ОТЗЫВ

официального оппонента кандидата физико-математических наук Сухова Сергея Владимировича на диссертационную работу Дударина Павла Владимировича «Исследование и разработка моделей и методов нечеткой кластеризации коротких текстов», представленную на соискание ученой степени кандидата технических наук по специальности 05.13.01 «Системный анализ, управление и обработка информации (информационные технологии и промышленность)»

Актуальность работы

Автоматическая обработка текстов является одним из приоритетных направлений автоматизации процессов обработки информации для решения широкого круга задач. Важным этапом интеллектуального анализа текстовых данных является кластеризация текста. Стремительный рост массивов информации, состоящих из наборов коротких текстовых фрагментов, вызывает необходимость развития методов обработки коротких текстов с применением машинного обучения. В отличие от текстов с большим количеством слов короткие тексты, состоящие из нескольких фраз, представляют большие трудности для кластеризации. Методы кластеризации больших текстов применительно к коротким текстам не дают удовлетворительных результатов. Проблемой кластеризации коротких текстов занимается большое число исследователей, однако большая часть проводимых исследований относится к текстам на английском языке. Для русского языка методы кластеризации разработаны на недостаточном уровне. Недостаточная эффективность имеющихся методов кластеризации коротких русскоязычных текстов затрудняет их использование в российских автоматизированных системах поддержки принятия решений и управления.

Обычно кластеризация текстов допускает несколько возможных вариантов для разбиения на группы. Методы, не позволяющие учесть намерения эксперта при кластеризации, могут оказаться неэффективными для конкретной задачи. Интерактивное участие эксперта позволяет корректировать кластеризацию, произведенную автоматической системой. Разрабатываемые в диссертационной работе П.В. Дударина методы машинного обучения должны заменить (или значительно облегчить) кропотливый ручной труд экспертов по кластеризации.

На основе вышеизложенного можно сказать, что исследования в области интерактивной кластеризации коротких русскоязычных текстов, выполненные в диссертационной работе П.В. Дударина, направлены на решение важной и актуальной задачи.

Научная новизна диссертации

В качестве новых результатов диссертационной работы следует отметить следующие:

1. Диссертантом предложена оригинальная архитектура искусственной нейронной сети, позволяющая решать задачу кластеризации на базе скрытого пространства признаков языковой модели русского языка.
2. С целью расширения словаря языковой модели разработан оригинальный метод обработки текстов, учитывающий семантическую близость слов и основанный на совместной работе нейронных сетей и нечеткого иерархического классификатора.
3. Предложен метод обработки обратной связи от эксперта, позволяющий интерактивно включать и исключать объекты из кластера, что позволило автоматически корректировать весовые коэффициенты нейронной сети.

Научная и практическая значимость работы

Научная значимость полученных в работе результатов заключается в развитии методов интеллектуальной обработки текстовой информации, в частности, в разработке новых моделей и методов с использованием нейронных сетей и методов нечеткой логики для решения задачи кластеризации коротких русскоязычных текстов. Предложенные методы являются универсальными и позволяют снизить нагрузку на экспертов при кластеризации больших объемов коротких текстов, что делает целесообразным их дальнейшее применение в прикладных исследованиях для проектирования информационно-аналитических систем.

Практическая значимость диссертационной работы заключается в разработке программного модуля системы поддержки принятия решений, позволяющего осуществлять интерактивную нечеткую кластеризацию наборов коротких текстов, и применение этого модуля для Системы стратегического планирования Министерства экономического развития Российской Федерации. Как показано в диссертации, интерактивные методы кластеризации обеспечили сокращение суммарных затрат времени эксперта на обработку результатов и позволили повысить точность кластеризации. Для разработанного автором программного модуля получено свидетельство о государственной регистрации программы для ЭВМ.

Достоверность результатов проведенных исследований

Основные положения, выводы и рекомендации, полученные в диссертационной работе, являются обоснованными и аргументированными. Достоверность полученных результатов подтверждается корректным использованием известных методов и подходов по обработке текста.

Полнота изложения материалов диссертации в печатных работах, опубликованных автором

Основные результаты исследований опубликованы в 19 работах, из которых 6 статей – в рецензируемых журналах, входящих в перечень ВАК при Минобрнауки РФ, 7 работ – в журналах, индексируемых в Scopus и Web of Science, 6 – в материалах научных конференций.

Структура и содержание диссертации

Диссертация состоит из введения, четырех глав, заключения, списка литературы и трех приложений. Объем работы составляет 136 страниц машинописного текста, включая 46 рисунков, 9 таблиц, и списка использованной литературы из 128 наименований.

Во введении обоснована актуальность темы диссертационного исследования, сформулированы цель и задачи работы, отражена научная новизна, практическая значимость, достоверность и обоснованность результатов исследований диссертации, приведены основные положения, выносимые на защиту, указана степень апробации и реализации результатов исследования.

Первая глава посвящена сравнительному анализу моделей и методов кластеризации коротких текстов и формулировке целей и задач исследования. В этой главе проведена систематизация интерактивных методов кластеризации. Рассмотрена задача кластеризации коротких текстов. Показана возможность повышения качества работы алгоритма кластеризации коротких текстов путем интерактивного взаимодействия с экспертом. Обоснована необходимость использования концепции переноса знаний (“transfer learning”) на основе предобученной нейронной сети. Проведено сравнение наиболее распространенных предварительно обученных моделей.

Во **второй главе** разработана архитектура искусственной нейронной сети, позволяющая решать задачу кластеризации коротких текстов, при использовании сжатых векторных представлений, получаемых с помощью кодировщика языковой модели. Предложены метод расширения словаря языковой модели, метод корректировки весов нейронной сети для “неизвестных” слов, а также метод корректировки весов нейронной сети, позволяющий учитывать ограничения, задаваемые экспертом при решении задачи кластеризации. Предложен оригинальный тип ограничений, заключающийся в запрете элементу находиться в определенном кластере.

В **третьей главе** рассмотрена федеральная информационная система “Стратегическое Планирование”. Приводится описание спроектированных диссертантом блоков программного модуля, позволяющего автоматизировать работу эксперта для решения задачи нечеткой интерактивной кластеризации

коротких текстов. Составлен алгоритм нечеткой интерактивной кластеризации коротких текстов.

В **четвертой главе** разработана программа для проведения испытаний по нечеткой интерактивной кластеризации коротких текстов. На модельных наборах коротких текстов (массив объявлений Авито) проведены эксперименты, подтверждающие работоспособность и эффективность разработанных методов кластеризации. Приводится описание результатов решения задачи кластеризации коротких текстов, содержащих показатели эффективности системы стратегического планирования Российской Федерации. Проведенные эксперименты продемонстрировали, что добавление экспертом всего нескольких ограничений значительно повышает точность кластеризации. Экспериментально подтверждено, что предлагаемый метод позволяет сократить временные затраты экспертов на этапе кластеризации и построения классификатора в 3 раза по сравнению с полностью ручной кластеризацией и в 2 раза – в случае использования методов тематической кластеризации. Показано, что время, затрачиваемое экспертами на проверку корректности классификации ключевых показателей эффективности во входящих документах, сокращается в 2-3 раза. В то же время показано, что применительно к длинным текстам интерактивная кластеризация оказывается малоэффективной.

В **заключении** приведены основные результаты работы, обсуждаются перспективы дальнейшего практического применения результатов. Список цитируемой литературы содержит обширную и достаточную библиографию по тематике диссертации. В **приложении** вынесены акт о внедрении результатов диссертационной работы, свидетельство о государственной регистрации программы для ЭВМ, результаты по кластеризации ключевых показателей эффективности системы “Стратегическое Планирование”.

К несомненным **достоинствам** работы следует отнести:

1. Гибкое использование передовых методов машинного обучения (таких как перенос знаний, тонкая настройка). Использование методов из различных областей машинного обучения, позволившее получить эффективные гибридные системы на основе интеграции нечеткой логики и нейронных сетей.
2. Большой объем работ по созданию многомодульной системы, интегрированной в систему поддержки принятия решений, позволяющей автоматизировать работу экспертов для решения задачи нечеткой интерактивной кластеризации коротких текстов.

Замечания по диссертационной работе:

1. Выводы главы 3 о том, что разработанный программный модуль позволил эффективно организовать совместную работу экспертов по кластеризации коротких текстов является преждевременным, так как эксперименты по эффективности модели проводятся только в следующей главе.
2. В связи с отсутствием объективных критериев правильности кластеризации в главе 4, было бы целесообразно получить оценку эффективности алгоритма от экспертов. Этого в диссертации сделано не было.
3. В формулах основного текста диссертации и в автореферате не определены некоторые переменные. Например, не определена переменная z_i в формуле (2.1) (стр.46 диссертации), не определена переменная u в уравнении (2.6) (стр.47 диссертации), не определена переменная μ_j (уравнения 5, 6) в тексте автореферата.
4. Рисунок 2.17 диссертации неинформативен. Подрисуночная подпись не объясняет рисунок. Рисунок 4 автореферата непонятен без пояснений в тексте.
5. В тесте диссертации отсутствуют ссылки на окончательные результаты кластеризации ключевых показателей эффективности системы стратегического планирования (Приложение 3).

Указанные замечания не снижают научной значимости результатов исследования.

Соответствие паспорту научной специальности

Тема и содержание диссертационной работы соответствует паспорту специальности 05.13.01. – «Системный анализ, управление и обработка информации (технические науки)», а именно:

п. 4 – разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации;

п. 13 – методы получения, анализа и обработки экспертной информации.

Заключение о соответствии диссертационной работы критериям, установленным Положением о порядке присуждения ученых степеней

Диссертационная работа Дударина П.В. содержит значимые научные результаты по специальности 05.13.01 «Системный анализ, управление и обработка информации (информационные технологии и промышленность)». Диссертационная работа Дударина П.В. является законченным исследованием, в котором содержится решение важной задачи по разработке математических методов и алгоритмов кластеризации коротких текстов с интерактивным участием эксперта.

По объему и научному уровню полученных результатов диссертационная работа удовлетворяет требованиям ВАК при Минобрнауки РФ, предъявляемым

к кандидатским диссертациям. Содержание автореферата соответствует основному содержанию диссертационной работы.

Считаю, что диссертационная работа удовлетворяет требованиям пп. 9-11, 13, 14 Положения о присуждении учёных степеней в редакции 28.08.2017 г., предъявляемым к кандидатским диссертациям, а ее автор, Дударин Павел Владимирович, заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.01 «Системный анализ, управление и обработка информации (информационные технологии и промышленность)».

Официальный оппонент,
Старший научный сотрудник, к.ф.-м.н. Сухов Сергей Владимирович
УФИРЭ им. В.А.Котельникова РАН

13.09.2021

Подпись к.ф.-м.н. Сухова С.В. заверяю:

Подпись Сухова С.В. заверяю.
Старший инспектор по кадрам
УФИРЭ им. В.А.Котельникова РАН
Солнцева
«13» сентября 2021 г.

Сухов Сергей Владимирович, к.ф.-м.н. (01.04.05), старший научный сотрудник
432071, г. Ульяновск, ул.Гончарова, 48/2
Ульяновский филиал Федерального государственного
бюджетного учреждения науки «Институт радиотехники
и электроники им. В.А.Котельникова Российской академии наук»
e-mail: ssukhov@ulireran.ru
Тел.: (8422) 44-29-96