Введение в информационный поиск

Введение в информационный поиск — это первый учебник, который содержит взаимосвязанное изложение проблем классического информационного поиска и поиска в вебе, включая смежные задачи классификации и кластеризации текстов. Учебник написан с позиций информатики и содержит современное изложение всех аспектов проектирования и реализации систем сбора, индексирования и поиска документов, методов оценки таких систем, а также введение в методы машинного обучения на базе коллекций текстов.

Несмотря на то что учебник задуман как вводный курс по информационному поиску, он будет интересен исследователям и профессионалам. Полный набор слайдов для лекций и упражнений, сопровождающих книгу, доступен в Интернете.

Кристофер Д. Маннинг — профессор информатики в Стэнфордском университете (Stanford University).

Прабхакар Рагхаван — директор исследовательского департамента корпорации Yahoo! Research и профессор-консультант по компьютерным наукам Стэнфордского университета.

Хайнрих Шютце — заведующий кафедрой теоретической вычислительной лингвистики Института обработки текстов на естественных языках (Штутгардский университет).

Предисловие к русскому изданию

Мы рады предоставленной возможности написать краткое предисловие к русскому изданию книги *Introduction to Information Retrieval*. Поскольку оригинальное издание вышло в 2008 году, важность поиска по текстовым и другим неструктурированным информационным источникам к текущему моменту еще больше возросла. Этот поиск важен и как техническая задача, и как главная часть социального и делового взаимодействия людей в современном информационном мире. Прошедший период ознаменовался широким распространением блогов, микроблогов и социальных сетей, а также все более широким применением инструментов, использующих методы машинного обучения и более глубокую интерпретацию текстов. В частности, в России это было восхитительное и, вероятно, поворотное время появления успешных компаний, занимающихся веб-поиском, оптическим распознаванием символов и автоматической обработкой текста. Кроме того, за прошедшее время активизировалась организационная и академическая деятельность. Мы хотели бы отметить, в частности, семинар РОМИП, упомянутый в главе 8, который организовал форум по оценке методов информационного поиска в русскоязычных источниках (`http://romip.ru/`), аналогичный конференциям TREC, а также недавно организованную Российскую летнюю школу по информационному поиску. Мы надеемся, что публикация этой книги будет способствовать дальнейшему развитию методов информационного поиска и веб-поиска в русскоязычном мире.

Кристофер Д. Маннинг
Прабхакар Рагхаван
Хайнрих Шютце

Предисловие редакторов перевода

Информационный поиск, родившись на стыке библиотечного дела и информатики в середине XX века, некоторое время оставался скромной научной и прикладной областью, в которой работало небольшое количество ученых. Бурный рост интернета в конце прошлого — начале нынешнего века стал мощнейшим стимулом развития дисциплины. Современный информационный поиск — это миллионы пользователей, огромные объемы данных, мощные вычислительные системы, изощренные алгоритмы. Для решения изначальной задачи — поиска информации, соответствующей потребности пользователя, — привлекаются методы машинного обучения, анализа мультимедийной информации, компьютерная лингвистика, геоинформационные сервисы, исследуется психология пользователей и их социальные связи, удобство интерфейсов и т.д.

Создание учебника для такой динамичной и разносторонней дисциплины становится очень сложной задачей. Кристофер Маннинг, Прабхакар Рагхаван и Хайнрих Шютце с этой задачей прекрасно справились. Книга представляет собой сбалансированное, последовательное и основательное введение в предметную область. До книги *Введение в информационный поиск*, оригинальное издание которой вышло в 2008 году, основными учебниками по информационному поиску были книги 1999 года: Witten et al., *Managing Gigabytes* и Baeza-Yates и Ribeiro-Neto, *Modern Information Retrieval* (см библиографию). С русскоязычной учебной и профессиональной литературой по информационному поиску дело обстояло намного хуже. За исключением отдельных вузовских учебных пособий и переводных изданий узкоприкладного характера, основательных книг по информационному поиску на русском языке не выходило с начала 80-х годов прошлого века.

Благодаря интернет-магазинам не составляет большого труда стать обладателем оригинального английского издания, к тому же электронная версия книги свободно доступна по адресу `http://www.informationretrieval.org/`. Несмотря на это, мы считаем, что издание книги на русском языке — важное и полезное дело. Русская версия призвана упросить знакомство с информационным поиском всем заинтересованным — студентам, аспирантам, исследователям и инженерам-практикам. Профессионалам, работающим в этой области, книга поможет структурировать их знания и послужит аннотированным библиографическим указателем. Еще одна важная задача книги — зафиксировать (а иногда и ввести в оборот) русскую терминологию по информационному поиску. Отсутствие профессиональной литературы по информационному поиску в течение долгого времени обусловило сложности, с которыми мы столкнулись в процессе перевода.

При переводе терминологии мы старались по возможности использовать устоявшиеся математические термины, термины, принятые в отечественной информатике (computer science), и те, что стали общепринятыми в практике разработки поисковой системы Яндекс, а также в рамках Российского семинара по оценке методов информационного поиска (РОМИП, `http://romip.ru`). При переводе мы искали не просто понятные, но и по возможности однозначные и непротиворечивые варианты терминов. Поэтому, например, в книге везде, где только можно, *dictionary* (структура данных) переводится как словарь, а *vocabulary* — как лексикон, *proximity* — близость, а *similarity* — сходство. Также мы старались использовать устоявшуюся терминологию из других наук, например из биологии, в тех случаях, когда она существует (например, *capture-recapture* — метод повторного захвата). Многие члены сообщества *ru_ir* в Живом Журнале откликнулись на

наш призыв поучаствовать в коллективном переводе предметного указателя книги на сайте `translated.by`, за что мы им очень благодарны.

Мы не стали переводить примеры на русский язык, для этого их пришлось бы слишком сильно переработать. Мы надеемся, что читатель сможет воспользоваться пояснениями в тексте, да и точного понимания текста примеров для понимания работы алгоритмов и методов, как правило, не требуется.

Наконец, мы взяли на себя смелость снабдить текст комментариями не только для указания эквивалентных русских источников, если они имеются, и комментариев по выбору терминологии, но и в случаях, когда наш опыт разработчиков веб-поиска и исследователей позволял дополнить оригинальный текст (в некоторых случаях это происходило из-за специфики русскоязычного поиска).

Мы хотим поблагодарить тех, кто принял участие в переводе терминологии и прислал свои предложения и замечания по тексту: Андрей Белов, Леонид Бойцов, Константин Воронцов, Максим Захаров, Юрий Зеленков, Евгений Кирпичев, Константин Коломеец, Наташа Лауфер, Лидия Пивоварова, Денис Расковалов, Григорий Сапунов, Александр Сигачёв, Павел Уваров, Евгений Харитонов.

Мы рады отметить, что интерес к информационному поиску в России растет. В качестве “точек кристаллизации” такого интереса можно назвать семинар РОМИП, серию летних школ RuSSIR (`http://romip.ru/russir2010/`), “Школу анализа данных” Яндекса (`http://shad.yandex.ru/`), сообщество “Информационный поиск” (`http://community.livejournal.com/ru_ir/`). Вопросы информационного поиска обсуждаются на конференциях “Электронные библиотеки” (`http://rcdl.ru/`) и “Диалог” (`http://www.dialog-21.ru/`). Надеемся, что эта книга поможет развитию информационного поиска — научной и прикладной дисциплины — в России и ближнем зарубежье.

Мы благодарим ООО “Яндекс” за поддержку русского издания книги.

Павел Браславский
Дмитрий Клюшин
Илья Сегалович

Предисловие

Еще в 1990-х годах результаты социологических исследований свидетельствовали о том, что большинство людей предпочитают получать информацию от других людей, а не с помощью информационно-поисковых (Information Retrieval — IR) систем. Например, в то время для бронирования билетов и гостиниц люди чаще обращались к сотрудникам туристических агентств. Однако за последние десять лет благодаря постоянному совершенствованию методов информационного поиска поисковые системы в вебе поднялись на новый качественный уровень, позволяющий лучше удовлетворять потребности все большего количества людей, а веб-поиск стал стандартным и часто предпочтительным механизмом поиска информации. Например, в 2004 году опрос Pew Internet Survey (Fallows, 2004) показал, что "92% пользователей сети Интернет считают ее удобной для получения повседневной информации". К удивлению многих, информационный поиск из преимущественно академической дисциплины стал базисом для средств доступа к информации, на который полагается большинство людей. В книге изложены научные основы этой дисциплины на уровне, доступном как студентам старших курсов университетов, так и способным студентам младших курсов.

Информационный поиск возник раньше веба. Его эволюция стимулировалась разнообразными проблемами, связанными с обеспечением поиска и доступа к информационным источникам. Сначала информационный поиск касался научных публикаций и библиотечных каталогов, однако вскоре он распространился и на другие сферы, в которых важна роль информации, — на журналистику, право и медицину. Многие исследования в области информационного поиска проводились именно в этом контексте, и до сих пор большая доля практических приложений этой дисциплины связана с обеспечением доступа к неструктурированной информации, хранящейся в многочисленных корпоративных и правительственных базах данных. Именно этим методам посвящена б*о*льшая часть книги.

Тем не менее в последние годы основным двигателем прогресса является веб, открывший возможность публиковать информацию десяткам миллионов пользователей. Эта лавина публикаций осталась бы недоступной, если бы информацию было невозможно найти, сопроводить аннотацией и проанализировать так, чтобы каждый пользователь мог быстро найти необходимые ему релевантные и исчерпывающие сведения. В конце 1990-х годов многие люди поняли, что дальнейшая индексация всего веба вскоре станет невозможной из-за его экспоненциального роста. Однако значительные научные инновации и превосходные инженерные решения, быстро снижающаяся стоимость компьютерного аппаратного обеспечения и появление коммерческой заинтересованности в веб-поиске в совокупности способствовали возникновению крупных поисковых систем, способных с высоким качеством и за доли секунды выполнить сотни миллионов запросов в день по базе, состоящей из миллиардов веб-страниц.

Структура книги и учебного курса

Книга является результатом объединения нескольких учебных курсов, прочитанных в Стэнфордском университете (Stanford University) и Штутгартском университете (University of Stuttgart) в разных вариантах: на протяжении одной четверти, одного семестра и двух четвертей. Эти курсы предназначались для старшекурсников, изучавших компьютерные науки, но оказались полезными и для студентов младших курсов, а также для студентов, осваивавших юриспруденцию, медицинскую информатику, статистику, лингвистику и разнообразные технические дисциплины. Книга организована так, чтобы осветить то, что мы считаем важным для студентов, изучающих информационный поиск на протяжении одного семестра. Кроме того, каждая глава содержит материал одной лекции продолжительностью 75–90 минут.

Главы 1–8 посвящены основам информационного поиска и, в частности, сущности поисковых систем; мы считаем, что этот материал является ядром любого курса по информационному поиску. В главе 1 введены инвертированные индексы (inverted indexes) и показано, как с их помощью можно обработать простые булевы запросы (Boolean queries). В главе 2 детально описываются способы предварительной обработки документов перед индексированием и методы усовершенствования индексов для расширения функциональных возможностей и повышения скорости поиска. В главе 3 рассматриваются поисковые структуры для словарей и методы обработки запросов, содержащих орфографические ошибки и другие неточности. В главе 4 описывается несколько алгоритмов построения инвертированного индекса по коллекции текстов с особым акцентом на масштабируемые и распределенные алгоритмы, допускающие применение к очень большим коллекциям. В главе 5 излагаются методы сжатия словарей и инвертированных индексов. Эти методы очень важны для обеспечения быстрой (за доли секунды) обработки пользовательских запросов в больших поисковых системах. Индексы и запросы, изучаемые в главах 1–5, касаются лишь *булева поиска* (Boolean retrieval), при котором документ либо соответствует запросу, либо нет. Желание измерить *степень* соответствия документа запросу, или релевантность (score) документа, стимулировало разработку методов взвешивания терминов (term weighting) и ранжирования (computation of scores), описанных в главах 6 и 7, и далее, к концепции списка документов, упорядоченных по степени соответствия запросу. Глава 8 посвящена оценке информационно-поисковых систем на основании экспертных оценок релевантности найденных документов, что позволяет сравнивать относительное качество систем на стандартных коллекциях документов и запросов.

Главы 9–21 основаны на материале, изложенном в главах 1–8, и охватывают широкий спектр более сложных тем. В главе 9 обсуждаются методы повышения эффективности поиска с помощью таких приемов, как обратная связь по релевантности (relevance feedback) и расширение запросов (query expansion), предназначенных для увеличения вероятности нахождения релевантных документов. В главе 10 рассматриваются методы информационного поиска по документам, структурированным с помощью языков разметки, таких как XML и HTML. Мы сводим поиск по структурированным документам к применению методов ранжирования на основе векторной модели (vector space scoring), изложенных в главе 6. В главах 11 и 12 для ранжирования документа по отношению к запросу используется теория вероятностей. Глава 11 посвящена традиционному вероятностному информационному поиску, позволяющему вычислить вероятность релевантности документа при заданном наборе слов запроса. Впоследствии эту вероятность можно использовать как показатель релевантности при ранжировании. В главе 12 иллюстриру-

ется альтернатива, в рамках которой для каждого документа в коллекции создается языковая модель, позволяющая оценить вероятность того, что она порождает заданный запрос. Эта вероятность является еще одним количественным показателем, с помощью которого осуществляется ранжирование документов.

В главах 13–18 излагаются методы машинного обучения и численные методы информационного поиска. Главы 13–15 посвящены проблеме классификации документов по известным категориям на основе набора документов и классов, которым они принадлежат. В главе 13 представлены доказательства того, что классификация на основе статистики представляет собой одну из ключевых технологий, необходимых для успешного функционирования поисковой системы. В ней излагается наивный байесовский подход (Naive Bayes), представляющий собой концептуально простой и эффективный метод классификации текстов, а также основы стандартной методологии оценки текстовых классификаторов. В главе 14 описано применение модели векторного пространства, введенной в главе 6, а также изложены два метода классификации: метод Роккио (Rocchio method) и метод k ближайших соседей (k nearest neighbor — kNN), применяемые к векторам документов. В ней также рассматривается компромисс между смещением и разбросом (дисперсией), представляющий собой важную характеристику задач обучения и позволяющий установить критерии для выбора подходящего метода классификации текстов. В главе 15 вводится метод опорных векторов (support vector machine), который многие исследователи в настоящее время считают наиболее эффективным методом классификации текстов. Кроме того, в данной главе исследуются связи между задачей классификации и, на первый взгляд, совершенно посторонними темами, таким как вывод функций ранжирования по набору обучающих примеров.

Главы 16–18 посвящены идентификации кластеров близких документов в коллекции. В главе 16 сначала приводится обзор нескольких важных приложений кластеризации в области информационного поиска, а затем рассматриваются два алгоритма плоской кластеризации (flat clastering): эффективный и широко используемый для кластеризации документов алгоритм *K* средних (*K*-means algorithm) и EM-алгоритм (expectation-maximization algorithm), который с вычислительной точки зрения является более затратным, но более гибким. В главе 17 обосновывается необходимость иерархически структурированной кластеризации (вместо плоской) для многих приложений в области информационного поиска, а также рассматриваются алгоритмы кластеризации, порождающие иерархии кластеров. В этой главе также рассматривается сложная проблема автоматической разметки кластеров. Глава 18 посвящена методам линейной алгебры, представляющим собой расширение методов кластеризации и открывающим захватывающие перспективы для применения алгебраических методов, разрабатываемых в рамках латентного семантического индексирования (latent semantic indexing).

Главы 19–21 посвящены проблемам поиска в вебе. В главе 19 приводятся краткий обзор основных задач, связанных с поиском в вебе, а также набор широко распространенных методов информационного поиска в вебе. В главе 20 описываются архитектура и требования, предъявляемые к веб-роботам (web-crawlers). В главе 21 рассматривается применение анализа ссылок для веб-поиска, где анализ проводится с использованием методов линейной алгебры и теории вероятностей.

Эта книга является исчерпывающим источником знаний по всем темам, связанным с информационным поиском. За ее пределами осталось множество тем, выходящих за рамки вводного курса по информационному поиску. Тем не менее все, кого интересуют эти темы, могут обратиться к перечисленным ниже учебникам.

Cross-language IR, Grossman and Frieder, 2004, ch. 4, and Oard and Dorr, 1996.

Image and multimedia IR, Grossman and Frieder, 2004, ch. 4; Baeza-Yates and Ribeiro-Neto, 1999, ch. 6; Baeza-Yates and Ribeiro-Neto, 1999, ch. 11; Baeza-Yates and Ribeiro-Neto, 1999, ch. 12; del Bimbo, 1999; Lew, 2001; and Smeulders et al., 2000.

Speech retrieval, Coden et al., 2002.

Music retrieval, Downie, 2006 and `http://www.ismir.net/`.

User interfaces for IR, Baeza-Yates and Ribeiro-Neto, 1999, ch. 10.[1]

Search User Interfaces, Marti A. Hearst, 2009.

Parallel and peer-to-peer IR, Grossman and Frieder, 2004, ch. 7; Baeza-Yates and Ribeiro-Neto, 1999, ch. 9; and Aberer, 2001.

Digital libraries, Baeza-Yates and Ribeiro-Neto, 1999, ch. 15, and Lesk, 2004.

Information science perspective, Korfhage, 1997; Meadow et al., 1999; and Ingwersen and Järvelin, 2005.

Logic-based approaches to IR, van Rijsbergen, 1989.

Natural language processing techniques, Manning and Schütze, 1999; Jurafsky and Martin, 2008; and Lewis and Jones, 1996.

Предварительные требования к уровню подготовки читателей

Чтобы понять содержание книги, достаточно знать основы структур данных и алгоритмов, линейной алгебры и теории вероятностей. Для удобства читателей и преподавателей, желающих сосредоточиться лишь на отдельных главах, мы приводим следующую информацию.

Для освоения глав 1–5 необходимо знать основы структур данных и алгоритмов. Главы 6 и 7 требуют в дополнение к этому знания основ линейной алгебры, включая векторы и скалярное произведение. Далее вплоть до главы 11 никаких дополнительных знаний не требуется. Для освоения материала, изложенного в главе 11, следует разбираться в основах теории вероятностей; краткий обзор понятий, используемых в главах 11–13, приведен в разделе 11.1. Глава 15 требует от читателя знания понятий нелинейной оптимизации, хотя ее можно читать, не имея глубоких познаний об алгоритмах нелинейной оптимизации. Глава 18 предполагает знание основ линейной алгебры, включая ранг матрицы и собственные векторы; краткий обзор этих понятий приведен в разделе 18.1. Знание определений собственных значений и собственных векторов требуется также при чтении главы 21.

Разметка

Примеры в тексте сопровождаются пиктограммой, на которой изображен карандаш.

Сложный материал выделяется пиктограммой с изображением ножниц.

?

Упражнения отмечены знаком вопроса. Сложность задачи указывается с помощью звездочек: простая — одна [*], средняя — две [**] и сложная — три [***] звездочки.

[1] Можно также посоветовать недавнюю книгу Marti A. Hearst, Search User Interfaces, 2009, см. http://searchuserinterfaces.com/. – *Примеч. ред.*

Благодарности

Мы благодарим издательство Cambridge University Press за то, что оно позволило опубликовать рабочий вариант книги в Сети, благодаря чему мы получили обратную связь с потенциальными читателями уже в процессе написания книги. Мы также выражаем признательность Лорен Коулз (Lauren Cowles), превосходному редактору, многократно изучившему рукопись каждой главы и сделавшему множество полезных замечаний относительно стиля, организации и охвата материала, а также ряд глубоких замечаний по существу книги. В том, что мы достигли поставленной цели, — большая ее заслуга.

Мы очень благодарны многим людям, сделавшим комментарии, предложения и исправления, основываясь на черновом варианте книги. Вот их имена: Шерил Аашейм (Cheryl Aasheim), Джош Аттенберг (Josh Attenberg), Бъерн Андрист (Björn Andrist), Люк Беланжер (Luc Bélanger), Том Бройель (Tom Breuel), Даниэль Буркхардт (Daniel Burckhardt), Георг Бушер (Georg Buscher), Фазли Кан (Fazli Can), Динцюань Чен (Dingquan Chen), Эрнест Дэвис (Ernest Davis), Педро Домингос (Pedro Domingos), Родриго Панчиниак Фернандес (Rodrigo Panchiniak Fernandes), Паоло Ферраджина (Paolo Ferragina), Норберт Фюр (Norbert Fuhr), Вигнеш Ганапати (Vignesh Ganapathy), Элмер Гардуно (Elmer Garduno), Сюбо Гэн (Xiubo Geng), Дэвид Гондек (David Gondek), Серджио Говони (Sergio Govoni), Миклош Эрдели (Miklós Erdélyi), Коринна Хабетц (Corinna Habets), Бен Хэнди (Ben Handy), Донна Харман (Donna Harman), Бенджамин Хаскелл (Benjamin Haskell), Томас Хюнн (Thomas Hühn), Дипак Джейн (Deepak Jain), Ральф Янкович (Ralf Jankowitsch), Динакар Джаяраджан (Dinakar Jayarajan), Винай Какаде (Vinay Kakade), Марек Ковалькевич (Marek Kowalkiewicz), Мэй Кобаяси (Mei Kobayashi), Вессель Краий (Wessel Kraaij), Рик Лефлер (Rick Lafleur), Флориан Лоус (Florian Laws), Хан Ли (Hang Li), Дэвид Манн (David Mann), Эннио Мази (Ennio Masi), Джуна Макконен (Juna Makkonen), Фрэнк Маккоун (Frank McCown), Пол Макнами (Paul McNamee), Свен Мейер цу Эссен (Sven Meyer zu Eissen), Александер Мурзаку (Alexander Murzaku), Гонзало Наварро (Gonzalo Navarro), Скотт Олссон (Scott Olsson), Даниэль Паива (Daniel Paiva), Тао Цинь (Tao Qin), Мега Рагхаван (Megha Raghavan), Картик Рагунатан (Karthik Raghunathan), Гулям Раза (Ghulam Raza), Михал Розен-Цви (Michal Rosen-Zvi), Клаус Ротенхауслер (Klaus Rothenhäusler), Кенью Л. Раннер (Kenyu L. Runner), Александер Саламанка (Alexander Salamanca), Григорий Сапунов (Grigory Sapunov), Тобиас Шеффер (Tobias Scheffer), Нико Шлафер (Nico Schlaefer), Евгений Шадчнев (Evgeny Shadchnev), Ян Соборофф (Ian Soboroff), Бенно Штейн (Benno Stein), Марцин Сидоу (Marcin Sydow), Эндрю Тёрнер (Andrew Turner), Джейсон Утт (Jason Utt), Хьюи Во (Huey Vo), Трэвис Уэйд (Travis Wade), Майк Уолш (Mike Walsh), Чанлян Ван (Changliang Wang), Жэньцзин Ван (Renjing Wang), Делл Джанг (Dell Zang) и Томас Цойме (Thomas Zeume).

Многие люди присылали свои замечания к отдельным главам как по нашей просьбе, так и по своей инициативе. Мы очень благодарны за это Джеймсу Аллану (James Allan), Омару Алонзо (Omar Alonso), Исмаилу Сеньору Альтинговде (Ismail Sengor Altingovde), Во Нгок Ану (Vo Ngoc Anh), Ру Бланко (Roi Blanco), Эрику Бреку (Eric Breck), Эрику Брауну (Eric Brown), Марку Карману (Mark Carman), Карлосу Кастильо (Carlos Castillo), Юнгху Чо (Junghoo Cho), Арону Кулотте (Aron Culotta), Дугу Каттингу (Doug Cutting), Мегане Деодхар (Meghana Deodhar), Сьюзан Дюмэ (Susan Dumais), Йоханнесу Фюрнк-

ранцу (Johannes Fürnkranz), Андреасу Хессу (Andreas Heß), Дьорду Хиемстра (Djoerd Hiemstra), Дэвиду Халлу (David Hull), Торстену Йоахимсу (Thorsten Joachims), Сиддхартхе Джонатану Дж. Б. (Siddharth Jonathan J. B.), Йаапу Кампсу (Jaap Kamps), Мунье Лалмас (Mounia Lalmas), Эми Лэнгвиль (Amy Langville), Николасу Лестеру (Nicholas Lester), Дэйву Льюису (Dave Lewis), Стивену Лиу (Stephen Liu), Даниэлю Лоуду (Daniel Lowd), Йоси Массу (Yosi Mass), Джеффу Мишелзу (Jeff Michels), Алессандро Москитти (Alessandro Moschitti), Амиру Найми (Amir Najmi), Марку Найорку (Marc Najork), Джорджио Марие Ди Нунцио (Giorgio Maria Di Nunzio), Полю Огилви (Paul Ogilvie), Приянке Патель (Priyank Patel), Яну Педерсену (Jan Pedersen), Кэтрин Педингс (Kathryn Pedings), Виссилису Плахурасу (Vassilis Plachouras), Даниэлю Рамаге (Daniel Ramage), Стивену Ризлеру (Stefan Riezler), Майклу Шиелену (Michael Schiehlen), Хельмуту Шмиду (Helmut Schmid), Фальку Николасу Шолеру (Falk Nicolas Scholer), Сабине Шульте им Вальде (Sabine Schulte im Walde), Фабрицио Себастиани (Fabrizio Sebastiani), Сарабжит Сингху (Sarabjeet Singh), Валентину Спитковскому (Valentin I. Spitkovsky), Александру Штрелю (Alexander Strehl), Джону Тейту (John Tait), Шивакумару Вайтианатану (Shivakumar Vaithyanathan), Эллен Ворхиз (Ellen Voorhees), Герхарду Вайкуму (Gerhard Weikum), Дэвиду Вайсу (Dawid Weiss), Джимингу Янгу (Yiming Yang), Йисонгу Ю (Yisong Yue), Джиану Чангу (Jian Zhang) и Джастин Цобель (Justin Zobel).

Кроме того, мы хотели бы упомянуть рецензентов, количество и качество комментариев которых трудно переоценить. Мы благодарны Павлу Берхину (Pavel Berkhin), Стефану Бютчеру (Stefan Büttcher), Джейми Каллан (Jamie Callan), Байрону Дому (Byron Dom), Торстену Зулю (Torsten Suel) и Эндрю Тротману (Andrew Trotman) за их значительное влияние на содержание и структуру книги.

Исходные черновики глав 13–15 были основаны на слайдах, созданных Рэем Муни (Ray Mooney). Несмотря на то что этот материал подвергся существенной переработке, мы благодарны Рэю за его вклад в эти три главы в целом и за описание вопросов, связанных с временно*й* сложностью алгоритмов классификации текстов, в частности.

К сожалению, мы не в состоянии перечислить всех, поскольку до сих пор получаем отзывы от наших читателей. Как и все самоуверенные авторы, мы не всегда прислушивались к их советам. Опубликованный вариант книги полностью остается на совести авторов.

Авторы выражают благодарность Стэнфордскому университету (Stanford University) и Штутгардскому университету (University of Stuttgart) за создание академической среды для плодотворных научных дискуссий и возможность читать учебные курсы, ставшие основанием для этой книги. Кристофер Маннинг (Christopher Manning) благодарит свою семью за многие часы, которые она позволила ему провести, работая на книгой, и надеется, что ему удастся наверстать упущенное на уик-эндах в следующем году. Прабхакар Рагаван (Prabhakar Raghavan) благодарит свою семью за терпение и поддержку во время работы над книгой, а также выражает признательность компании Yahoo! Inc. за плодотворную атмосферу, в которой проходила эта работа. Хайнрих Шютце (Hinrich Schütze) хотел бы поблагодарить своих родителей, семью и друзей за поддержку в процессе работы над книгой.

Веб-адреса и контактная информация

Этой книге посвящен веб-сайт `http://informationretrieval.org`, который, помимо ссылок на различные ресурсы, содержит слайды для каждой главы, которые можно использовать для лекций. Мы приветствуем обратную связь с читателями и с благодарностью примем их замечания и предложения по улучшению книги, которые можно отсылать авторам по адресу `informationretrieval@yahoogroups.com`.