まえがき

調査によれば 1990 年代の最近まで，ほとんどの人が情報検索システム (information retrieval (IR) systems) よりも他の人たちから情報を手にいれる方を好んでいた．もちろんその時代では，旅行予約も，ほとんどの人は、人手による旅行斡旋を利用していた．しかし，この 10 年来の情報検索の効果に対する絶え間のない最適化の進歩はウエブのサーチエンジンの品質をほとんどの人がほとんどの場合に満足する新しいレベルまで押し上げ，ウエブサーチが，情報発見の標準になり，しばしば，選択肢としても好まれるようになった．2004 年の Pew Internet Survey (Fallows 2004) によると“インターネットユーザーの 92%が，日々の情報を手にいれるのにインターネットが良い場所であるといっている”そうである．多くの人にとって驚きだが，情報検索分野は第一義的な学問の一分野から，ほとんどの人にとっての情報アクセスの手段の基礎になった．この本は当分野の科学的な基礎を，大学院生と理解の進んだ学部生にとって理解できるレベルで，紹介する．

情報検索はウエブの世界とともに始まったわけではない．情報アクセスを提供の際のさまざまな課題に答えるために，IR の分野はさまざまな形式のコンテンツをサーチする原則的な（？？？？？ principled）アプローチを与え進化してきた．この分野は，科学的出版や図書館の記録から始まった．そしてすぐに他の形式のコンテンツ，とくに，ジャーナリストや弁護士や医者のような情報の専門家のコンテンツまで広がった．IR に関する科学的研究の多くはこれらの文脈で行われており，途切れなく行われてきた IR の数多くの実践によって，さまざまな企業や政府の領域での非定型の情報へのアクセスが提供されつづけてきた．さらに，これらの仕事が本書の多くの基礎を形作っている．

それにもかかわらず，近年ウエブ (the World Wide Web) は，数千万規模のコンテンツ作成者に出版を解き放ち，イノベーションの主な推進力でありつづけている．各ユーザーが，必要性に関連した包括的な情報をすばやく見つけられるように，情報を見つけたり，注釈を施したり，分析されなければ，この出版された情報の爆発は，現実的な価値をもたないだろう．1990 年代の後半には，ウエブの世界全体をインデックス化し続けるのは，ウエブの世界のサイズの指数関数的な成長のために，急速に不可能になるだろうと多くの人が感じていた．しかし，重要な科学的なイノベーションや優れたエンジニアリングやコンピューターハードウエアの価格の急速な減少や，ウエブ検索の商業的な基盤

の登場などすべてが一緒になって今日のウエブ検索に力を与え，数十億ものウエブページに対して一日数億回もの検索を秒以下で高品質の結果を提供する事ができるようになった．

本の構成と講義の開発

この本は私たちがスタンフォード大学 (Stanford University) やシュテュットガルト大学 (the University of Stuttgart) で，1 学期間，1 セメスター，2 学期間という範囲で教えた一連の講義の結果である．これらの講義はコンピューターサイエンスの大学院の初期の学生を対象としたが，学部のコンピューターサイエンスの高学年の学生や，法律，医療情報，統計，言語学，さまざまなエンジニアリングの分野からの学生もまた参加していた．したがって，この本の重要な設計原則は，私たちが IR について 1 学期間の大学院の講義で重要と信じるものを扱うことである．追加の原則は，75 分から 90 分の 1 講義のなかで扱えると信じる教材に基づいて各章を組み立てた．

この本の最初の 8 章は情報検索の基礎，とくにサーチエンジンの真髄にあてた．つまり，ここでの教材が情報検索のどの講義でも核であると考える．第 1 章は逆インデックスを紹介し，単純なブールクエリーがそのようなインデックスを使って処理できることを紹介する．第 2 章はこの紹介をもとに，文書がインデックス化される前の前処理の方法を詳細に扱い，逆インデックスがいろいろな方法で機能的にもスピードも補強できることを議論する．第 3 章は，辞書のための検索の構造を議論し，綴り訂正や，サーチされる文書コレクションの語彙と不正確にしか適合しない他のクエリーをどのように処理するかを議論する．第 4 章は，テキストのコレクションから逆インデックスを構築する多くのアルゴリズムを記述する．とくに，非常に大きなコレクションに適用できる高度にスケーラブルな分散アルゴリズムに格別に注意をはらう．第 5 章は辞書と逆インデックスを圧縮する技法を扱う．これらの技法は大規模なサーチエンジンで，ユーザーのクエリーに秒単位以下で応答するために必須である．第 1 章から 5 章までで考えられたインデックスとクエリーは**ブール検索** (*Boolean retrieval*) のみを扱っている．そこでは文書はクエリーに一致するか否かのどちらかである．文書がクエリーに一致する**程度** (*extent*) や，あるいは，クエリーに対する文書の得点付けを計測したいので，第 6 章および第 7 章での用語の重み付けの開発や得点付けの計算の発想が産まれ，これがクエリーに対するランキングされた文書のリストという着想につながった．第 8 章は，情報検索が取得した文書の関連度に基づいた情報検索システムの評価に焦点をあてている．これによって，ベンチマークとなる文書コレクションやクエリーについて，異なるシステムの相対的な性能の比較ができる．

第 9 章から 21 章は最初の 8 章の基礎をもとに，より先進的なさまざまな話題を扱っている．第 9 章は，関連する文書を取り出す確率を増やすことを目的にした関連性のフィードバック (relevance feedback) やクエリー拡張 (query expansion) のような技法を用いて検索が強化できる手法を議論している．第 10 章は XML や HTML のようなマークアップ言語によって構造化された文書か

らの情報検索 (IR) を扱っている．本書では，構造化検索を第 6 章で開発されるベクトル空間得点付け手法 (the vector space scoring methods) に還元して扱っている．第 11 章および第 12 章は，クエリーに対する文書の得点付けを計算するのに確率論を使っている．第 11 章は，クエリー用語集合が与えられたとき，文書の関連性の確率を計算するためのフレームワークを提供する伝統的な確率的 IR を開発している．そして，この確率はランキングにおける得点付けとして使われるだろう．第 12 章は，別の選択肢を描き，そこでは，コレクション中の各文書に対して，言語モデルを構築し，そのモデルから与えられたクエリーを生成する確率を評価できる． この確率は，文書をランク順にできるもう一つの数量である．

第 13 章から 18 章は，情報検索でのさまざまな形の機械学習と数値手法を取り扱う．第 13 章から 15 章は，文書集合とそれらが属すクラスが与えられたとき，知られているカテゴリー集合に文書を分類する問題を扱う．第 13 章は，検索エンジンを成功させる主な技術の一つとして，統計的分類の動機付けを行う．概念的に単純で効率の良いテキストの分類手法である単純ベイズ分類器 (Naive Bayes) を紹介する．そして，テキスト分類器を評価する標準的な手法を概観する．第 14 章は，第 6 章のベクトル空間モデルを扱い，文書のベクトルを操作する二つの分類手法，ロッキオ (Rocchio) と k 近傍 (kNN，$k$nearest neighbor) を紹介する．また，テキストの分類問題に対して適切な手法の選択基準を提供する学習問題の重要な特徴づけとしてバイアスバリアンストレードオフ (bias-variance tradeoff) を紹介する．第 15 章は，多くの研究者が現在もっとも有効なテキスト分類手法と見ているサポートベクターマシン (support vector machine) を紹介する．この章で，分類問題と，訓練用の事例集合から得点付け関数を誘導するといった一見違った話題とのつながりも扱おう．

第 16，17，18 章はコレクションの関連文書からクラスターを導く問題を考える．第 16 章では，最初に IR におけるクラスター化の多くの重要な応用を概観する．そして，2 つのフラットクラスタリングアルゴリズムを記述する．それらは効率がよく広く使われている文書のクラスタリング手法である K-平均 (K-means) アルゴリズムと，計算量的にはより高価だが，より柔軟でもある期待値最大化 (EM) アルゴリズム (expectation-maximization algorithm) である．第 17 章は，（フラットクラスタリングの代わりに）IR の多くの応用分野で階層的構造化クラスタリング (hierarchically structured clusterings) の必要性を説き，クラスターの階層を産み出す多くのクラスタリングアルゴリズムを紹介している．この章はまた，クラスターに対して自動的にラベルを計算する難問も扱っている．第 18 章は，クラスタリングの拡張を行う線形代数からの手法を開発し，また IR における代数的手法の興味をそそる側面も扱っており．潜在的意味論インデキシングのアプローチの中でさらに追求する．

第 19 章から 21 章はウエブサーチの問題を扱っている．第 19 章でウエブサーチにおける基本的な課題の要約を，ウエブ情報検索において一般的な技法の集合も一緒に与えている．次に，第 20 章は基本的なウエブクローラーのアーキテクチャーと必要な要素を記述している．最後に，第 21 章は，線形代数と先進的な確率論からのいくつかの手法を用いたウエブサーチにおけるリンク解析の実力を考察している．

本書は IR に関連したすべての話題を扱うといった包括的なものではない．IR 入門クラスで取り扱いたい範囲を超えていると考えられる多くの話題ははずしている．にもかかわらず，これらの話題に興味のある読者のために，主に教科書の範疇で次のポインターを提供する．

多言語 IR (Cross-language IR) Grossman and Frieder 2004, ch. 4 と Oard and Dorr 1996.
画像とマルチメディアの 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; Smeulders et al. 2000.
音声検索 (Speech retrieval) Coden et al. 2002.
音楽検索 (Music retrieval) Downie 2006 と http://www.ismir.net/.
IR のためのユーザーインターフェース Baeza-Yates and Ribeiro-Neto 1999, ch. 10.
並列とピア・ツー・ピア (peer-to-peer IR) Grossman と Frieder 2004, ch. 7; Baeza-Yates と Ribeiro-Neto 1999, ch. 9; そして**?**.
デジタル図書館 Baeza-Yates and Ribeiro-Neto 1999, ch. 15 と Lesk 2004.
情報科学全体 (information science perspective) Korfhage 1997; Meadow et al. 1999; そして Ingwersen and Järvelin 2005.
IR への論理学的アプローチ van Rijsbergen 1989.
自然言語処理技法 Manning and Schütze 1999; Jurafsky and Martin 2008; そして，Lewis and Jones 1996.

前提条件

データ構造とアルゴリズムや線形代数や確率論それぞれの入門コースは，21 すべての章の前提条件として十分である．いくつかの章だけを読むように調整したい読者や教師のためにより詳細なものをここで触れよう．

第 1 章から 5 章は，アルゴリズムとデータ構造の基礎のコースを前提条件にしている．第 6 章および第 7 章は，さらに，ベクトルとドット積を含む基本線形代数の知識が必要である．確率論の基礎コースが必要だが，第 11 章まで他の追加の前提条件はない．11.1 節は第 11，12，13 章に必要な概念の手早い復習を行っている．第 15 章は，読者が非線形最適化の観念に通じていることを仮定しているが，この章は非線形最適化についてのアルゴリズムの詳細の知識がなくとも読むことができるだろう．第 18 章は，線形代数の最初のコースと行列の階数と固有値に慣れ親しんでいることが必要である．18.1 節で簡単な復習を与えている．固有値と固有ベクトルの知識は，第 21 章でも必要である．

本の構成

テキスト中の実行可能な例は，それらの左側の余白に鉛筆のしるしを示している．節や副節に現れる高度な，あるいは難しい題材は，余白に鋏の絵で示している．演習問題は余白に疑問符でしるしをつけている．演習問題の難易度は，簡単 [⋆]，中程度 [⋆⋆]，難題 [⋆⋆⋆] で示している．

謝辞

著者たちは，この本の草稿のオンラインでの開示を許可してくれたCambridge University Press に感謝する．それによって，この本の執筆中に多くのフィードバックを得ることができた．私たちはまた Lauren Cowles に感謝したい．彼女は素晴らしい編集者だった．数回にもわたって，スタイルに関してや構成や取り扱いの範囲について各章のコメントを，さらに，本の内容についても詳細のコメントを提供してくれた．この本の執筆という目標の達成にあたって，彼女は重要な功績を果たした．

この本の草稿にコメントや助言や修正を与えてくれた非常に多くの人たちにとても感謝している．いろいろな修正やコメントを与えてくれた次の人たちに感謝している．

Cheryl Aasheim, Josh Attenberg, Luc Bélanger, Tom Breuel, Daniel Burckhardt, Georg Buscher, Fazli Can, Dinquan Chen, Ernest Davis, Pedro Domingos, Rodrigo Panchiniak Fernandes, Paolo Ferragina, Norbert Fuhr, Vignesh Ganapathy, Elmer Garduno, Xiubo Geng, David Gondek, Sergio Govoni, Corinna Habets, Ben Handy, Donna Harman, Benjamin Haskell, Thomas Hühn, Deepak Jain, Ralf Jankowitsch, Dinakar Jayarajan, Vinay Kakade, Mei Kobayashi, Wessel Kraaij, Rick Lafleur, Florian Laws, Hang Li, David Mann, Ennio Masi, Frank McCown, Paul McNamee, Sven Meyer zu Eissen, Alexander Murzaku, Gonzalo Navarro, Scott Olsson, Daniel Paiva, Tao Qin, Megha Raghavan, Ghulam Raza, Michal Rosen-Zvi, Klaus Rothenhäusler, Kenyu L. Runner, Alexander Salamanca, Grigory Sapunov, Tobias Scheffer, Nico Schlaefer, Evgeny Shadchnev, Ian Soboroff, Benno Stein, Marcin Sydow, Andrew Turner, Jason Utt, Huey Vo, Travis Wade, Mike Walsh, Changliang Wang, Renjing Wang, and Thomas Zeume.

多くの人たちが，私たちの依頼か，彼等自身から，個別の章に詳細なフィードバックを与えてくれた．この点で，とくに次の人たちに感謝する．

James Allan, Omar Alonso, Ismail Sengor Altingovde, Vo Ngoc Anh, Roi Blanco, Eric Breck, Eric Brown, Mark Carman, Carlos Castillo, Junghoo Cho, Aron Culotta, Doug Cutting, Meghana Deodhar, Susan Dumais, Johannes Fürnkranz, Andreas Heß, Djoerd Hiemstra, David Hull, Thorsten Joachims, Siddharth Jonathan J. B., Jaap Kamps, Mounia Lalmas, Amy Langville, Nicholas Lester, Dave Lewis, Stephen Liu, Daniel Lowd, Yosi Mass, Jeff Michels, Alessandro Moschitti, Amir Najmi, Marc Najork, Giorgio Maria Di Nunzio,