Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК_____

Ковалевский
Александр Михайлович

Алгоритмы профилирования пользователей посредством нейронных сетей

АВТОРЕФЕРАТ

на соискание степени магистра технических наук
по специальности: 1-40 80 02 – Системный анализ, управление и обработка информации

_____

Научный руководитель
Гуринович Алевтина Борисовна
канд. физ.-мат. н., доцент

_____

Минск 2018

ВВЕДЕНИЕ

В настоящее время в обществе, компьютеры и интернет широко используется для доступа к различным видам информации. Сейчас недостаточно того, чтобы компьютеры выполняли сложные задачи в нужные сроки и хранили большие объемы информации. С появлением интернета, объем информации увеличился во много раз и ежедневно этот объем информации увеличивается. Сейчас в интернете множество разнообразных ресурсов, начиная от социальных сетей, видео сервисов до новостных лент различной тематики. И современный человек обязан быть в курсе событий, а для этого необходимо ежедневно просматривать большой объем информации. Поэтому можно говорить о такой важной задаче как профилирование пользователей, так как большая часть информации в интернете не представляет интереса для конкретного пользователя, что её просматривает. Это связано с тем, что каждый человек уникален.

Профилирование – разумное ограничение предъявляемой посетителю информации с целью выделения более важного для него содержания.

К примеру профилирование необходимо для решения следующих задач:

1. Поиск информации.
2. Распознавание эмоциональной окраски текстов.
3. Разделение сайтов по тематическим каталогам.
4. Борьба со спамом.
5. Персонификация рекламы.

Решение данной задачи позволит потребителям услуг тратить меньше времени на поиск, просмотр и усвоение контента и больше на практическое применение. Тем самым увеличивается эффективность работы во многих сферах связанных с масс медиа: информирование о свежих новостях, подбор информации по заданной тематике при поиске, предложение конкретных товаров и услуг, необходимых потребителю. А также при сокращении времени на данные виды деятельности, у пользователей будет больше времени на другие, что эффективно скажется на их работе, и экономике государств в целом.

В работе показано, на чем основано профилирование пользователя, какую роль играют нейронные сети в решении данной задачи, а также задач классификации и кластеризации текстов. И каких успехов можно добиться при использовании сверточных нейронных сетей в основе своих алгоритмов, и при этом используя векторную и семантическую репрезентацию слов при их кодировании.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цели и задачи исследования

Целью данной работы является исследование методов глубокого обучения (deep learning) и его применений к задаче профилирования пользователей. Для достижения поставленной цели были решены следующие задачи:

– Произведен обзор современных методик профилирования пользователя, основанных на применении многослойных нейронных сетей, и их применений для задач обработки текстов и документов;

– Проведено исследование существующих реализаций нейронных сетей для задачи классификации текстов и способов векторного кодирования данных;

– Разработаны новые алгоритмы для использования в задаче профилирования пользователя;

– Произведено экспериментальное сравнение разработанных алгоритмов с наиболее часто используемым алгоритмом для профилирования пользователей с помощью нейронных сетей.

Научная новизна

Представленные в магистерской диссертации способы работы со сверточными нейронными сетями для задач профилирования пользователя и классификации текстов, наряду с разработанными алгоритмами векторной и семантической репрезентации показывают большую эффективность по сравнению с текущими наиболее популярными способами решения данных проблем и задач. Это показывает актуальность работы в выбранном направлении исследования.

Положения выносимые на защиту

1) В ходе научного исследования выявлено, что используемые на текущий момент методы профилирования пользователя не эффективны.

2) Причина низкой эффективности лежит в использовании устаревших методов для решения этой задачи при увеличении объемов информации, требуемых для обработки на текущий момент.

3) Для повышения эффективности при решении задачи профилирования пользователя следует использовать в своей основе сверточные нейронных сети.

4) В ходе исследования были разработаны два алгоритма: алгоритм векторной репрезентации на основе сверточных нейронных сетей и алгоритм семантической репрезентации на основе сверточных нейронных сетей.

Апробация результатов диссертации

Результаты исследования были представлены в «54-ой научной конференции аспирантов, магистрантов и студентов» учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Личный вклад соискателя ученой степени

Представленные в работе новые научные результаты получены автором лично. Научный руководитель А.Б. Гуринович принимала участие в постановке задач и обсуждении результатов. Разработка алгоритмов векторной и семантической репрезентации слов с помощью сверточных нейронных сетей, изложенных в данной работе, целиком и полностью лежит на авторе.

Опубликование результатов диссертации

Результаты исследования были опубликованы в виде материалов указанной выше научной конференции.

Структура и объем диссертации

Магистерская диссертация представлена в виде пояснительной записки на 62 страницах, состоящей из введения, шести разделов и заключения, списка использованных источников и приложения.

ОСНОВНАЯ ЧАСТЬ

В первом разделе приведено описание понятия семантической паутины, а также профилирования пользователя.

Семантическая паутина – это общедоступная глобальная семантическая сеть, формируемая на базе интернета путём стандартизации представления информации в виде, пригодном для машинной обработки. Семантическая паутина состоит из машинно-читаемых элементов – узлов семантической сети, с опорой на онтологии.

Информация о пользователе может быть представлена несколькими способами. Как правило, если мы говорим о более абстрактной и обобщенной информации, мы говорим о "профиле пользователя" или "модели пользователя", которые включают в себя основную характеристику пользователя и данные о поведении пользователя. С точки зрения данных, пользователь является ключевым источником получения мета-данных.

Второй раздел посвящён работе с веб-данными, описанию источников данных и их обработке.

Есть несколько видов данных, которые являются наиболее важными в веб-персонализации. Эти данные разделены на четыре основные категории:

- данные из журналов веб-доступа

- данные контента
- веб-структура данных сайта
- демографические данные.

Затем данные для персонализации подлежат предварительной подготовке.

Фаза подготовки данных может быть разделена на две фазы: (I) получение данных из интернета и (II) подготовка данных.

С данными производятся операции:
- Очистка от генерируемых автоматически данных
- Удаление записей, не отражающих активность пользователя
- Определение каждого отдельного пользователя
- Идентификация пользовательской сессии
- Нахождение полного пути
- Идентификация транзакции.

Третий раздел содержит некоторые подходы к автоматическому анализу информации на основе профиля пользователя. Здесь указываются способы классификации данных пользователя с требуемой перед этим обработкой, такой как: индексация, извлечение термов, назначение весов термам и перевод текста в векторное представление. А также в данном разделе разбирается кластеризация данных пользователя, и какие методы и алгоритмы могут для этого использоваться.

В четвёртом разделе приводится описание понятия нейронной сети и её архитектуры.

Искусственные нейронные сети представляют собой семейство моделей, построенных по принципу организации и функционирования биологических нейронных сетей – сетей нервных клеток живого организма.

Также в этом разделе описаны виды функций активации для модели искусственного нейрона, и описано понятие функции потерь для нейронной сети.

Пятый раздел посвящен описанию понятия и архитектуры сверточных нейронных сетей.

Сверточные нейронные сети были предложены Яном Лекуном и изначально нацелены на эффективное распознавание изображений. Свое название архитектура сети получила из-за наличия операции свёртки, суть которой в том, что каждый фрагмент изображения умножается на матрицу (ядро) свёртки поэлементно, а результат суммируется и записывается в аналогичную позицию выходного изображения. В архитектуру сети заложены априорные знания из предметной области компьютерного зрения: пиксель изображения

сильнее связан с соседним (локальная корреляция) и объект на изображении может встретиться в любой части изображения.

Успех применения сверточных нейронных сетей к классификации изображений привел к множеству попыток использовать данный метод к другим задачам. В последнее время их стали активно использоваться для задачи классификации текстов.

Сверточная нейронная сеть обычно представляет собой чередование сверточных слоев, субдискретизирующих слоев и при наличии полносвязных слоев на выходе. Все три вида слоев могут чередоваться в произвольном порядке.

В сверточном слое нейроны, которые используют одни и те же веса, объединяются в карты признаков, а каждый нейрон карты признаков связан с частью нейронов предыдущего слоя. При вычислении сети получается, что каждый нейрон выполняет свертку некоторой области предыдущего слоя (определяемой множеством нейронов, связанных с данным нейроном).

Пример архитектуры сверточной нейронной сети представлен на рисунке 1.

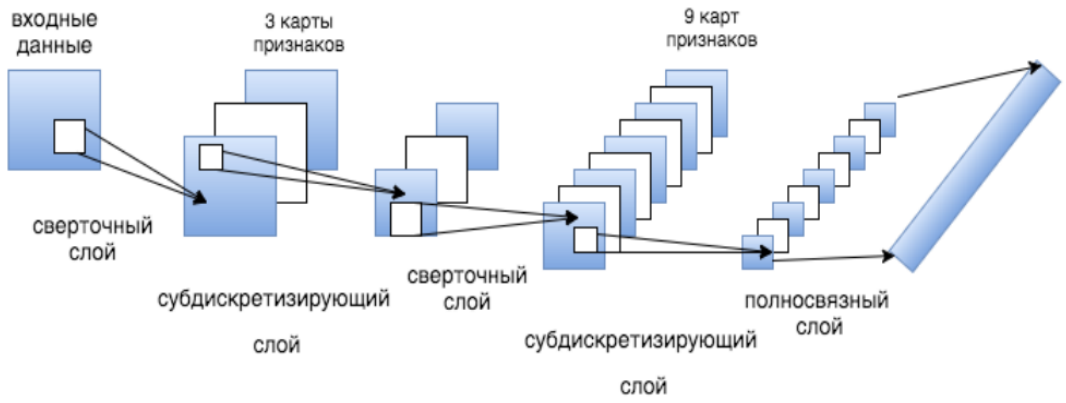

Рисунок 1: Архитектура сверточной нейронной сети

Далее в этом разделе подробно разбираются каждый в слой по отдельности, и приводятся примеры использования данной архитектуры для задачи классификации текстов.

- Посимвольный подход
- Подход с кодированием слов

При посимвольном подходе, каждый символ кодируется отдельно, пример сверточной сети с посимвольным подходом можно посмотреть на рис. 2.

В подходе с кодированием слов каждому слову в тексте сопоставляется вектор фиксированной длины, затем из полученных векторов для каждого объ-

екта выборки составляется матрица, которая аналогично изображениям подается на вход сверточной нейронной сети. На рисунке 3 приведен пример сверточной нейронной сети с использованием кодирования слов.

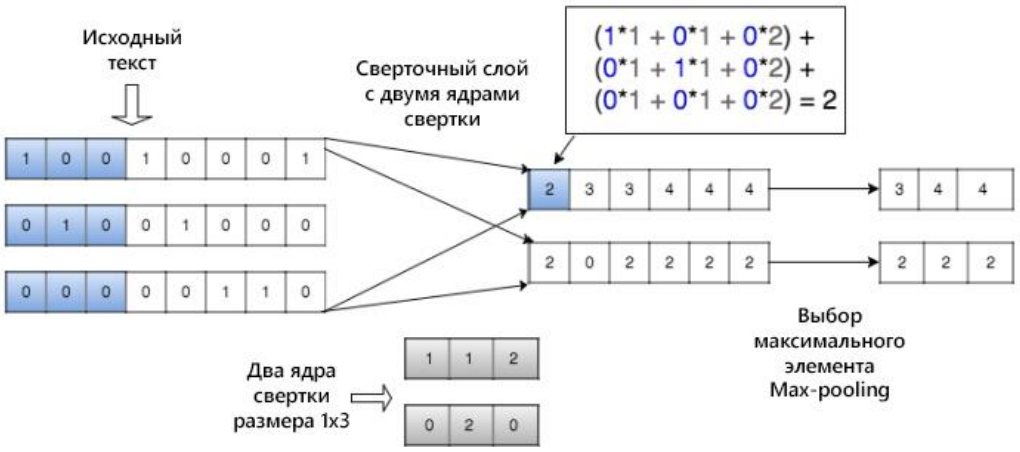

Рисунок 2: Посимвольный подход

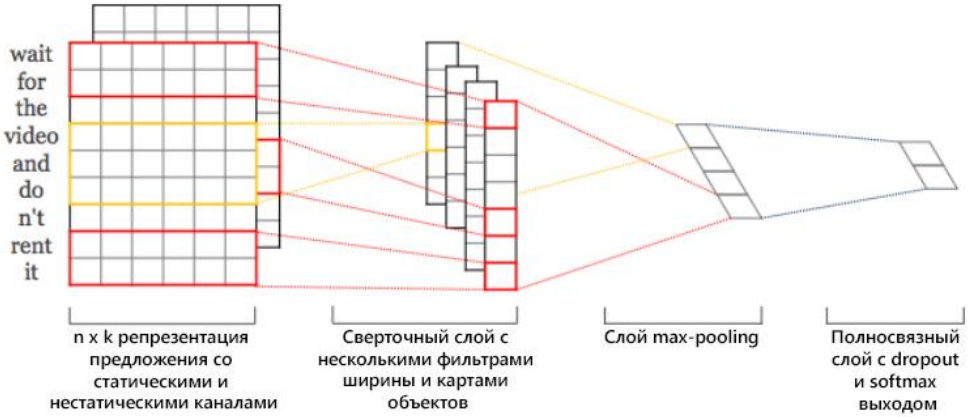

Рисунок 3: Кодирование слов

В шестом разделе описаны разработанные алгоритмы и приведены результаты их работы.

При профилировании пользователей, а точнее классификации текстов с использованием сверточной нейронной сети:

1) фильтр имеет такую же ширину m, как матрица (на вход нейронной сети поступают целые слова);

2) можно применять одновременно разные фильтры с разными высотами для выделения различных признаков.

Сверточная нейронная сеть является очень эффективным методом представления текстов. Поэтому при работе с ней мы будем использовать два алгоритма классификации текстов: векторной репрезентации и семантической репрезентации.

Алгоритм при векторной репрезентации слов и текстов

Есть множество методов и технологий представления текстовой информации в виде векторного представления: GloVe, AdaGram, Text2Vec, Seq2Vec и другие, но наиболее популярной технологией является Word2Vec.

Word2Vec – технология от компании Google, которая заточена на статистическую обработку больших массивов текстовой информации. Он собирает статистику по совместному появлению слов в фразах, после чего методами нейронных сетей решает задачу снижения размерности и выдает на выходе компактные векторные представления слов, в максимальной степени отражающие отношения этих слов в обрабатываемых текстах.

В Word2Vec можно использовать две различных архитектуры нейронной сети для перевода слова в вектор: Continuous Bag of Words и Skipgram. Модель Continuous Bag of Words(CBOW) представляет собой метод поиска ассоциированных слов, подробнее рисунок 4. Модель Skipgram – поиска взаимозаменяемых слов, смотрите рисунок 5.

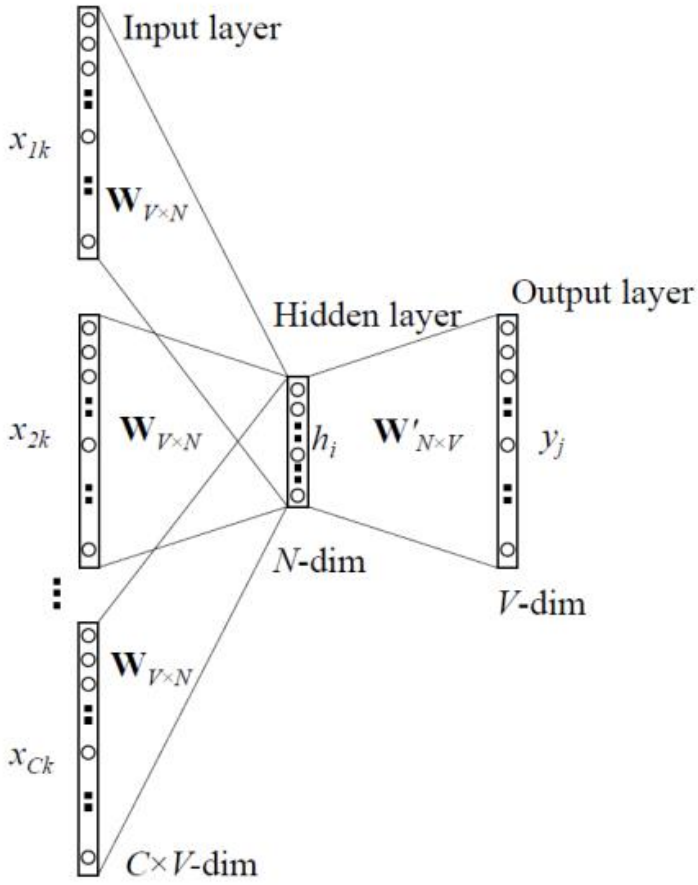

Рисунок 4: Общий случай модели CBoW

Skipgram модель работает медленнее, но обычно с помощью нее достигается лучшее качество классификации текстов.

В моем случае используется сверточная нейронная сеть с векторным представлением слов по данной технологии, таким образом получая совершенно новый способ для профилирования пользователя и классификации текстов.

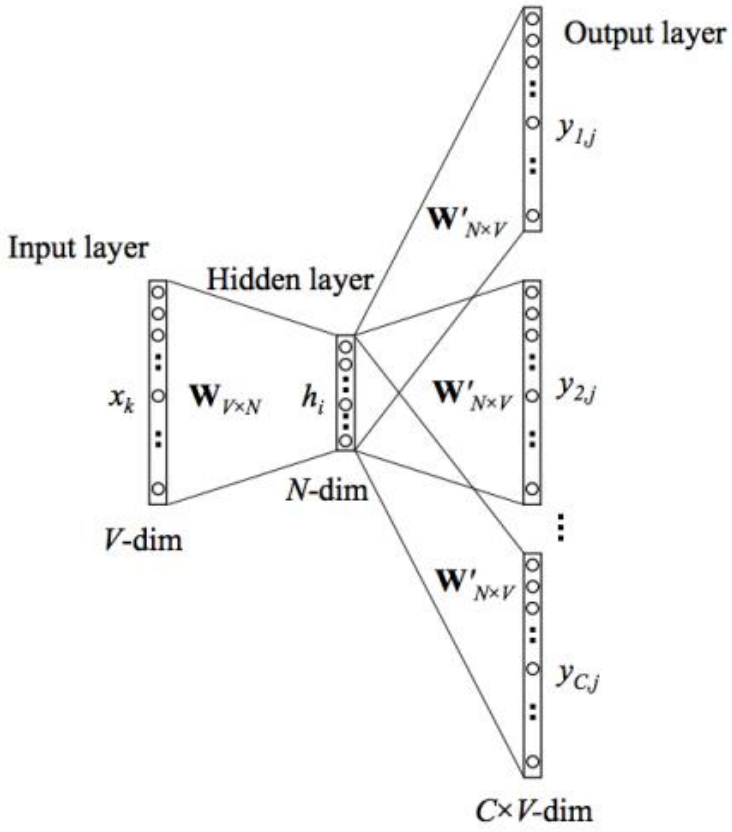

Рисунок 5: Общий случай модели Skipgram

Преимуществами являются:

– Сверточная нейронная сеть с данным алгоритмом учится весьма быстро на неподготовленных текстах.

– Снижаемая размерность, т.е. словарь с 2.5 млн. токенов ужимается до 256 элементов вектора действительных чисел;

Негативным эффектом векторной репрезентации является быстрая деградация векторов при операциях над ними. Сложить слова одной фразы ещё выполнимо, сложить слова нескольких фраз уже нет.

Алгоритм при семантической репрезентации слов и текстов

Учтя преимущества и недостатки предыдущего метода, проведено избавление от вторых, таким образом повысив эффективность.

Для этого из векторных репрезентаций слов создаётся семантический вектор смыслов слов. Чтобы это сделать, проводится кластеризация вектора наших слов. В качестве алгоритма кластеризации в данной работе наиболее

подходящим под условия задачи был выбран алгоритм k-means++, который является улучшением алгоритма k-means. Суть улучшения k-means++ по сравнению с его предком заключается в нахождении более «хороших» начальных значений центроидов кластеров. Оригинальный k-means не регламентирует то, как выполняется этот этап алгоритма, и поэтому является нестабильным.

В нашем случае для алгоритма семантической репрезентации количество устойчивых кластеров выбирается таким образом, чтобы однозначно отражать стиль и тематику текста. Т.е. если оно будет большим, можно ожидать, что каждый кластер будет указывать на достаточно узкую тематику текста, а точнее – на узкий признак тематики или стиля. Длина семантического вектора текста численно равна количеству кластеров, и каждый элемент этого вектора может быть объяснён через соответствующие данному кластеру слова.

Относительная плотность слов из каждого кластера в исследуемом тексте хорошо описывает текст. Разумеется, каждое конкретное слово имеет отношение ко многим кластерам, к каким-то больше, к каким-то меньше. Поэтому, в первую очередь необходимо вычислить семантический вектор слова – вектор, описывающий расстояние от слова до центра соответствующего кластера в векторном пространстве Word2Vec, при этом отбрасываются значения векторов менее 0. Полученные расстояния до центра и есть искомый семантический вектор. Каждый элемент данного вектора имеет свой смысл, задаваемый теми словами, что образуют соответствующий кластер. Сложение таких векторов деградирует намного медленнее сложения оригинальных репрезентаций слов.

На получении исключительно семантических векторов слов не останавливаемся. Производим сложение семантических векторов отдельных слов, составляющих текст, получая семантический вектор всего текста. Так как каждому тексту поставлен в соответствие вектор в семантическом пространстве, возможно вычислить расстояние между любыми двумя текстами как косинусную меру между ними. Имея расстояние между текстами, можно провести классификацию в векторном пространстве текстов, а не отдельных слов. Это необходимо для фильтрации самих текстов согласно требуемым тематикам профилирования.

Если описать чуть подробнее, то при наборе документов, поделённых на два или более классов. Каждый из документов будет обладать семантическим вектором, результатом сложения векторов, входящих в него слов, отнесённых к кластерам. Вычислив математическое ожидание и дисперсию каждого элемента суммарного вектора по всей обучающей выборке документов, и после

отнормировав вектора документов, получается весьма хорошая сигнатура документа. Она интерпретируема в человеческих терминах, где каждый элемент – это количество слов, ассоциированных с соответствующей темой в документе. При количестве слов больше нуля, значит в документе тема поднимается чаще, чем в среднем. Меньше нуля – реже. При значении -1 тема не поднимается вовсе.

Результаты работы алгоритмов

Тестирование алгоритмов профилирования пользователя проводились на данных указанных в таблице ниже.

Таблица 1: Данные

Выборка	Число классов	Размер обучающей выборки	Размер тестовой выборки
Ag News	4	100.000	7600
DBPedia	13	500.000	60.000
Amazon Review Full	5	3.000.000	500.000

1. Ag news − новостные интернет-статьи. Объем обучающей выборки 100.000 объектов, объем тестовой выборки 8000 объектов. Статьи необходимо классифицировать на 4 класса − мировые, спортивные, бизнес и научные новости.
2. DBPedia − название и аннотации статей из Википедии. Объем обучающей выборки 500.000 объектов, объем тестовой выборки 60.000 объектов. Тексты необходимо классифицировать на 13 классов − компания, образовательное учреждение, политик, спортсмен, актер, средство передвижения, здание, природное место, город, животное, художественное произведение, фильм, литературное произведение.
3. Amazon Review Full − комментарии с сайта Amazon.com. Объем обучающей выборки 3.000.000 объектов, объем тестовой выборки 500.000 объектов. Тексты необходимо классифицировать на 5 классов − отзывы пользователей от отрицательного до положительного по пятибалльной шкале

Параметры использованной свёрточной нейронной сети для работы с представленными алгоритмами были заданы такими: по 128 фильтрам высотой 3, 4 и 5, размер векторов слов n = 256, число эпох обучения = 100. Обучение и тестирование были реализованы на основе работы Denny Britz с использованием библиотеки TensorFlow.

В качестве алгоритма для сравнения эффективности был выбран стандартный для профилирования пользователей и классификации текстов: Bag of Words & TF IDF.

Таблица 2: Результаты

Данные	Bag of Words & TF IDF	CW2V	CSR
Ag News	0.878	0.861	0.925
DBPedia	0.922	0.953	0.989
Amazon Review Full	0.552	0.569	0.613

Обозначения, использованные в таблице 2:

– Bag of Words & TF IDF – наиболее популярный способ перевода текста в векторное представление с помощью обычной нейронной сети.

– CW2V – свёрточная нейронная сеть с кодированием слов и векторной репрезентацией слов с помощью алгоритма Word2Vec;

– CSR – свёрточная нейронная сеть с использованием семантической репрезентации.

Из таблицы 2 видно, что точность классификации при использовании сверточных нейронных сетей и конкретно алгоритма семантической репрезентации гораздо выше, чем при использовании алгоритма векторной репрезентации, и тем более они оба превосходят наиболее популярный алгоритм классификации текстов Bag of Words & TF IDF с помощью обычных нейронных сетей. Хоть, тот и показал себя лучше на отдельной выборке по сравнению с алгоритмом векторной репрезентации, однако общая эффективность на всех выбранных данных его значительно ниже, чем разработанные в ходе работы алгоритмы. Это достигается тем, что при разработке данных алгоритмов мы использовали сверточную нейронную сеть, которая имеет указанные преимущества перед полносвязными нейронными сетями.

Таким образом, можно смело утверждать, что разработанные алгоритмы показали свою состоятельность и подходят для практического применения в задачах классификации текстов, а также профилирования пользователей.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

– В ходе научного исследования был произведен обзор современных методик профилирования пользователя, основанных на применении многослойных нейронных сетей, и их применений для задач обработки текстов и документов, и было выявлено, что те являются на текущий момент с многократной возрастающей информацией на входе не эффективны;

– Проведены исследования существующих реализаций сверточных нейронных сетей для обработки текстов, а также нейронных сетей с векторным представлением слов и текстов, на основе технологии Word2Vec;

– Разработаны новые алгоритмы: алгоритм векторной репрезентации и алгоритм семантической репрезентации, построенные на основе сверточных нейронных сетей и использующих векторное представление данных для использования в задаче профилирования пользователя;

– Произведено экспериментальное сравнение разработанных алгоритмов с наиболее часто используемым алгоритмом для профилирования пользователей с помощью нейронных сетей.

Рекомендации по практическому использованию результатов

В результате исследования, при реализации подходов из данной работы и использовании разработанных алгоритмов становится возможным:

– повысить эффективность поиска и получения информации для каждого пользователя сети Интернет. Это стало возможным благодаря предварительной обработке разработанными в этом исследовании алгоритмами его предпочтений и предыдущих поисковых запросов;

– повысить производительность средств распространения рекламных информационных материалов в сети Интернет и эффективность рекламного и информационного воздействия на пользователей с помощью точного определения желаний пользователя на основе его профиля;

– могут быть использованы в качестве методов для борьбы со спамом, а точнее его распознавании при проверке получаемых писем с помощью определения тематики и стиля их написания. Ведь многие спам-письма используют одинаковые шаблоны;

– улучшить фильтрацию документов как по автору, поднимаемой теме в тексте, так и по художественному стилю или стилистике написания предложений;

– добавить персонализацию информации при автоматическом переводе текстов, выявление смысловых намеков в переводимом тексте и добавлении

стилистических и языковых особенностей пользователя-переводчика, тем самым облегчая труд и адаптацию машинного перевода;

– помочь в навигации по большим информационным ресурсам со сложной структурой, благодаря разделению сайта по темам (кластерам) со схожим смыслом;

– улучшить индексацию поисковых запросов, используя подбор предложений по аналогичным запросом из смежных синонимичных тем;

– повысить точность автоматического аннотирования и реферирования текстов, так как при использовании алгоритма тема текста будет определяться гораздо точнее и при этом будут указываться поднимаемые еще в данном документе темы, а также предлагаться смежные темы к уже указанным;

– и др.

Таким образом, сфера возможного применения результатов научного исследования весьма широка, что показывает высокую ценность проделанной работы.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

[1] Ковалевский, А.М. Влияние сети 4G на здоровье человека / А.М. Ковалевский // Компьютерное проектирование и технология производства электронных систем: сб. материалов 52-ой науч. конф. аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» (Минск, 25-30 апреля 2016 года). – Минск: БГУИР, 2016

[2] Ковалевский, А.М. Автоматизированная информационная система развлекательных мероприятий / А.М. Ковалевский // Информационные технологии и управление: сб. материалов 53-ой науч. конф. аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» – Минск: БГУИР, 2017

[3] Ковалевский, А.М. Сверточные нейронные сети в решении задачи профилирования / А.М. Ковалевский // Информационные технологии и управление: сб. материалов 54-ой науч. конф. аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлек-троники» – Минск: БГУИР, 2018