УТВЕРЖДАЮ

Проректор по исследованиям и разработкам
Московского физико-технического института (государственного университета)

Гаричев Сергей Николаевич

«4» мая 2018 г

ОТЗЫВ

ведущей организации на диссертацию **Александра Александровича Пономаренко** «Исследования и разработка алгоритмов поиска в распределенных масштабируемых хранилищах данных», представленную на соискание учёной степени кандидата технических наук по специальности 05.13.17 «Теоретические основы информатики»

1. Актуальность темы

В настоящий момент времени скорость роста информации доступной в цифровом виде значительно опережает закон Мура. В свою очередь, требования индустрии диктуют необходимость использования алгоритмов машинного обучения применительно к большим массивам данных. При этом для обработки данных в реальном времени требуется, чтобы, вычислительная сложность алгоритмов не превышала полилогарифмическую. Одним из широко используемых методов машинного обучения, особенно важным для решения задачи многоклассовой классификации, является метод k-ближайших соседей. Существующие методы поиска k-ближайших соседей с

полилогарифмической вычислительной сложностью имеют ограниченную сферу применения и могут быть использованы только для векторных пространств. В этой связи диссертационная работа Пономаренко А. А., посвященная алгоритмам поиска k-ближайших соседей не требующих векторного представления информационных объектов, представляется актуальной.

Важным новым элементом диссертация является то, что в качестве основы для алгоритмов поиска предлагается использовать специальный класс графов, что делает возможным использовать алгоритмы, эффективно работающие в распределенной среде.

2. Новизна проведённых исследований и полученных результатов

К наиболее значимым новым научным результатам работы может быть отнесено следующее:

- Предложена математическая модель оптимальной конфигурации рёбер графа для поиска ближайшего соседа алгоритмом жадного направленного поиска GreedyWalk.
- Предложен и исследован алгоритм построения графа MSWConstruction, позволяющий для метрического пространства строить граф, обладающий навигационными свойствами тесного мира.
- Предложен и исследован алгоритм поиска k-ближайших соседей K-NNSearch, основанный на построенной структуре графа с навигационными свойствами тесного мира.
- Разработан алгоритм улучшения поисковых свойств графа на этапе исполнения запросов RepairByQuery

Особого внимания заслуживает алгоритм формирования графа MSWConstruction. За его внешней простой скрывается несколько оригинальных эффективных решений:

1) использовать k-граф в качестве аппроксимации графа Делоне,

2) использовать стратегию добавления вершины в граф одну за другой в случайном порядке, оставляя при этом рёбра по принципу рандомизированного динамического способа формирования графа, что обеспечивает навигационные свойства тесного мира.

К сильной стороне работы можно также отнести предложенную в четвёртой главе математическую модель оптимальной конфигурации рёбер для фиксированного множества точек. Данный подход видится новым и оригинальным для изучения задачи поиска ближайшего соседа.

3. Степень обоснованности и достоверности полученных результатов и выводов

Достоверность полученных результатов подтверждается корректным использованием математического аппарата, а также воспроизводимыми результатами экспериментов на общедоступных наборах данных. Полученные результаты сравнивались с имеющимися аналогами с использованием общедоступной библиотеки с открытым кодом non-metric space library.

4. Краткая характеристика работы

Диссертационная работа состоит из введения, пяти глав, заключения и библиографического списка. Библиографический список содержит 69 наименований. Общий объём работы 136 страниц.

Во введении обоснована актуальность выбранной темой, сформулированы основные цели и задачи исследования.

В первой главе приведён содержательный обзор работ, касающихся алгоритмической составляющей построения распределённых хранилищ информации на основе структурированных p2p сетей.

Во второй главе предлагаются новые алгоритмы для построения графов обладающих свойством навигационного тесного мира, а также

предлагаются новый алгоритм для приближенного поиска k-ближайших и алгоритм для улучшения поисковых свойств графа на этапе исполнения запросов.

В третьей главе экспериментально исследуется свойства предложенных алгоритмов. Рассматриваются следующие характеристики графов: диаметр, коэффициент кластеризации, распределение длин эффективных путей. Точность и вычислительная сложность алгоритма поиска k-ближайших сравнивается с имеющимися аналогами.

Четвёртая глава посвящена модели математического программирования описывающей оптимальную конфигурацию рёбер графа. При этом предполагается, что в качестве алгоритма поиска используется жадный направленный поиск (GreedyWalk). Модель также накладывает ограничения, чтобы поиск от произвольной вершины заканчивался успехом. Приводятся решения модели полученные точными и эвристическими методами для некоторых случаев целочисленной решётки.

В пятой главе описываются технические нюансы программы, которая использовалась для проведения численных экспериментов.

5. Публикация и апробация материалов диссертации

По теме диссертации опубликовано 16 работ, в том числе 6 работ в изданиях, рекомендованных ВАК, автором получено два свидетельства о государственной регистрации программ для ЭВМ.

6. Значимость результатов для науки и производства

Значимость подтверждается многочисленным цитированием работ в международном сообществе. Статья «Approximate nearest neighbor algorithm based on navigable small world graphs», в которой были опубликованы алгоритмы MSWConstruction и K-NNSearch к настоящему моменту времени была процитирована уже 55 раз.

Используя результаты диссертационной работы перспективным представляется создание распределённых систем хранения и поиска информации, а также улучшение существующих систем много классовой классификации за счёт сокращения времени необходимого для поиска k-ближайших соседей.

Кроме этого результаты работы используются в учебном процессе НИУ ВШЭ при подготовке и проведении курсов магистерской программы «Интеллектуальный Анализ Данных»

7. Рекомендация по внедрению результатов

Полученные в диссертации результаты могут быть использованы как для создания масштабируемых хранилищах данных, так и для построения распределённых децентрализованных систем с нетривиальной поисковой функциональностью. Результаты диссертации рекомендуются к внедрению в компаниях занимающихся хранением и обработкой, в том числе классификацией, больших массивов данных.

8. Замечания

1. Работа имеет экспериментальную направленность, однако многие вопросы остаются открытыми и нуждаются в теоретическом исследовании. Например, остаётся открытым вопрос, почему алгоритм поиска knn-search имеет наблюдаемую полилогифмическую вычислительную сложность?

2. Интерес также представляет поведение алгоритма поиска в зависимости от параметров характеризующих сложность множества в паре с функцией расстояния. При этом возможно использование такой характеристики множества как intrinsic dimensionality (Chávez E. et al. Searching in metric spaces //ACM computing surveys (CSUR). – 2001. – Т. 33. – №. 3. – С. 273-321), doubling dimension (Krauthgamer R., Lee J. R. The black-box complexity of nearest-neighbor search //Theoretical Computer Science. – 2005. – Т. 348. – №. 2-3. – С. 262-276.). Для случая, когда абсолютные

расстояния не важны, было бы интересно использовать комбинаторный фреймворк предложенный в статье (Goyal N., Lifshits Y., Schütze H. Disorder inequality: a combinatorial approach to nearest neighbor search //Proceedings of the 2008 International Conference on Web Search and Data Mining. – ACM, 2008. – С. 25-32.).

3. Теоретически не исследован выбор k-ближайших в алгоритме построения графа MSWConstruction. В частности, является ли выбор k-ближайших соседей алгоритма в некотором смысле оптимальным способом для формирования множества рёбер?

4. Имеются некоторые неточности и описки при изложении. Например: пропущена запятая в первом предложении второго абзаца на странице 95; на странице 22 в первом предложении опечатка в слове «последовательно»; на странице 3 в последнем абзаце в 4-м предложении пропущена запятая после слова «алгоритм».

Сделанные замечания не затрагивают основного содержания диссертационного исследования, не снижают его теоретической ценности и практической значимости.

Работы выстроена логично, её структура отражает цели и задачи исследования. Автореферат полностью раскрывает основные положения диссертации.

В целом можно заключить, что диссертация А. А. Пономаренко является законченной научно-квалификационной работой, в которой изложены новые научно обоснованные технические решения, внедрение которых представляет собой значительный вклад в развитие информационно-телекоммуникационных технологий.

Автореферат полностью и точно отражает содержание диссертации.

Представленная диссертационная работа удовлетворяет всем требования, предъявляемым ВАК РФ к диссертации на соискание кандидата технических наук, а её автор Пономаренко Александр Александрович,

заслуживает присуждение учёной степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики»

Отзыв на диссертацию обсуждён и одобрен на расширенном заседании кафедры дискретной математики МФТИ «04» мая 2018 г., протокол № 05/04

Заведующий кафедрой
дискретной математики МФТИ,
д-р физ. – мат. наук, профессор

Райгородский
Андрей Михайлович

Почтовый адрес: 141700, Московская область, г. Долгопрудный, Институтский пер.,9
Телефон: 8 (495) 408 408-78-81
Адрес электронной почты: raigorodskii.am@mipt.ru
Организация – место работы: федеральное государственное автономное образовательное учреждение высшего образования «Московский физико–технический институт (государственный университет)», кафедра дискретной математики
Должность: заведующий кафедрой
Web-сайт организации: https://mipt.ru/