

УДК 004.021
ББК 32.973.1
Н60

Н60 Марк Нидхем, Эми Ходлер

Графовые алгоритмы. Практическая реализация на платформах Apache Spark и Neo4j. / пер. с англ. В. С. Яценкова – М.: ДМК Пресс, 2020. – 258 с.: ил.

ISBN 978-5-97060-799-2

Каждую секунду во всем мире собирается и динамически обновляется огромный объем информации. Графовые алгоритмы, которые основаны на математике, специально разработанной для изучения взаимосвязей между данными, помогают разобраться в этих гигантских объемах. И, что особенно важно в наши дни, они улучшают контекстную информацию для искусственного интеллекта.

Эта книга представляет собой практическое руководство по началу работы с графовыми алгоритмами. В начале описания каждой категории алгоритмов приводится таблица, которая поможет быстро выбрать нужный алгоритм и ознакомиться с примерами его использования.

Издание предназначено для разработчиков и специалистов по анализу данных. Для изучения материала книги желателен опыт использования платформ Apache Spark™ или Neo4j, но она пригодится и для изучения более общих понятий теории графов, независимо от выбора графовых технологий.

УДК 004.021
 ББК 32.973.1

Original English language edition published by O'Reilly Media, Inc. Copyright © 2019 Mark Needham and Amy E. Hodler. All rights reserved. Russian-language edition copyright © 2019 by DMK Press. All rights reserved.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

Материал, изложенный в данной книге, многократно проверен. Но, поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

ISBN 978-1-49204-768-1 (англ.)
 ISBN 978-5-97060-799-2 (рус.)

© 2019 Amy Hodler and Mark Needham.
 © Оформление, перевод на русский язык, издание,
 ДМК Пресс, 2020

Оглавление

Предисловие.....	10
Вступительное слово рецензента	14
Глава 1. Введение.....	17
Что такое графы?.....	18
Что такое графовые алгоритмы и анализ графов?	19
Обработка графов, базы данных, запросы и алгоритмы.....	22
OLTP и OLAP.....	23
Почему мы должны изучать графовые алгоритмы?	25
Где и когда применяется анализ графов?	29
Заключение	30
Глава 2. Теория и концепции графов	31
Терминология.....	31
Типы и структуры графов.....	32
Случайные, локализованные и безмасштабные сети	32
Разновидности графов	34
Связные и несвязные графы.....	35
Невзвешенные и взвешенные графы.....	35
Ненаправленные и ориентированные графы	36
Ациклические и циклические графы.....	38
Разреженные и плотные графы.....	39
Однокомпонентные, двудольные и k -дольные графы.....	40
Типы графовых алгоритмов.....	42
Поиск пути	42
Определение центральности.....	43
Обнаружение сообщества	43
Заключение	44
Глава 3. Графовые платформы и обработка.....	45
Графовые платформы и особенности обработки	45
Подходы к выбору платформы	45
Подходы к обработке данных	46
Объединяющие платформы	47
Выбор платформы	48
Apache Spark.....	49
Графовая платформа Neo4j.....	51
Заключение	54

Глава 4. Алгоритмы поиска по графу и поиска пути	55
Пример данных: транспортный граф	58
Импорт данных в Apache Spark	59
Импорт данных в Neo4j	60
Поиск в ширину	61
Поиск в ширину с помощью Apache Spark	61
Поиск в глубину	63
Кратчайший путь	64
Когда следует использовать алгоритм кратчайшего пути?	66
Реализация алгоритма кратчайшего пути с Neo4j	67
Поиск кратчайшего взвешенного пути с Neo4j	69
Поиск кратчайшего взвешенного пути с Apache Spark	70
Вариант алгоритма кратчайшего пути: A^*	73
Вариант алгоритма кратчайшего пути: k -кратчайший путь Йена	75
Алгоритм кратчайшего пути между всеми парами вершин	76
Подробный разбор алгоритма APSP	77
Когда следует использовать APSP?	79
Реализация APSP на платформе Apache Spark	79
Реализация APSP на платформе Neo4j	80
Кратчайший путь из одного источника	82
Когда следует использовать алгоритм SSSP?	83
Реализация алгоритма SSSP на платформе Apache Spark	83
Реализация алгоритма SSSP на платформе Neo4j	86
Минимальное остовное дерево	87
Когда следует использовать минимальное остовное дерево?	88
Реализация минимального остовного дерева на платформе Neo4j	89
Алгоритм случайного блуждания	91
Когда следует использовать алгоритм случайного блуждания?	91
Реализация алгоритма случайного блуждания на платформе Neo4j	92
Заключение	93
Глава 5. Алгоритмы вычисления центральности	94
Пример графовых данных – социальный граф	96
Импорт данных в Apache Spark	98
Импорт данных в Neo4j	98
Степенная центральность	98
Охват вершины	99
Когда следует использовать степенную центральность?	100
Реализация алгоритма степенной центральности с Apache Spark	100
Центральность по близости	102
Когда следует использовать центральность по близости?	103
Реализация алгоритма центральности по близости с Apache Spark	103
Реализация алгоритма центральности по близости с Neo4j	106

Вариант центральности по близости: Вассерман и Фауст.....	107
Вариант центральности по близости: гармоническая центральность.....	109
Центральность по посредничеству.....	110
Когда следует использовать центральность по посредничеству?.....	113
Реализация центральности по посредничеству с Neo4j.....	113
Вариант центральности по посредничеству: алгоритм Брандеса.....	116
PageRank.....	118
Влияние.....	118
Формула алгоритма PageRank.....	119
Итерация, случайные пользователи и ранжирование.....	119
Когда следует использовать PageRank?.....	122
Реализация алгоритма PageRank с Apache Spark.....	122
Реализация алгоритма PageRank с Neo4j.....	125
Вариант алгоритма PageRank: персонализированный PageRank.....	126
Заключение.....	127
Глава 6. Алгоритмы выделения сообществ.....	128
Пример данных: граф зависимостей библиотек.....	131
Импорт данных в Apache Spark.....	132
Импорт данных в Neo4j.....	133
Подсчет треугольников и коэффициент кластеризации.....	134
Локальный коэффициент кластеризации.....	134
Глобальный коэффициент кластеризации.....	135
Когда следует использовать подсчет треугольников и коэффициент кластеризации?.....	136
Реализация подсчета треугольников с Apache Spark.....	136
Реализация подсчета треугольников с Neo4j.....	137
Локальный коэффициент кластеризации с Neo4j.....	137
Сильно связанные компоненты.....	139
Когда следует использовать сильно связанные компоненты?.....	140
Реализация поиска сильно связанных компонентов с Apache Spark.....	141
Реализация поиска сильно связанных компонентов с Neo4j.....	142
Связанные компоненты.....	144
Когда следует использовать связанные компоненты?.....	144
Реализация алгоритма связанных компонентов с Apache Spark.....	145
Реализация алгоритма связанных компонентов с Neo4j.....	145
Алгоритм распространения меток.....	147
Обучение с частичным привлечением учителя и начальные метки.....	148
Когда следует использовать распространение меток?.....	149
Реализация алгоритма распространения меток с Apache Spark.....	150
Реализация алгоритма распространения меток с Neo4j.....	151
Лувенский модульный алгоритм.....	152
Когда следует использовать Лувенский алгоритм?.....	157
Реализация Лувенского алгоритма с Neo4j.....	158

Проверка достоверности сообществ.....	162
Заключение	162
Глава 7. Графовые алгоритмы на практике	164
Анализ данных Yelp на платформе Neo4j	165
Социальная сеть Yelp.....	165
Импорт данных.....	166
Графовая модель.....	166
Краткий обзор данных Yelp	167
Приложение для планирования поездки	171
Туристический бизнес-консалтинг.....	177
Поиск похожих категорий	182
Анализ данных о рейсах авиакомпании с помощью Apache Spark.....	187
Предварительный анализ	188
Популярные аэропорты	189
Задержки вылетов из аэропорта Чикаго.....	190
Плохой день в Сан-Франциско	193
Взаимосвязи аэропортов через авиакомпанию	194
Заключение	201
Глава 8. Графовые алгоритмы и машинное обучение	202
Машинное обучение и важность контекста.....	202
Графы, контекст и точность	203
Извлечение и отбор связанных признаков.....	205
Графовые признаки.....	207
Признаки и графовые алгоритмы.....	207
Графы и машинное обучение на практике: прогнозирование связей ..	209
Инструменты и данные.....	210
Импорт данных в Neo4j.....	211
Граф соавторства	213
Создание сбалансированных наборов данных для обучения и тестирования	214
Как мы предсказываем недостающие связи	220
Разработка полного цикла машинного обучения.....	221
Прогнозирование связей: основные признаки графа	222
Прогнозирование связей: треугольники и коэффициент кластеризации	235
Прогнозирование связей: выделение сообществ	239
Заключение	245
Итог книги	246
Приложение А. Дополнительная информация и ресурсы	247
Дополнительные алгоритмы.....	247
Массовый импорт данных Neo4j и Yelp.....	248

АРОС и другие инструменты Neo4j	249
Поиск наборов данных	249
Помощь в освоении платформ Apache Spark и Neo4j.....	250
Дополнительные курсы	250
Об авторах	252
Об изображении на обложке.....	253
Предметный указатель	254