МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное учреждение науки
Институт автоматики и процессов управления
Дальневосточного отделения Российской академии наук
(ИАПУ ДВО РАН)
Радио ул., д. 5, Владивосток, 690041
Телефон (423) 2310439, факс (423) 2310452
E-mail: director@iacp.dvo.ru, http: www.iacp.dvo.ru
ОКПО 02698217, ОГРН 1022502127878
ИНН/КПП 2539007627/253901001
07.09.2021 № 16141/338
На _____ от _____

УТВЕРЖДАЮ
Директор
Института автоматики и процессов управления Дальневосточного отделения Российской академии наук,
член-корреспондент РАН
_____ Р.В. Ромашко
"01" 09 2021 г.

Отзыв ведущей организации

Федеральное государственное бюджетное учреждение науки «Институт автоматики и процессов управления» Дальневосточного отделения Российской академии наук о диссертационной роботе Дударина Павла Владимировича на тему «Исследование и разработка моделей и методов нечеткой кластеризации коротких текстов», представленной на соискание ученой степени кандидата технических наук по специальности 05.13.01 – «Системный анализ, управление и обработка информации (информационные технологии и промышленность)»

1. *Актуальность темы диссертации*

Стремительный рост объемов генерируемой человечеством текстовой информации приводит к необходимости поиска новых эффективных методов машинной обработки. Большие объемы уже накопленных и пополняемых наборов коротких текстов могут быть изучены только с использованием современных инструментов автоматизированной обработки текстов. Примерами таких наборов могут служить посты в социальных сетях, наборы заголовков новостей и научных статей и наборы коротких заметок или объявлений. Работа с такими текстами осложнена частичным или полным отсутствием контекста и большей чувствительностью алгоритмов машинной обработки коротких текстов к опечаткам, аббревиатурам и неологизмам.

Многие из этих примеров с точки зрения машинной обработки и системного анализа начали вызывать интерес относительно недавно, чем

отчасти обусловлено отсутствие уже готовых и законченных стандартных методов анализа. Обзоры работ для русского языка и существующих пакетов обработки текстов показывают, что эффективность существующих методов кластеризации коротких текстов не достаточна для их широкого использования в системах поддержки принятия решения и управления.

Стремительный рост объемов создаваемой в рамках разных видов деятельности текстовой информации приводит к необходимости поиска новых эффективных методов ее обработки. Большие объемы накапливаемых массивов текстов могут быть изучены только с использованием современных инструментов автоматизированной обработки текстов. Примерами таких наборов могут служить наборы заголовков новостей и научных статей, наборы коротких заметок, объявлений, высказываний, в том числе относящихся к некоторой профессиональной деятельности или отрасли, например, сфере деятельности образовательных учреждений или планирования экономического развития. Работа с такими текстами осложнена частичным или полным отсутствием контекста и большей чувствительностью алгоритмов машинной обработки коротких текстов к опечаткам, аббревиатурам и неологизмам.

Обзоры работ для русского языка показывают, что пока для использования в системах поддержки принятия решения недостаточно законченных стандартных методов анализа коротких текстов и готовых пакетов обработки текстов, а эффективность существующих методов кластеризации текстов не достаточна для их широкого использования.

Поэтому тематика данной диссертации, посвященная исследованию и разработке моделей и методов, нечеткой кластеризации наборов коротких текстов с возможностью учета экспертной информации является актуальной задачей для исследования.

2. *Научная новизна результатов диссертационной работы*

Научная новизна исследований и полученных результатов диссертационной работы заключается в том, что ее автором получены новые научные результаты.

1. Предложенная архитектура искусственной нейронной сети, отличающаяся от известных тем, что позволяет решать задачу кластеризации на базе скрытого пространства признаков языковой модели;
2. Предложенный метод обработки текстов для расширения словаря языковой модели на базе нейронной сети с использованием нечеткого иерархического классификатора, отличающийся от известных тем, что позволяет учитывать семантическую близость слов;
3. Предложенный метод обработки обратной связи от эксперта, отличающийся от известных тем, что позволяет корректировать весовые коэффициенты нейронной сети и проводить интерактивную кластеризацию наборов коротких текстов;
4. Разработанный алгоритм, автоматизирующий применение предложенных модели и методов для выполнения нечеткой интерактивной кластеризации наборов коротких текстов, интегрированный в систему поддержки принятия решений (СППР).

3. *Значимость полученных результатов для науки и практики*

Теоретическая значимость полученных в диссертационной работе результатов заключается в разработке метода интерактивной кластеризации позволяющего учитывать обратную связь от эксперта в процессе обучения нейронной сети методом обратного распространения ошибки, а также в разработке метода расширения словаря языковой модели позволяющего конструировать векторы признаков слов с учетом различных смысловых оттенков.

Практическая ценность результатов диссертационной работы состоит в разработанных программных средствах для интерактивной нечеткой

кластеризации текстов на языке Python, позволяющих осуществлять интерактивную нечеткую кластеризацию коротких текстов, и применение разработанного программного обеспечения в задаче анализа набора коротких текстов в рамках НИР в интересах Министерства экономического развития РФ для Системы стратегического планирования РФ.

4. Рекомендации по использованию результатов и выводов диссертации

Полученные в ходе диссертационного исследования результаты можно рекомендовать к использованию на практике для решения задачи построения автоматических классификаторов как результата применения метода кластеризации для больших объемов коротких текстов, а также для проведения верификации классификатора и его уточнения по результатам проведенной кластеризации. В частности могут выявляться новые кластеры, старые кластеры могут исчезать или объединяться, также может меняться иерархическая структура дерева кластеров.

Разработанный метод интерактивной кластеризации является универсальным и может быть применен для различных наборов коротких текстов, желательно его включение в какой-либо открытый пакет обработки текстов.

5. Достоверность результатов и обоснованность выводов

Достоверность результатов, полученных в ходе выполнения диссертационных исследований, обеспечивается корректными постановками задач, результатами проведенных вычислительных экспериментов и их анализом, а также подтверждается результатами проверки работоспособности разработанных моделей и методов при апробации в качестве элементов системы поддержки принятия решений.

6. *Структура и основное содержание диссертационной работы*

Диссертационная работа включает в себя введение, 4 главы, заключение, список литературы и приложения. Общий объем диссертации – 136 страниц, включая 46 рисунков и 9 таблиц. Библиография включает в себя 128 наименований.

Во введении охарактеризована актуальность темы диссертационной работы, определены объект, предмет и цель работы, поставлены исследовательские задачи, сформулированы теоретическая значимость, научная новизна и практическая ценность полученных результатов, перечислены выносимые на защиту положения.

В первой главе проводится сравнительный анализ моделей и методов нечеткой кластеризации коротких текстов. Отмечается рост количества и объемов наборов коротких текстов, а также увеличение числа работ посвященных методам интерактивной кластеризации. Проводится анализ современных языковых моделей основанных на искусственных нейронных сетях.

Во второй главе представлена разработанная архитектура искусственной нейронной сети, позволяющая решить задачу кластеризации коротких текстов на базе языковой модели. Далее в разработанную архитектуру добавляется слой позволяющий расширить словарь языковой модели, для чего представлен метод на базе кластеризации нечеткого графа. Для решения задачи учета экспертной информации была проведена модификация целевой функции предложенной нейронной сети, что позволило включить обратную связь, получаемую от эксперта, в процедуру обучения сети методом обратного распространения ошибки.

В третьей главе дается описание Федеральной Информационной Системы “Стратегическое Планирование” (ФИС СП), в рамках которой проводилась апробация результатов представленной работы. Разработанные в рамках второй главы методы сведены в единый алгоритм нечеткой интерактивной кластеризации коротких текстов для возможности

автоматизации этой процедуры в рамках ФИС СП. Показано качественное изменение эффективности функционирования ФИС СП.

В четвертой главе представлены результаты ряда экспериментов, подтверждающие работоспособность и эффективность предлагаемого метода нечеткой интерактивной кластеризации коротких текстов. Определены границы применимости метода, проведено сравнение с аналогами и обозначены возможные дальнейшие шаги по развитию предложенного метода.

В заключении отражены основные результаты диссертации.

В приложениях приведены документы подтверждающие внедрение результатов диссертационной работы, свидетельство о государственной регистрации программы для ЭВМ и подробные результаты основного эксперимента четвертой главы.

7. *Соответствие требованиям по выполнению, оформлению и апробации диссертационной работы*

Основные результаты работы докладывались и получили одобрение на 12 международных и национальных конференциях. Содержание диссертации полно отражено в 19 научных работ, в том числе в 6 статьях в изданиях из перечня ВАК и в 7 статьях в изданиях, индексируемых в базе данных Web of Science и Scopus. Получено 1 свидетельство о государственной регистрации программ для ЭВМ,

Основные положения и результаты диссертационной работы доложены и обсуждены на конференциях и конгрессах: Всероссийская научно-практическая конференция “Нечеткие системы и мягкие вычисления” (Санкт-Петербург, 2017); Международная конференция “Интеллектуальные информационные технологии в технике и на производстве” IITI (Варна, 2017; Сочи, 2018; Острава, 2019); Всероссийская научная конференции «Нечеткая логика и мягкие вычисления в промышленности» (Ульяновск, 2017, 2018, 2019); Национальная Конференция по Искусственному Интеллекту (Москва,

2018; Ульяновск, 2019); Международная конференция “World Conference on Soft Computing” (Баку, 2018); Международная конференция “Mexican International Conference on Artificial Intelligence” (Гвадалахара, 2018); Международная конференция “European Society for Fuzzy Logic and Technology” (Прага, 2019); Международная конференция по компьютерной лингвистике и интеллектуальным технологиям “Диалог” (Москва, 2019); I Национальный конгресс по когнитивным исследованиям, искусственному интеллекту и нейроинформатике (Москва, 2020).

Основные теоретические и практические результаты диссертационной работы использованы в рамках фундаментальных и прикладных научных исследований Министерства экономического развития РФ по темам: “Разработка рекомендаций по совершенствованию информационного обеспечения участников стратегического планирования в части осуществления мониторинга и контроля реализации документов стратегического планирования с использованием Федеральной информационной системы стратегического планирования (ФИС СП)” и “Разработка методического обеспечения интеллектуальной системы проверки уведомления об утверждении (одобрении) документа стратегического планирования или внесении в него изменений при ведении федерального государственного реестра документов стратегического планирования Федеральной информационной системы стратегического планирования (ФИС СП)”. Результаты НИР внедрены в системе ГАС “Управление”.

Архитектура искусственной нейронной сети и алгоритм нечеткой кластеризации коротких текстов, методы расширения словаря языковой модели и корректировки весов нейронной сети для учета обратной связи эксперта в интерактивной кластеризации, а также программная реализация метода нечеткой интерактивной кластеризации на языке Python внедрены в системе Планета.Аналитика 4.0 (включена в реестр отечественного ПО) компании ООО “ИБС “Экспертиза”.

Автореферат соответствует установленным стандартам, достаточно полно отражает содержание диссертации и содержит 24 страницы.

8. Замечания по диссертационной работе

По работе следует отметить следующие замечания и рекомендации.

1. Указанные значения метрики перплексии для разных языковых моделей не рассчитывались автором самостоятельно, а приводились из сторонних исследований, в результате перплексия считалась на различных наборах текстов, что не позволяет достоверно провести сравнение качества обучения языковых моделей между собой.

2. В работе не приведено убедительное обоснование выбора корпуса Тайга и обученной на нем модели word2vec для построения нечеткого графа.

3. При изложении метода расширения языкового словаря использован двусмысленный термин “незнакомое слово”, который в контексте изложения может быть воспринят как незнакомый для исследователя, так и как незнакомый только для предварительно обученной языковой модели (т.е. не содержащийся в первоначальном наборе данных на котором проводилось обучение).

4. В своей работе автор не приводит аргументацию, почему достигнутый в одном из экспериментов показатель качества в 80,1% является хорошим. Остается неясным возможно ли получить большую точность.

5. При расчете метрики точности использовались формулы расчета для точной кластеризации, при этом предложенный метод является нечетким, и было бы интересно рассчитать точность с учетом нечеткого распределения по кластерам.

6. Для экспериментов по определению точности и границ применимости метода использовались наборы по 2000 текстов, при этом в корпусе значительно больше текстов. Возможно, результаты отличались бы другими для полного набора текстов.

7. Для серии экспериментов с набором текстов ключевых показателей системы Стратегического планирования приводятся только оценки эффективности работы алгоритма и не приводятся оценки точности.

Отмеченные недостатки носят частный характер и, по нашему мнению, не влияют на общую положительную оценку хорошего уровня диссертационный работы Дударина П.В.

9. Заключение

Диссертация является законченной научно-квалификационной работой, выполненной на актуальную тему. Новые научные результаты, полученные в диссертации, направлены на решение научной задачи, имеющей важное значение для развития методов решения задачи системного анализа и обработки экспертной информации.

Работа Дударина Павла Владимировича является самостоятельным научно-исследовательским трудом для областей исследований, перечисленных в паспорте специальности 05.13.01 – «Системный анализ, управление и обработка информации (информационные технологии и промышленность)», прежде всего, пункту 4 «Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации» и пункту 13 «Методы получения, анализа и обработки экспертной информации».

Представленная диссертация на тему «Исследование и разработка моделей и методов нечеткой кластеризации коротких текстов» удовлетворяет требованиям предъявляемым ВАК РФ к диссертациям на соискание ученой степени кандидата технических наук, а ее автор, Дударин Павел Владимирович, достоин присуждения ученой степени кандидата технических наук по специальности 05.13.01 - «Системный анализ, управление и обработка информации (информационные технологии и промышленность)».

Отзыв подготовлен Грибовой Валерией Викторовной, доктором технических наук, заместителем директора по научной работе Федерального

государственного бюджетного учреждения науки «Институт автоматики и процессов управления» Дальневосточного отделения Российской академии наук, 690041, г. Владивосток, ул. Радио, 5; тел: (423) 2310439, факс: (423) 2310452, web-сайт: http://www.iacp.dvo.ru, e-mail: gribova@iacp.dvo.ru

Диссертационная работа Дударина Павла Владимировича на тему «Исследование и разработка моделей и методов нечеткой кластеризации коротких текстов», представленная на соискание ученой степени кандидата технических наук по специальности 05.13.01 - «Системный анализ, управление и обработка информации (информационные технологии и промышленность)», рассмотрена и обсуждена на расширенном заседании семинара лаборатории интеллектуальных систем «Института автоматики и процессов управления» Дальневосточного отделения Российской академии наук, протокол № 2/21 от 31.08.2021 г.

Заместитель директора по научной работе,
научный руководитель лаборатории
интеллектуальных систем, д.т.н. В.В. Грибова

Ученый секретарь семинара
к.т.н. Е.А. Шалфеева