ОТЗЫВ

официального оппонента Стояновой Ольги Владимировны
на диссертацию Козлова Павла Юрьевича
«Нейро-нечеткие методы и алгоритмы анализа электронных неструктурированных текстовых документов»,
представленную на соискание учёной степени кандидата технических наук
по специальности 05.13.17 – Теоретические основы информатики

Актуальность темы диссертации

Проблема обработки неструктурированной информации относится к числу приоритетных проблем, решение которых необходимо для достижения цели перехода к цифровой экономике, сформулированной в Стратегии развития информационного общества в Российской Федерации на 2017-2030 годы. Созданная для реализации Стратегии программа «Цифровая экономика», утвержденная правительством в июле 2017 года, рассматривает в качестве одной из важнейших задач «улучшение доступности и качества государственных услуг для граждан». Для решения указанной задачи необходимо совершенствование процедур обмена электронной информацией между гражданами и государственными службами. Эффективность такого взаимодействия оценивается с применением показателей оперативности реакции на обращения граждан и удовлетворенности последних результатами оказанных услуг. Оба показателя напрямую зависят от скорости обработки поступающих информационных запросов и точности распознавания, содержащихся в них сообщений.

То, что информация в запросах представлена в неструктурированном виде не позволяет использовать классические методы обработки реляционных, объектных, иерархических, а также размеченных текстовых данных. Для анализа неструктурированной текстовой информации могут быть рекомендованы методы семантического поиска, но их применение осложняется спецификой информационных сообщений, содержащих обращения граждан, и особенностями функционирования обрабатывающих эти сообщения государственных служб. К специфике сообщений относится: разнообразие форм обращений, изложение проблем от первого лица, возможность наличия нескольких тематических аспектов в одном обращении. Среди особенностей функционирования государственных служб следует отметить: частые изменения структуры ответственности, регламентов работы, расширение спектра оказываемых в электронном виде услуг.

Сказанное обуславливает необходимость разработки методов, моделей и алгоритмов автоматизированного анализа неструктурированной текстовой информации, учитывающих специфику ее содержания и использования в

системе электронных государственных услуг, что подтверждает актуальность темы диссертации.

Научная новизна диссертации

К основным результатам, содержащим элементы научной новизны, относятся следующие.

1. Метод анализа и рубрицирования неструктурированных электронных текстовых документов. Новизна метода заключается в возможности выбора модели, используемой для решения задачи рубрицирования (классификации) документов, с помощью системы нечетких продукционных правил, входными переменными которой являются размер документа, степень пересечения тезаурусов рубрик и объем статистической информации. Учитывая состав моделей-классификаторов (нечетко-логические, нейро-нечеткие, вероятностные), выбор входных переменных обоснован.

2. Каскадная нейро-нечеткая модель рубрицирования неструктурированных электронных текстовых документов небольшого размера, преобразованных к унифицированному представлению. Модель может использоваться при условии наличия необходимого для ее обучения числа статистических данных (ранее поступивших и классифицированных по рубрикам документов) и незначительном пересечении тезауруса рубрик. Новизна модели заключается в использовании предварительного разбиения множества значимых слов на синтаксические группы, расчете показателей близости синтаксических групп к рубрикам, использовании указанных показателей в нейро-нечетких моделях классификации документов.

3. Модель в виде нечеткого дерева решений, предназначенная для рубрицирования неструктурированных электронных текстовых документов в условиях пересечения тезаурусов рубрик и небольшого объема статистической информации. Отличительной особенностью модели является возможность учета синтаксической роли значимых слов при вычислении степени их принадлежности к рубрикам. Научной новизной обладает также предложенная процедура построения дерева решений, основанная на вычислении расстояния между тезаурусами рубрик.

4. Метод формирования рубрикатора на основании нечеткой динамической кластеризации имеющегося массива электронных текстовых документов. Новизна метода заключается в использовании трех предложенных показателей: степень соответствия документа рубрике, степень неопределенности отнесения документа к рубрике, степень несоответствия документа рубрике, анализ которых служит основой принятия решений об изменении структуры рубрикатора.

Весь комплекс разработанных методов и моделей автоматизированной обработки неструктурированной текстовой информации и ее рубрицирования

в системе электронных государственных услуг в условиях нестационарности состава и структуры рубрик составляет научную новизну диссертации. Полученные научные результаты соответствуют п.5 паспорта специальности в части «разработка и исследование методов и алгоритмов анализа текста» и п.6 в части «разработка принципов и методов извлечения данных из текстов на естественном языке».

Обоснованность и достоверность научных результатов диссертации

Обоснованность полученных в диссертации результатов обеспечивается корректным применением методов искусственных нейронных сетей, нечеткой логики, экспертного оценивания, теории вероятностей, синтаксического и морфологического анализа электронных текстовых документов. Полученные результаты не противоречат общепризнанным научным положениям.

Достоверность результатов подтверждается данными вычислительных экспериментов, при проведении которых использовалась информация, полученная из достоверных источников (администрации Смоленской области и международной базы текстовых данных Newsgroup 20).

Значимость результатов диссертации для теории и практики

Предложенные методы и модели анализа и рубрицирования текстовых документов вносят вклад в развитие теории обработки цифровой информации, представленной на естественных языках.

Разработанные алгоритмы и программные средства, реализующие предложенные методы и модели, могут найти практическое применение в системах электронных услуг для повышения эффективности обработки поступающих пользовательских сообщений. Кроме того, данные решения можно рекомендовать при построении поисковых систем для сужения информационного поля путем точного рубрицирования запросов.

Замечания по диссертации

1. Важной особенностью предложенного метода обработки и рубрицирования текстовых документов является возможность выбора модели-классификатора на основании анализа характеристик документов и рубрик. Одна из характеристик - объем (размер) документа. По данной характеристике на рис.1.6 (стр.28) и в тексте раздела 2.1 (стр. 30) представлены различные наборы категорий без описания связей между ними. Кроме того, выбор числовых значений для границ категорий, представленных на стр.28, ничем не обоснован, а для категорий на стр. 30 критерии вообще отсутствуют. Учитывая, что указанная характеристика влияет на выбор модели и,

следовательно, на результат рубрицирования, следовало строже подойти к формированию критериев ее оценивания.

2. В предложенной каскадной нейро-нечеткой модели, представленной на рис. 2.2 (стр.51), для повышения точности рубрицирования предложено использовать анализ синтаксических характеристик значимых слов. Это существенно усложняет работу классификатора, в то время как эффект от подобного усложнения в условиях слабого пересечения тезаурусов рубрик не очевиден и требует теоретического обоснования, особенно с учетом отсутствия в тексте диссертации примера практического использования данной модели.

3. Из описания шага 2 (стр. 64) процедуры рубрицирования документов с помощью нечеткого дерева решений (в диссертации номер формулы отсутствует, в автореферате - выражение (7)) следует, что реализуется процедура полного перебора на множестве узлов нечеткого дерева решений. В то же время среди достоинств модели (стр. 66) отмечается меньшая трудоемкость «вследствие направленного анализа по отдельной ветви НДР».

4. Из рис.3.1 (стр. 79) следует, что при реализации предложенного мультимодельного метода рубрицирования результаты морфологического анализа используются нечеткой продукционной системой выбора способа рубрицирования для проверки выполнения условий П1-П6, определяющих выбор модели. Из дальнейшего описания алгоритмов и результатов морфологического анализа (стр.85-90) не ясно, какие именно результаты данного анализа и каким образом влияют на указанный выбор.

Перечисленные замечания не снижают степень научной новизны и практической значимости полученных результатов.

Заключение о соответствии диссертации критериям, установленным Положением о порядке присуждения ученых степеней

Содержание диссертации логически последовательно изложено в четырех главах, введении и заключении. Автореферат хорошо отражает содержание диссертации.

Основные положения диссертации своевременно опубликованы в рецензируемых научных журналах и прошли необходимую апробацию на международных научных конференциях.

По актуальности, научной новизне и практической значимости диссертация П. Ю. Козлова представляет собой законченную научно-квалификационную работу. В ней решена научная задача разработки методов, моделей и алгоритмов анализа неструктурированных текстовых документов в

системе электронных государственных услуг, имеющая существенное значение для развития отрасли знаний, связанной с анализом текстов на естественном языке, что соответствует требованиям Положения о присуждении ученых степеней, утвержденного Постановлением Правительства Российской Федерации от 24.09. 2013 г., №842 в части раздела II, п.9.

На основании изложенного считаю, что Козлов Павел Юрьевич заслуживает присуждения учёной степени кандидата технических наук по специальности 05.13.17 – Теоретические основы информатики.

Официальный оппонент
доктор технических наук по специальности 05.13.06, доцент, старший преподаватель кафедры информационных систем в экономике Санкт-Петербургского государственного университета — Ольга Владимировна Стоянова

почтовый адрес: 191123, Санкт-Петербург, ул.Чайковского, д.62, кафедра ИСЭ
e-mail: o.stoyanova@spbu.ru
телефон: +7(921)4495090

ЛИЧНУЮ ПОДПИСЬ Стояновой О.В. ЗАВЕРЯЮ
ВЕДУЩИЙ СПЕЦИАЛИСТ Е. ПРОТАСОВА
20.12.2017