий, приводят к расширению и углублению изучаемых предметных областей интеграции изучаемых предметов или отдельных тем. Это обусловливает изменение критериев отбора содержания учебного материала. Они основываются на необходимости интенсификации процесса интеллектуального становления и саморазвития личности обучаемого, формирования умений формализовать знания о предметном мире, извлекать знания, пользуясь различными современными методами обработки информации [1].

В настоящее время нет единой общепринятой классификации информационных и программных средств, хотя во многих работах в зависимости от целей, реализация которых оправдывает применение компьютеров, выделяются специфические типы [2].

Актуальность работы

Развитие современного общества характеризуется процессом информатизации всех сфер деятельности. Тенденции постиндустриального развития общества таковы, что основной акцент перемещается с простого восприятия информации на развитие умения использовать информацию для решения практических и профессиональных задач в условиях, которые быстро изменяются. Наиболее перспективным направлением интенсификации усвоения информационных потоков является их классификация, однако при этом возникают проблемы, связанные с различием специфических типов информации. В этой связи задача анализа принципов и алгоритмов классификации текстовой информации, определение их положительных аспектов и отрицательных последствий является актуальной.

Цель работы – анализ методов и обоснование принципов классификации информации, циркулирующей в составе баз данных компьютерных обучающих систем.

Основное содержание работы

Важность для теории и практики рассматриваемой задачи определяет достаточно широкий круг исследователей, разрабатывающих методы решения в различных отраслях.

Рассмотрим несколько примеров фундаментальных задач, отражающих специфические типы информации и требования к её классификации и обработке.

Базовой отраслью Донецкого региона является угольная промышленность, при этом всё добываемое топливо отрабатывается подземным способом.

Информация о состоянии массива горных пород является важнейшей базой для принятия решений о выборе систем разработки и способов поддержания горных выработок [3].

Одним из первых фундаментальных этапов при проектировании угольной шахты является обоснование принятия решения о выборе системы разработки. На этом этапе

проводится анализ больших объемов разнообразной информации о состоянии массива горных пород, при этом в ряде случаев традиционные методы не дают достаточно надежных результатов, что определяет проблему модификации применяемых математических методов.

Современная концепция мониторинга каких-либо процессов или явлений, в том числе и анализ информации о состоянии массивов горных пород, включает в себя следующие обязательные компоненты:

– первоочередную разработку математических или иных моделей контролируемых процессов;
– выбор и расчет приоритетных контролируемых параметров;
– измерение этих параметров в натурных условиях;
– сопоставление расчетных и измеренных величин с целью внесения необходимой коррекции принятых моделей;
– оценку современного состояния контролируемого объекта путём сопоставления измеренных и прогнозно-критических значений наблюдаемых параметров;
– разработку технических мер по обеспечению эффективности и безопасности горных работ;
– контроль реализации разработанных технических мер и их корректировка.

Любой контролируемый процесс сводится к анализу и обработке временного ряда, полученного путем измерения наблюдаемых значений.

Другой важнейшей задачей является информатизация деятельности промышленного предприятия.

Современное состояние рынка характеризуется повышением значимости и ценности информации в процессе формирования управленческих решений. Машиностроительное предприятие является сложной открытой системой, взаимодействующей с субъектами с помощью информационных потоков, которые способствуют образованию конкурентных преимуществ предприятия в занимаемом сегменте рынка. Поэтому качество и скорость передачи информационных потоков является важным элементом в информационной системе, которая составляет контур внутренней и внешней среды машиностроительного предприятия [4].

В управленческой политике машиностроительного предприятия необходимо четко определять цели и задачи, чтобы оценить эффект совершенствования управления с помощью информационных технологий. Допустим, между стадиями принятия решения покупателем и эффективными коммуникативными каналами, доводящими информацию о товаре, существует определенная зависимость (табл. 1).

Таблица 1 – Характер стадий принятия решения покупателем и цели информационной политики

Стадия принятия решения покупателем	Цели (эффекты) информационной политики
Незнание (потребитель не знает о существовании марки)	Осведомленность о марке
Осведомленность, знание (потребитель ознакомлен, но эмоциональная оценка его случайна)	Отношение к марке
Положительное отношение (потребитель ознакомлен с информацией и разделяет данную ей оценку)	Намерение купить
Предпочтение, желание купить (потребитель ознакомлен с информацией, разделяет данную ей оценку, и готов транслировать ее)	Содействие покупке

Для достижения эффективности коммуникационная политика должна строиться не только на поиске уникальных коммерческих аргументов для продвижения товара, но и концентрации на его уникальных покупательских свойствах; причинах, по которым потенциальный потребитель отдаст предпочтение товару, сравнив его с аналогичными продуктами других предприятий. Уникальные покупательские свойства, находящие выражение в ощутимых выгодах для покупателя, главным образом являются субъективными и напрямую не зависят от объективных свойств товара. Потребитель сравнивает аналогичные продукты различных предприятий и отдает предпочтение тому, который для него «более приятен и полезен».

В этой связи понятие образа имеет определяющий смысл. Поэтому создание и поддержание оптимального образа компании и выпускаемых продуктов является основой не только специализированных видов коммуникации, но влияет на весь комплекс продвижения продукции (реклама, персональная продажа, стимулирование сбыта).

Третий пример, который рассмотрим в рамках данной статьи, посвящён проблеме развития дистанционного образования.

В мировой системе образования быстрыми темпами развивается дистанционная форма обучения. Анализируя сайты как российских, так и зарубежных учебных заведений, можно заметить большое количество программ, создаваемых для разработки дистанционных комплексов [3].

В 2012 г. стартовал совместный проект Гарвардского университета и Массачусетского технологического института (MTI), получивший название edXcourse. Как заявили авторы проекта, они поставили задачей разработку концепции и оптимальных стратегий дистанционного обучения. Особенностью проекта является полная открытость выложенных на сайты вузов курсов; лекции и другие учебные материалы можно скачать бесплатно, есть возможность ознакомиться с уже завершенными курсами или же стать участником действующих. Однако, как заявили авторы проекта, у них пока не выработана окончательная стратегия дистанционного образования.

По мере развития технологий образования необходимо постепенно отказываться от репродуктивного метода, от запоминания знаний, от усвоения умений – «ум заключается не только в знании, но и в умении применять знание на деле» (Аристотель). В период модернизации процесса обучения необходимо воспитывать у нового поколения студентов установку личности на самообразование, самовоспитание, саморазвитие, самоусовершенствование, творческое отношение к любому виду деятельности и развитие критического мышления. Логическое мышление, умственная деятельность являются основой, на которой держится весь научно-образовательный процесс – «не мыслям надобно учить, а мыслить» (И. Кант).

Применение информационных технологий в процессе обучения различным учебным дисциплинам в вузах требует от преподавателя знаний как в области подготовки сценария учебного курса с учетом возможностей инструментальных средств разработки программ, так и в области методики преподавания конкретной дисциплины [4]. Этот аспект должен учитываться и при подготовке преподавателя, который призван не только обучать одному предмету, но и быть проводником использования распределенных информационных ресурсов в обучении другим дисциплинам [5].

Обучение с использованием компьютерных технологий постепенно становится новым образовательным стандартом, который внедряется во все структуры, проводящие подготовку и переподготовку специалистов.

Как видно из вышеизложенного, объёмы и разнообразие информационных потоков требуют решения задачи классификации текстовых документов.

Одной из технологий обработки текстовой информации является автоматическая классификация текстовых документов. Использование этой технологии позволяет сократить время на обработку электронных документов.

Для решения этого вопроса часто применяются различные тематические классификаторы, рубрикаторы и т.д., которые облегчают поиск информации автоматически или вручную.

Одной из самых эффективных становится идея алгоритма объединения нескольких классификаторов в так называемую композицию.

В настоящее время достаточно часто используются системы управления знаниями, которые применяются для решения широкого круга задач. В данной работе рассмотрен подход к решению одной из таких задач – классификации документов.

Несмотря на разницу в методах, большая их часть пытается сделать примерно одно и то же: используя некоторую эвристику (например, расстояние между словами, частоту использования слов или заранее определённые отношения между словами), найти группу слов, которая точно определяет темы или описывает информацию, содержащуюся в исходном тексте.

С учётом частоты встречаемости общей лексики в текстах разного программного обеспечения наиболее естественный путь решения вышеуказанных задач состоит в использовании известной статистической меры TF-IDF для выделения среди слов исходной фразы общей лексики и слов терминов (в том числе в составе сочетаний).

В настоящей работе рассматриваются возможности сокращения размерности векторов для поиска в текстовом корпусе описаний близких фрагментов знаний и языковых форм их выражения.

Рассмотрим некоторые из методов автоматической классификации текста, выделим их преимущества и недостатки, что может послужить отправной точкой для разработки более эффективных подходов.

Уменьшение размерности признакового пространства

Одной из проблем анализа текстов является большое количество слов в документе. Если каждое из этих слов подвергать анализу, то время поиска новых знаний резко возрастет. В то же время не все слова в тексте несут полезную информацию. Таким образом, удаление неинформативных слов, а также приведение близких по смыслу слов к единой форме значительно сокращают время анализа текстов.

Уменьшение размерности позволяет снизить вычислительную сложность алгоритма. При этом важно не ухудшить качество классификации. Предполагается, что ряд термов практически неинформативны.

Иногда предполагается, что ряд термов зависит друг от друга. То есть они с большой вероятностью попадут в один класс или в один документ. Очевидно, что за счет таких признаков можно уменьшить размерность пространства без ухудшения точности алгоритма.

Уменьшение размерности может проходить локально или глобально. Локальность означает, что процедура проводится для каждой категории отдельно, затем оставшиеся слова объединяются в единое пространство. Глобальность подразумевает, что участие принимают сразу все документы обучающей выборки, без учета их принадлежности к той или иной категории.

Отбор неинформативных признаков

Существуют различные способы определения признаков, не влияющих на качество классификации.

Разумно предположить, что, если какой-либо терм практически не встречается в документах обучающей выборки, он не несет специфической информации, определяющей класс объекта. Когда терм, наоборот, встречается во всех документах и много раз, то он также будет нести мало полезной информации о принадлежности документа к тому или иному классу. Если все классы статистически независимы от какого-либо терма, то он также неинформативен. Такие простые правила, конечно, нуждаются в формализации. Приведем некоторые возможные правила, которые позволяют определить наиболее ненужные признаки.

Интерес в данной статье представляет сравнение влияния использования стемминга, Стоп-слов, Нижних границ и *TF-IDF* в разных модификациях алгоритмов.

Все системы тестируются на наборе данных русского языка с использованием английских терминов и названий, при этом тестирование происходит с помощью использования алгоритмов.

Алгоритмы машинного обучения требуют специально подготовленной обучающей выборки, то есть специально размеченных документов, в которых указаны слова, которые являются самыми важными в данном тексте.

Мера *TF-IDF*

В задачах анализа текстов и информационного поиска *TF-IDF* есть статистическая мера, используемая для оценки важности слова в контексте документа, входящего в некоторый текстовый корпус. Согласно определению, данная мера есть произведение *TF*-меры (отношения числа вхождений слова к общему числу слов документа) и инверсии частоты встречаемости слова в документах корпуса (*IDF*).

TF-IDF – это статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса.

Вес некоторого слова пропорционален количеству употребления этого слова в документе и обратно пропорционален частоте употребления слова в других документах коллекции. Вычисляется по формуле:

$$TF - IDF(w, d.D) = TF(w, d) \times IDF(w, D),$$

где *TF* (*term frequency* – частота терма) – отношение числа вхождения некоторого терма к общему количеству термов документа. Таким образом, оценивается важность терма w в пределах отдельного документа d, частота слова оценивает важность слова w_i в пределах отдельного документа.

$$TF(w, d) = \frac{n_i}{\sum_k n_k},$$

где n_i – число вхождений слова i в документ.

$\sum_k n_k$ – общее число слов в данном документе.

IDF (inverse document frequency – обратная частота документа) – инверсия частоты, с которой некоторое слово встречается в документах коллекции. Учёт *IDF* уменьшает вес широкоупотребительных слов. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение *IDF*.

$$IDF(w, D) = \log \frac{|D|}{|(d_i \supset w_i)|},$$

где $|D|$ — количество документов в корпусе.

$|(d_i \supset w_i)|$ — количество документов, в которых встречается слово w_i.

Когда *TF-IDF* функция применена ко всем словам во всех документах корпуса, слова можно отсортировать по полученным весам. Более высокий *TF-IDF* вес говорит о том, что слово важно для данного документа, и в то же время достаточно редко употребляется в других документах корпуса. Это зачастую можно интерпретировать как знак того, что слово является важным для данного конкретного документа и может быть использовано, чтобы точно описать документ. *TF-IDF* предоставляет хорошую эвристику для определения кандидатов в ключевые слова, и этот метод (и многие его модификации) за годы исследований показал свою эффективность. В связи с эффективностью и простотой, *TF-IDF* продолжает активно применяться и сегодня.

Метод «Нижняя граница»

В качестве нижней границы будем использовать алгоритм, возвращающий первые 100 слов. Данный алгоритм является весьма эффективным, так как требует наименьшего количества вычислительных ресурсов, не требует моделей. При этом его точность и полнота зачастую превышает сложные алгоритмы.

Метод «Шумовые слова»

Шумовые слова – это слова, являющиеся вспомогательными и несущие малую смысловую нагрузку о содержании документа. Чаще всего заранее составляется список таких слов, и в процессе предварительной обработки они удаляются из текста. К стоп-словам относят предлоги, союзы и местоимения и т.д.

Метод «Стемминг»

Стеммнигом является морфологический поиск. Он заключается в преобразовании каждого слова и приведении его к общей форме, главным образом используется отбрасывание окончаний. Эта технология алгоритма морфологического разбора, учитывающая языковые особенности, вследствие чего является языково-зависимым алгоритмом.

Для апробации предложенных композиций методов был создан тестовый корпус, состоящий из пятидесяти текстовых документов, которые описывали факты предметной области «Компьютерная графика».

Модификации вышеописанных методов можно представить следующими дополняющими друг друга комбинациями обработки текста (табл. 2).

Таблица 2 – Композиция методов векторного представления текста

Сокращение	Композиция методов	Краткое описание
Т-I	TF-IDF	Использование метода TF-IDF для преобразования текста в вектор без какой-либо еще обработки.
Ш+ Т-I	Шумовые слова+ TF-IDF	Удаление слов, не несущих смысловой нагрузки, и использование алгоритма TF-IDF для приведения текста к векторному виду.

Продолж. табл. 2

Сокращение	Композиция методов	Краткое описание
С+ Т-I	Стемминг + TF-IDF	Приведение слов в тексте к единой основе и преобразование текста при помощи алгоритма TF-IDF.
С + Ш + Т-I	Стемминг + Шумовые слова + TF-IDF	Проведение слов в тексте к одинаковой основе удаление шумовых слов, не несущих в себе смысловую нагрузку текста, и использование алгоритма TF-IDF.
НГ + TF-IDF	Нижняя граница + TF-IDF	Используются сразу два алгоритма. Нижняя граница и TF-IDF в виде линейной комбинации.
Ш+НГ+Т-I	Шумовые слова + Нижняя граница + TF-IDF	Удаление Шумовых слов и использование сразу двух алгоритмов. Нижняя граница и TF-IDF.
С+ Ш+ НГ + TF-IDF	Стемминг + Шумовые слова + Нижняя граница + TF-IDF	Приведение слов в тексте к общей основе, затем удаление слов, которые самостоятельно не несут никакой смысловой нагрузки, использование сразу двух алгоритмов. Нижняя граница и TF-IDF.

В качестве исходных данных для приведения текста в векторный вид используются документы на русском языке. Документ, подвергающийся обработке и анализу, в результате которых получаем вектор признаков, подается на вход классификатора.

Выводы

В результате проведенной работы по выявлению наиболее эффективной композиции методов приведения документов к векторному виду для их дальнейшего использования классификатором, установлена необходимость предварительной обработки текста для уменьшения размерности при сохранении смысловой нагрузки документа.

В ходе исследования была выявлена оптимальная композиция, использующая такие вспомогательные процедуры, как удаление стоп-слов из документов, стемминга, определение важности терма в корпусе документов по *TF-IDF* характеристикам, работа которой позволит ускорить получение результата.

Список литературы

1. Павлыш В. Н. Программные средства при использовании новых информационных технологий в процессе обучения [Текст] / В. Н. Павлыш, М. Н.Зайцева, В. В. Хохлаткина // Машиностроение и техносфера XXI века : сборник трудов XIII международной научно-технической конференции. – Т. 3. – Донецк, 2006. – С.136–139.
2. Павлыш В. Н. Постановка задачи разработки компьютерной системы обучения для дисциплин естественнонаучного профиля [Текст] / В. Н. Павлыш, А. А.Каплюхин, Т. А. Ушакова // Машиностроение и техносфера XXI века : сборник трудов XIV международной научно-технической конференции в г. Севастополе 17 – 22 сентября 2007 г. : в 5-ти т. – Донецк : ДонНТУ, 2007. – Т. 3. – С. 128–134.

3. Павлыш В. Н. Совершенствование методов статистического анализа для обработки информации о состоянии массивов горных пород [Текст] / В. Н. Павлыш, С. С. Гребенкин, О. А. Тихонова // Проблемы горного дела и экологии горного производства : матер. IX Междунар. науч.-практ. конф. (24-25 апреля 2014 г., г. Антрацит). – Донецк : Донбасс, 2014. – С. 24–29.
4. Павлыш В. Н. Информационное обеспечение управления экономической деятельностью машиностроительных предприятий в современных условиях [Текст] / В. Н. Павлыш, М. В. Миньковская, С. Б. Лагойко // Известия ТТИ ЮФУ– ДонНТУ : Материалы Одиннадцатого Международного научно-практического семинара «Практика и перспективы развития партнерства в сфере высшей школы» : в 3 кн. – Кн.2. – 2010, № 10. – Таганрог : Изд-во ТТИ ЮФУ. – С. 171–175.
5. Павлыш В.Н., Анохина И.Ю. Информационные технологии в дистанционном образовании // Известия ЮФУ – ДонНТУ : материалы Пятнадцатой Международной научно-практической конференции «Практика и перспективы развития партнерства в сфере высшей школы». – Кн. 2. – 2014, № 14. – Таганрог : Изд-во ЮФУ,2014. – С.120–125.
6. Павлыш В. Н. Математическое моделирование процессов функционирования специализированных аппаратов конвективного типа [Текст] / В. Н. Павлыш, Е. В. Перинская // Проблемы искусственного интеллекта. – Донецк : ГУ ИПИИ. – 2015. – № 0(1). – С. 89–98.
7. Pavlysh V. N. Modification of computer methods of presentation and analysis of geotechnical information [Text] / V. N. Pavlysh, G. I. Turchanin, O. A. Tikhonova // Проблемы искусственного интеллекта. – Донецк : ГУ ИПИИ. – 2016. – № 1 (2). – С. 15–24.
8. Павлыш В. Н. Проект построения алгоритма классификации текстовых документов [Текст] / В. Н. Павлыш, Е. И. Бурлаева // Проблемы искусственного интеллекта. – Донецк : ГУ ИПИИ. – 2017. – № 4 (7). – С. 24–32.
9. Павлыш В. Н. Математическое моделирование нестационарных процессов в среде с нечётко определёнными параметрами [Текст] / В. Н. Павлыш, Г. Б. Перетолчина// Проблемы искусственного интеллекта. – Донецк : ГУ ИПИИ. – 2018. – № 2 (9). – С. 33–45.
10. Павлыш В. Н. Математическое моделирование процесса движения газовоздушной смеси в сплошной среде (на примере угольного пласта) [Текст] / В. Н. Павлыш, И. В. Тарабаева // Проблемы искусственного интеллекта. – Донецк : ГУ ИПИИ. – 2018. – № 3 (10). – С. 104–111.

References

1. Pavlysh V.N., Zaytseva M.N., Hohlatkina V.V. Programmnyye sredstva pri ispol'zovanii novykh informatsionnykh tekhnologiy v protsesse obucheniya [Program means for new information technologies using in teaching process]. *Mashinostroyeniye i tekhnosfera XXI veka : sbornik trudov XIII mezhdunarodnoy nauchno-tekhnicheskoy konferentsii* [Machine building and technosphere of XXI sentury". The book of the works of XIII international scie-techn. Conference], V. 3, Donetsk, 2006, pp. 136-139.
2. Pavlysh V.N., Kaplyukhin A.A., Ushakova T.A. Postanovka zadachi razrabotki komp'yuternoy sistemy obucheniya dlya distsiplin yestestvennonauchnogo profilya [The definition of the task of computer teaching system creation for nature science disciplines]. *Mashinostroyeniye i tekhnosfera XXI veka : sbornik trudov KHIV mezhdunarodnoy nauchno-tekhnicheskoy konferentsii v g. Sevastopole 17 – 22 sentyabrya 2007 g.* : v 5-ti t. [Machine building and technosphere of XXI sentury". The book of the works of XIV international scie-techn. Conference in t. Sevastopol 17-22 sept. 2007], 5 vol., V. 3, Donetsk, 2007, pp. 128-134.
3. PavlyshV.N., GrebyonkinS.S., Tikhonova O.A. Sovershenstvovaniye metodov statisticheskogo analiza dlya obrabotki informatsii o sostoyanii massivov gornykh porod [The perfection of statistics analysis methods for working of information about mounting massive stage]. *Problemy gornogo dela i ekologii gornogo proizvodstva : mater. IX Mezhdunar. nauch.-prakt. konf. (24-25 aprelya 2014 g., g. Antratsit)* [The problems of mining and mine ecology: Mater. of IX intern. Scie-pract. Conf. (24-25 apr. 2014, t. Antracit)], Donetsk-Donbass, 2014, pp. 24-29.
4. PavlyshV.N.,Minkovskaya M.V., Lagoyko S.B. Informatsionnoye obespecheniye upravleniya ekonomicheskoy deyatel'nost'yu mashinostroitel'nykh predpriyatiy v sovremennykh usloviyakh [The informatics providing of machine building enterprise economics activity control in modern conditions]. *Izvestiya TTI YUFU– DonNTU : Materialy Odinnadtsatogo Mezhdunarodnogo nauchno-prakticheskogo seminara «Praktika i perspektivy razvitiya partnerstva v sfere vysshey shkoly» : v 3 kn.* [TTI SFU – DonNTU Sciences News. Mater. ofXI intern. Scie.-pract. Sem. "Practice and prospects of high school sphere partners development" In 3 books], Taganrog, ed. TTI SFU, b. 2, 2010, No. 10, pp. 171-175.

5. PavlyshV.N., Anokhina I.J. Informatsionnyye tekhnologii v distantsionnom obrazovanii [Information technology in distant education]. *Izvestiya YUFU–DonNTU : materialy Pyatnadtsatoy Mezhdunarodnoy nauchno-prakticheskoy konferentsii «Praktika i perspektivy razvitiya partnerstva v sfere vysshey shkoly»*. [TTI SFU – DonNTU Sciences News. Mater. of XV intern. Scie.-pract. conf. "Practice and prospects of high school sphere partners development"], b.2,2014, No. 14, Taganrog, ed. TTI SFU, 2014, pp. 120-125.
6. Pavlysh V. N., Perinskaya E. V. Matematicheskoye modelirovaniye protsessov funktsionirovaniya spetsializirovannykh apparatov konvektivnogo tipa [Mathematical modeling of functioning processes of special convective type apparatus]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence, Donetsk, 2015, no. 0(1), pp. 89-98.
7. Pavlysh V. N., Turchanin G. I, Tikhonova O. A. Modification of computer methods of presentation and analysis of geotechnical information. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2016, no. 1 (2), pp. 15–24.
8. Pavlysh V. N., Burlayeva Ye. I. Proyekt postroyeniya algoritma klassifikatsii tekstovykh dokumentov [Draft of the algorithm for the classification of text documents] *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2017, no. 4 (7), pp. 24–32.
9. Pavlysh V. N., Peretolchina G. B. Matematicheskoye modelirovaniye nestatsionarnykh protsessov v srede s nechotko opredelonnymi parametrami [Mathematical modeling of non-stationary processes in a medium with indistinctly defined parameters] *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2018, no. 2 (9), pp. 33–45.
10. Pavlysh V. N., Tarabayeva I. V. Matematicheskoye modelirovaniye protsessa dvizheniya gazovozdushnoy smesi v sploshnoy srede (na primere ugol'nogo plasta) [The Mathematical Modeling of Gas-Air Mix Moving Process In Continuous Environment (with Coal Stratum As Example)] *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2018, no. 3 (10), pp. 104–111.

RESUME

V.N. Pavlysh, S.A. Zori, E.I. Burlaeva

The Problem of Information Classification when Data Base Forming in Computer Teaching Systems

Background: the most promising direction of intensifying the assimilation of information flows is their classification, butexist problems about specific types; in this connection the task of analysis of principles and algorithms of text information classification, definition of its positive aspects and negative properties is actual;

thepurposeofthework– the analysis of methods and providing of classification principles of information, which exists in data base content of computer teaching systems.

Materials and methods: in the article the methods of comparative and statistical analysis, imitation modeling and laboratory experiment are used.

Results: as a result of the work done to identify the most effective composition of methods for reducing documents to a vector view, for their further use the classifier establishes the need for preprocessing of the text to reduce the dimension while maintaining the semantic load of the document.

Conclusion: at the entrance of the study, an optimal composition was identified, using such auxiliary procedures as removing stop words from documents, stemming, determining the importance of a term in a corpus of documents on TF-IDF characteristics, which will speed up the receipt of the result.

РЕЗЮМЕ

В.Н. Павлыш,С.А. Зори, Е.И. Бурлаева

Задача классификации информации при формировании баз данных в компьютерных обучающих системах

История вопроса, исходные данные: наиболее перспективным направлением интенсификации усвоения информационных потоков является их классификация, однако при этом возникают проблемы, вызванные спецификой типов информационных потоков; в этой связи задача анализа принципов и алгоритмов классификации текстовой информации, определение их положительных аспектов и отрицательных последствий является актуальной;

Цель работы – анализ методов и обоснование принципов классификации информации, циркулирующей в составе баз данных компьютерных обучающих систем.

Материалы и методы: в статье использованы методы сравнительного и статистического анализа, имитационного моделирования, лабораторного эксперимента.

Результаты: в результате проведенной работы по выявлению наиболее эффективной композиции методов приведения документов к векторному виду для их дельнейшего использования классификатором установлена необходимость предварительной обработки текста для уменьшения размерности при сохранения смысловой нагрузки документа.

Заключение: входе исследования была выявлена оптимальная композиция, использующая такие вспомогательные процедуры, как удаление стоп-слов из документов, стемминга, определение важности терма в корпусе документов по *TF-IDF* характеристикам, работа которой позволит ускорить получение результата.

Статья поступила в редакцию 01.07.2018.