# 目次

記号表 *page* x
まえがき xiii

*1 Boolean retrieval* 1

1.1 情報検索問題の例 3
1.2 逆インデックス構築の最初の試み 6
1.3 ブールクエリーの処理 10
1.4 拡張されたブールモデル対ランク付けされた検索 14
1.5 引用文献と参考図書 16

*2* 用語語彙とポスティングリスト 18

2.1 文書輪郭と文字列解読 18
2.2 用語の語彙を決定する 21
2.3 スキップポインターによる高速なポスティングリストの共通集合操作 32
2.4 位置的ポスティングと句クエリー 35
2.5 引用文献と参考図書 41

*3* 辞書と寛容な検索 43

3.1 辞書検索の構成 43
3.2 ワイルドカード検索 45
3.3 スペル修正 49
3.4 音声上の修正 56
3.5 引用文献と参考図書 57

*4* インデックスの構築 59

4.1 ハードウェア概説 60
4.2 ブロックソートインデキシング 61
4.3 単一パスインメモリインデキシング 64
4.4 分散インデキシング 66

4.5 動的インデキシング 69
4.6 他のタイプのインデックスについて 72
4.7 引用文献と参考図書 75

***5 Index compression*** 77

5.1 検索システムにおける用語の統計的性格 78
5.2 辞書の圧縮 82
5.3 ポスティングファイルの圧縮 86
5.4 引用文献と参考図書 96

**6 点数付け，語重みづけ，ベクトル空間モデル** 98

6.1 パラメーター及びゾーンインデックス 98
6.2 語頻度と重みづけ 104
6.3 点数付けのベクトル空間モデル 107
6.4 改良型 tf–idf 関数 113
6.5 引用文献と参考図書 118

**7 完全な検索システムでの計算スコア** 119

7.1 効果的なスコアリングとランキング 119
7.2 情報検索システムのコンポーネント 127
7.3 ベクトル空間スコアリングとクエリ演算子の相互作用 131
7.4 引用文献と参考図書 132

**8 情報検索の評価** 134

8.1 情報検索システムの評価 135
8.2 一般的に使われているテスト集合 136
8.3 順位なし検索集合の評価 137
8.4 順位つき検索結果の評価 140
8.5 妥当性の評価 146
8.6 さらに広い観点から：システムの質とユーザーの実用性 149
8.7 検索結果のスニペット 152
8.8 引用文献と参考図書 154

***9 Relevance feedback and query expansion*** 157

9.1 Relevance feedback and pseudo relevance feedback 158
9.2 Global methods for query reformulation 168
9.3 References and further reading 172

***10 XML* 検索** 173

10.1 基本的な XML 概念 175
10.2 XML 検索のチャレンジ 178
10.3 XML 検索のベクトル空間モデル 183
10.4 XML 検索の評価 187

10.5 テキスト中心対データ中心の XML 検索 191
10.6 引用文献と参考図書 193

***11*** **確率的情報検索** 195

11.1 基本的な確率論のレビュー 196
11.2 確率ランキング原理 197
11.3 バイナリ独立モデル 198
11.4 評価といくらかの拡張 205
11.5 引用文献と参考図書 209

***12*** **検索のための言語モデル** 211

12.1 言語モデル 211
12.2 クエリ尤度モデル 216
12.3 情報検索における言語モデル手法と他の手法との対比 222
12.4 拡張言語モデル手法 223
12.5 引用文献と参考図書 224

***13*** **テキストの分類とナイーブベイズ** 226

13.1 テキスト分類問題 229
13.2 ナイーブベイズテキスト分類 231
13.3 ベルヌーイモデル 235
13.4 ナイーブベイズの性質 237
13.5 特徴選択 243
13.6 テキスト分類の評価 250
13.7 引用文献と参考図書 257

***14*** **ベクトル空間分類** 259

14.1 文書の表現とベクトル空間での関係性の指標 260
14.2 ロッキオ分類 262
14.3 $k$ 最近傍法 266
14.4 線形分類器 対 非線形分類器 270
14.5 クラス数が 2 より多い分類 274
14.6 バイアス–バリアンス・トレードオフ 276
14.7 引用文献と参考図書 283

***15*** **支持ベクトル機械と文書の機械学習** 286

15.1 支持ベクトル機械 – 線形的に分離可能な場合 286
15.2 支持ベクトルモデルへの拡張 292
15.3 テキスト文書の分類における問題 299
15.4 アドホック情報検索での機械学習手法 304
15.5 引用文献と参考図書 308

***16*** **平坦クラスタ化** 311

16.1 情報検索でのクラスター化 312
16.2 問題記述 316
16.3 クラスター化の評価 317
16.4 $K$ 平均法 321
16.5 モデル基盤クラスター化 327
16.6 引用文献と参考図書 332

**17 階層的クラスター化** 335

17.1 階層的集塊性クラスター化 336
17.2 単一リンクと完全リンククラスター化 339
17.3 グループ平均集塊クラスター化 345
17.4 重心クラスター化 346
17.5 階層的集塊クラスター化の最適性 348
17.6 分割可能クラスター化 351
17.7 クラスターラベル付け 351
17.8 実装ノート 353
17.9 引用文献と参考図書 355

**18 行列の分解と潜在意味インデキシング** 357

18.1 線形代数の復習 357
18.2 用語—文書行列と特異値分解 361
18.3 低階数近似 363
18.4 潜在意味インデキシング 366
18.5 引用文献と参考図書 371

**19 ウエブ検索の基礎** 372

19.1 背景と歴史 372
19.2 ウエブの特徴 374
19.3 経済モデルとしての宣伝 380
19.4 サーチのユーザー体験 382
19.5 インデックスのサイズと推定 384
19.6 近複製とシングリング 388
19.7 引用文献と参考図書 392

**20 ウエブのクローリングとインデックス化** 393

20.1 概説 393
20.2 クローリング 394
20.3 インデックスを分散化する 403
20.4 接続サーバー 405
20.5 引用文献と参考図書 407

**21 リンク解析** 409

21.1 グラフとしてのウエブの世界 410

21.2 PageRank 412
21.3 ハブと権威者 421
21.4 引用文献と参考図書 427

参考文献 429
索引 456

# 記号表

| 記号 | ページ | 意味 |
|---|---|---|
| $\gamma$ | 89 | $\gamma$ コード |
| $\gamma$ | 229 | 分類，あるいは，クラスター関数: $\gamma(d)$ は $d$ のクラス，あるいは，クラスター |
| $\Gamma$ | 229 | 第 13 章および第 14 章の監督付き学習法（supervised learning method）: $\Gamma(\mathbb{D})$ は，トレーニング集合 $\mathbb{D}$ から学んだ分類関数 $\gamma$ である． |
| $\lambda$ | 358 | 固有値 (eigenvalue) |
| $\vec{\mu}(.)$ | 262 | （ロッキオ (Rocchio) 分類での）1 つのクラス，あるいは，（$K$-平均と重心クラスタリングでの）1 つのクラスターの重心 (centroid) |
| $\Phi$ | 103 | トレーニング例 (training example) |
| $\sigma$ | 361 | 特異値 (singular value) |
| $\Theta(\cdot)$ | 10 | アルゴリズムの計算複雑度のきっちりとした限界 (tight bound) |
| $\omega, \omega_k$ | 318 | クラスタリング中のクラスター |
| $\Omega$ | 318 | クラスタリング，あるいは，クラスターの集合 $\{\omega_1, \ldots, \omega_K\}$ |
| $\arg\max_x f(x)$ | 159 | $f$ が最大値に達するような $x$ の値 |
| $\arg\min_x f(x)$ | 159 | $f$ が最小値に達するような $x$ の値 |
| $c, c_j$ | 229 | 分類におけるクラス，あるいは，カテゴリー |
| $\mathrm{cf}_t$ | 81 | 用語 $t$ のコレクション中の頻度（その用語が，文書コレクション中に現れる総数） |
| $\mathbb{C}$ | 229 | すべてのクラスの集合 $\{c_1, \ldots, c_J\}$ |
| $C$ | 240 | $\mathbb{C}$ の要素を値とするランダム変数 |
| $C$ | 357 | 用語–文書行列 (term–document matrix) |
| $d$ | 4 | コレクション $D$ の $d$ 番目の文書のインデックス |
| $d$ | 63 | 文書 (document) |
| $\vec{d}, \vec{q}$ | 158 | 文書ベクトル，クエリーベクトル |
| $D$ | 316 | すべての文書からなる集合 $\{d_1, \ldots, d_N\}$ |
| $D_c$ | 262 | クラス $c$ にある文書集合 |

| | | |
|---|---|---|
| $\mathbb{D}$ | 229 | 第 13 章から第 15 章のすべてのラベル付き文書集合 $\{\langle d_1, c_1\rangle, \ldots, \langle d_N, c_N\rangle\}$ |
| $\mathrm{df}_t$ | 105 | 用語 $t$ の文書頻度（コレクション中でその用語が現れる文書の全数） |
| $H$ | 90 | エントロピー (entropy) |
| $H_M$ | 92 | $M$ 番目の調和数 ($M$th harmonic number) |
| $I(X;Y)$ | 244 | ランダム変数 $X$ と $Y$ の相互情報 (mutual information) |
| $\mathrm{idf}_t$ | 106 | 用語 $t$ の逆文書頻度 (inverse document frequency) |
| $J$ | 229 | クラスの数 |
| $k$ | 260 | 集合からの上位 $k$ 個の要素．例えば，kNN での $k$ 個の最近傍，上位 $k$ 個の検索された文書，語彙 $V$ からの上位 $k$ 個の選択された特徴 |
| $k$ | 48 | $k$ 個の文字列 |
| $K$ | 316 | クラスターの数 |
| $L_d$ | 207 | 文書 $d$ の長さ（トークン数でみたとき） |
| $L_\mathrm{a}$ | 234 | トークン数でみたときのテスト文書（あるいはアプリケーションの文書）の長さ |
| $L_\mathrm{ave}$ | 62 | トークン数でみたときの文書の平均長 |
| $M$ | 4 | 語彙のサイズ ($\lvert V\rvert$) |
| $M_\mathrm{a}$ | 234 | テスト文書（または，アプリケーション文書）の語彙のサイズ |
| $M_\mathrm{ave}$ | 70 | コレクション中の文書での語彙の平均サイズ |
| $M_d$ | 211 | 文書 $d$ の言語モデル |
| $N$ | 4 | 検索，あるいは，トレーニング用のコレクション中の文書数 |
| $N_c$ | 232 | クラス $c$ 中の文書数 |
| $N(\omega)$ | 268 | イベント $\omega$ の発生数 |
| $O(\cdot)$ | 10 | アルゴリズムの計算複雑度の上界 (rupper bound) |
| $O(\cdot)$ | 197 | イベントの発生確率 |
| $P$ | 138 | 精度 (precision) |
| $P(\cdot)$ | 196 | 確率 (probability) |
| $P$ | 413 | 遷移確率行列 (transition probability matrix) |
| $q$ | 51 | クエリー |
| $R$ | 138 | 再現率 (recall) |
| $s_i$ | 51 | 文字列 (string) |
| $s_i$ | 100 | ゾーン得点に対するブール値 (boolean values for zone scoring)？？？？？？？？？？？？ |
| $\mathrm{sim}(d_1, d_2)$ | 108 | 文書 $d_1, d_2$ の類似度得点 |
| $T$ | 38 | 文書コレクション中の全トークン数 |
| $T_{ct}$ | 232 | クラス $c$ の文書の中の語 $t$ の発生数 |
| $t$ | 4 | 語彙 $V$ のなかの $t$ 番目の用語のインデックス |
| $t$ | 54 | 語彙の中の用語 |
| $\mathrm{tf}_{t,d}$ | 105 | 文書 $d$ での用語 $t$ の用語頻度（$d$ での $t$ の全発生数） |

| | | |
|---|---|---|
| $U_t$ | 238 | （用語 $t$ があれば）値 0,（$t$ がなければ）値 1 を取るランダム変数 |
| $V$ | 185 | コレクション (a.k.a. レキシコン) 中の用語 $\{t_1, \ldots, t_M\}$ の語彙 |
| $\vec{v}(d)$ | 108 | 長さで正規化された文書ベクトル |
| $\vec{V}(d)$ | 108 | 長さで正規化されていない文書 $d$ のベクトル |
| $\mathrm{wf}_{t,d}$ | 112 | 文書 $d$ での用語 $t$ の重さ |
| $w$ | 100 | 重さ（例えば，ゾーンや用語の重さ） |
| $\vec{w}^{\mathrm{T}}\vec{x} = b$ | 262 | 超平面 (hyperplane)；$\vec{w}$ はこの長平面の法線ベクトル (normal vector) で $w_i$ は $\vec{w}$ の要素 $i$ |
| $\vec{x}$ | 198 | 用語接続ベクトル (term incidence vector) $\vec{x} = (x_1, \ldots, x_M)$；もっと一般的には，文書特徴表現 (document feature representation) |
| $X$ | 238 | 語彙 $V$ の値を取るランダム変数（例えば，文書中与えられた $k$ の位置で） |
| $\mathbb{X}$ | 229 | テキスト分類での文書空間 |
| $\lvert x \rvert$ | 123 | $x$ の絶対値（absolute value） |
| $\lvert A \rvert$ | 54 | 集合の要素数 (set cardinality)：集合 $A$ の要素の数 |
| $\lvert S \rvert$ | 358 | 正方行列 $S$ の行列式 (determinant) |
| $\lvert s_i \rvert$ | 51 | 文字列 $s_i$ の文字数 |
| $\lvert \vec{x} \rvert$ | 123 | ベクトル $\vec{x}$ の長さ |
| $\lvert \vec{x} - \vec{y} \rvert$ | 116 | $\vec{x}$ と $\vec{y}$ のユークリッド距離 (Euclidean distance)（つまり，$(\vec{x} - \vec{y})$ の長さ） |