



# 情報アクセス論 第2回

## 「情報検索システムの構成」

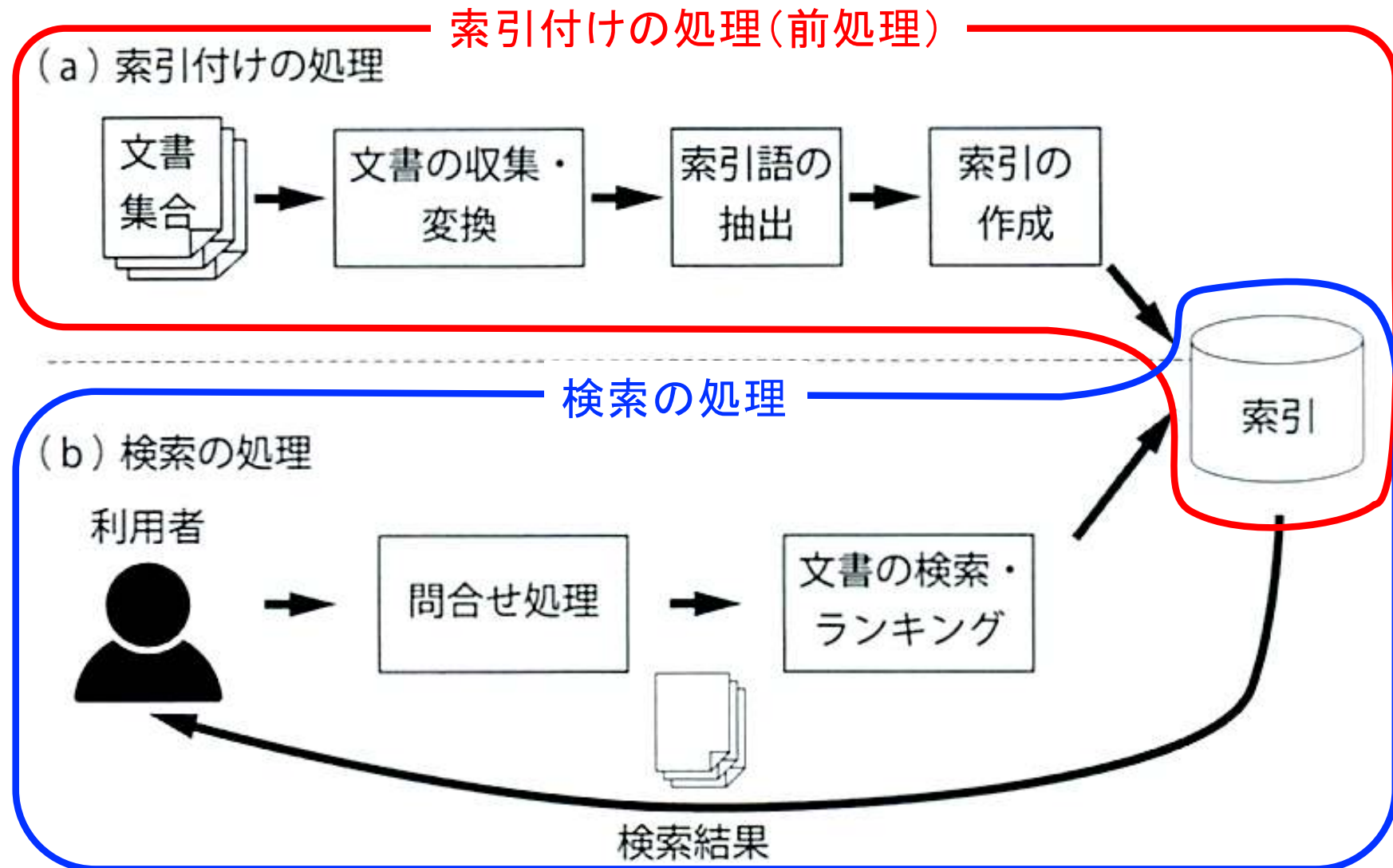
情報理工学部  
前田 亮



# 情報検索システムの構成

- 大きく分けて以下の二つの処理から構成される
- **索引付けの処理**
  - 検索の前処理として事前に行う
  - 検索を実現するための**索引**を作成
  - 文書集合が更新された場合, 索引も更新
- **検索の処理**
  - 利用者が問合せをシステムに入力した時に行う
  - **索引**を用いて, 問合せに適合する文書の検索・ランキングを行う

# 情報検索システムの全体構成



# 索引とは？

- まず、本の索引を考える
  - 語句を一定の順序で並べ、その**所在**を示した表
  - **所在**：その語句が含まれるページ番号
- 索引があるおかげで、本の先頭から調べなくても、素早く目的のページに到達できる

## 148 索引

グラフ 129  
クローラ 14, 19  
クローリング 19  
群平均法 82  
訓練データ 75  
形態素 15, 31, 116  
形態素解析 15, 31  
形態素解析器 15, 116  
言語横断情報検索 109  
検索 1  
検索結果の出力 54  
検索語 41  
検索式 8  
構文トランスファ方式 113  
効率性 62  
コサイン値 46  
コサイン類似度 43, 46  
コーパス 110  
コミュニティ 86  
コーンツリー 131  
コンテンツベースフィルタリング 93

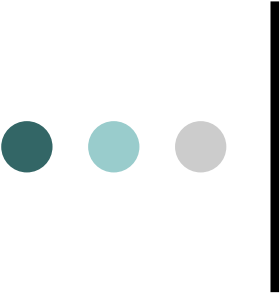
### ▶さ 行

再現率 64  
サイトリンク 60  
索引 7, 9, 31  
索引語 9, 40, 87  
索引語頻度 33  
索引語 - 文書行列 44  
索引付け 9  
サジェスト 60  
サーチ 2  
サポートベクタマシン 121  
サンキヤダイアグラム 133  
支持度 122  
辞書 117  
シソーラス 55, 89  
自動分類 3  
自動要約 59  
シフト JIS 25  
重心法 82

重要度 117  
樹形図 80  
情報 1  
情報アクセス技術 1  
情報可視化 4, 129  
情報検索 1  
情報推薦 3, 91  
情報フィルタリング 91  
情報要求 8  
スキーマ 106  
ストップワード 32  
ストリームグラフ 132  
スニペット 59  
スペル修正 55, 56  
精度 64  
接辞処理 32  
全文検索 7  
相関ルール 122  
相互情報量 119  
ソーシャル検索 86  
ソーシャルメディア 87

### ▶た 行

タグ 87  
タグクラウド 90  
タグのリスト 90  
多言語情報アクセス 3, 108  
単語の極性 120  
単純ベイズ分類器 76  
単連結法 81  
逐次探索方式 6  
ツリーマッピング 131  
ディレトリ型検索エンジン 10  
適合性フィードバック 17, 49, 54, 58  
適合率 64  
テキストマイニング 3, 115  
デスクトップ検索 15  
テストデータ 76  
データベース管理システム 62  
転置索引 16, 35  
デンドログラム 80, 130



# 情報検索における索引

- 情報検索の索引でも同様に、索引を使うことで、ある語句が含まれる文書を素早く見つけることができる
- 本の索引における**所在**は、**ページ**
  - ・ 一つの本の中のどのページにあるか
- 情報検索の索引における**所在**は、**文書**
  - ・ たくさんある文書の中のどの文書にあるか



# なぜ索引が必要か？

- 索引がないと、検索のたびにすべての文書を調べなくてはならない
  - たとえばWebの場合、少なくとも数百億の文書があるため、検索のたびにすべての文書を一つ一つ調べるのは非現実的
  - そのため、情報検索システムでは**事前に**索引を作成しておく



# 索引付けの処理手順

## 1. 文書の収集・変換

- 索引付けの対象となる文書を収集し、テキストデータに変換

## 2. 索引語の抽出

- 各文書のテキストデータから索引語を抽出

## 3. 索引の作成

- 索引語の集合から、高速な検索を行うためのデータ構造である索引を作成



# 文書の収集・変換

- 索引付けの処理で最初に行うことは、検索対象とする文書の収集
  - 文書の種類や性質によって方法が異なる
    - Web上の文書, PC内の文書, 企業内の文書, データベース上の文書, etc.
- ここでは、代表的な文書の種類について、収集・変換の方法を説明
  - 詳しくは第3回で説明



# ● ● ● | Web文書の場合

- Web文書は, Webページ間のハイパーリンクをたどることで収集
- Webクローラ (Web crawler)
  - Web文書を, リンクをたどって収集・保存
  - Webでは常に文書が追加・更新・削除される
    - 膨大なWebページを効率的に収集
    - 最新の状態を保持する必要
  - Web検索エンジンによるクローラの例:  
Googlebot, Bingbot



# 個人のPC内の文書の場合

- 個人PC内の文書検索の機能は、最近のOSでは標準搭載されている
  - Windows Search, macOSのSpotlight
- **デスクトップ検索**と呼ばれる
  - 常にバックグラウンドで動作し、ハードディスクを走査して索引を構築・更新
  - テキスト・Office文書・PDF・電子メール・各種メディアファイルなどが対象



# 文書のテキストへの変換

- Web・PCなどにはさまざまな文書形式が混在
  - 索引付けを行うには、これらの文書からテキスト部分(およびメタデータ)を抽出する必要
    - HTML, Word, PowerPoint, PDF → テキスト
    - テキスト抽出のためのソフトウェアが必要
  - 文字コードの変換が必要になる場合もある
    - 日本語の場合, シフトJIS, EUC, JISコード, Unicodeが混在
    - 特定の文字コードに統一する必要
      - たとえばUnicode



# 索引語の抽出

- テキストを単語に分割し、索引語として抽出
- 言語によって方法が大きく異なる
- 英語の場合
  - 不要語の除去
    - “and”, “or”, “the”, “in”など
  - 語幹の抽出
    - “computer”, “computers”, “computing”, “compute”
- 日本語の場合
  - 形態素解析, 品詞による選別



# 形態素解析

- 日本語では語の区切りがないため、単語の分割は容易ではない
  - 形態素解析という処理を行う
- 「形態素 (morpheme)」は文の構成要素であり、意味を持つ最小の言語単位
  - 形態素と単語は必ずしも一致しないが、本講義では同等のものとして扱う
- 日本語の形態素解析器
  - ChaSen (茶筌), MeCab (和布蕪), JUMAN++, Sudachi, etc.

# 形態素解析の例

- 形態素解析器は、文を形態素に分割するだけでなく、形態素の品詞・読み・活用なども解析
- 助詞，代名詞，連体詞などは，索引に不要な語として使用しない場合が多い

文

この条約の締約国は、地球の気候の変動及びその悪影響が人類の共通の関心事であることを確認し、...



形態素解析

形態素の列

<del>この</del>	条約	<del>の</del>	締約	国	.....
連体詞	名詞・一般	助詞・連体化	名詞・サ変接続	名詞・接尾	



# 索引の作成

- 索引付けの中心となる部分
- 情報検索で用いる索引は**転置索引**と呼ばれる
  - 「**文書-索引語**」の情報を「**索引語-文書**」の情報に変換する
  - ある索引語がどの文書に含まれるか？
  - 文書集合中の各文書には固有のIDを付与
    - ・ **文書ID**と呼ぶ
- 転置索引によって、利用者が入力した問合せの各単語がどの文書に含まれるかを高速に見つけることができる



# 転置索引の例

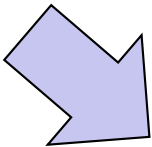
文書1  
(ID=1)

この条約の締約国は、地球の気候の変動及びその悪影響が人類の共通の関心事であることを確認し...

文書2  
(ID=2)

「気候変動の悪影響」とは、気候変動に起因する自然環境又は生物相の変化であって、...

索引語



## 転置索引

単語	文書ID
条約	1
締約	1
国	1
気候	1, 2
変動	1 2





# 索引に含まれるその他の情報

- 索引には文書IDの情報が必須だが、その他の情報を含める場合がある
- 索引語の**出現頻度**
  - 索引語が文書中に何回出てきたか？
  - 検索結果の文書のランキングに用いられる
- 索引語の**出現位置**
  - 索引語が文書のどこにでてきたか？
  - フレーズ検索などに用いられる
- 索引について詳しくは第4回で説明



# 検索の処理

## ○ 問合せ処理

- 利用者による問合せの入力や改良の支援

## ○ 文書の検索・ランキング

- 問合せと索引を用いて文書を検索・ランキングする



# 問合せ処理

## ○ 問合せの入力

- 問合せは、単なるキーワードの羅列だけではない
- 利用者が**問合せ言語**を用いて入力した問合せを解釈し、検索・ランキング処理に渡す
- 問合せ言語には様々な演算子が用いられる
  - AND, OR, NOT, フレーズ検索, 検索対象の限定, 部分一致など



# 問合せの改良

## ○ スペル修正

- 問合せ中のスペルミスを自動的に修正（あるいは修正候補を提示）

## ○ 問合せ候補の提示機能

- 問合せを入力している途中で、候補となる問合せを提示

## ○ 問合せ拡張

- 問合せの同義語などを自動的に追加

## ○ 適合性フィードバック

- 最初の検索結果を基に、より良い検索結果を得るために問合せを修正

## ○ 詳しくは第6回で説明

# スペル修正の例



serect

すべて

動画

画像

ショッピング

地図

もっと見る

約 2,280,000,000 件 (0.55 秒)

次の検索結果を表示しています **select**

元の検索キーワード: serect

英語 ▼



日本語 ▼

select

編集

sə'lekt

選択

Sentaku

# 問合せ候補の提示機能の例



立

- 立命館大学
- 立命館大学 偏差値
- 立命館守山高校
- 立命館
- 立命館大学 図書館
- 立花孝志
- 立木観音
- 立木神社
- 立山黒部アルペンルート
- 立川

Google 検索 I'm Feeling Lucky

不適切な検索候補の報告

# 問合せ拡張の例

The screenshot shows a Google search for 'インタフェイス' (Interface). The search bar contains 'インタフェイス'. Below the search bar, the results for 'インタフェイス' are displayed. The first result is from Weblio国語辞典, titled '「インタフェイス」の意味や使い方 わかりやすく解説 ...'. The second result is from Wikipedia, titled 'インターフェイス - Wikipedia'. Red arrows point from the search bar to the Wikipedia result, and from the Wikipedia result to the search bar. Red boxes highlight the search bar, the Wikipedia result, and the text 'インターフェイス' in the Wikipedia snippet. The Wikipedia snippet also contains the text 'インタフェイス' and 'インターフェイス'.

Google インタフェイス - Google 検索

google.com/search?q=インタフェイス&hl=ja&ei=CSI-ZK7IIJnl2roP6PWUyAU&oq=インタフェイス&g...

Google

インタフェイス

すべて 画像 ショッピング ニュース 動画 もっと見る ツール

約 1,460,000 件 (0.42 秒)

Weblio国語辞典  
https://www.weblio.jp/content/インタフェイス

「インタフェイス」の意味や使い方 わかりやすく解説 ...

1 異なる種類のものをつなぐときの共用部分。界面。接触面。2 コンピューターで、機器やプログラムどうしをつなぐ装置、または部分。ハードウェアを接続するハード ...

Wikipedia  
https://ja.wikipedia.org/wiki/インターフェイス

インターフェイス - Wikipedia

インターフェイス (英: interface) はインタフェイス、インターフェイスとも書き、英語で界面や接触面、中間面などといった意味を持ち、転じて ...

他の人はこちらも質問 :

インターフェイスとはなにか?

インタフェイス

情報技術

情報技術において、インタフェイスは、情報行うシステム間のプロトコル、または、その部分をいう。コンピュータシステムの各部分はシステム間の接続や、人間と機械の間の入道がある。インターフェイスあるいはインタスなどと表記することもある。ウィキペディ

フィ



# 文書の検索・ランキング

- 問合せに一致する結果を得るだけでなく、より適切な結果を上位に持ってくることが重要
  - ある問合せに対して、各文書がどれだけ適切であるかを示すスコアを計算し、ランキング
- 検索システムの核心となる最も重要な部分
- ブーリアンモデル、ベクトル空間モデル、確率モデルなど、様々な検索モデルがある
  - 詳しくは第5回で説明



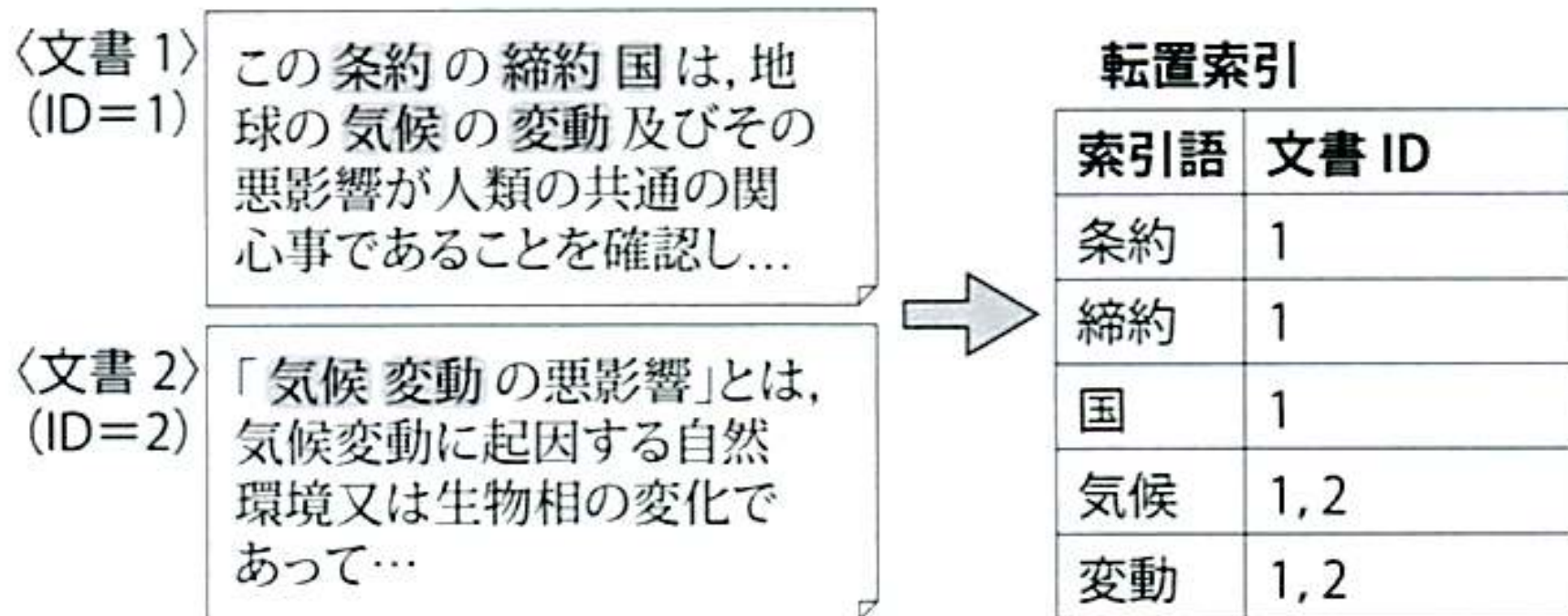


# 教科書 第2章の演習問題の 解説(1/3)

- 2.1 2.2.2項(スライドでは12~14枚目)で述べた方法で日本語の文書から索引語の抽出を行った場合, 実際に検索する際に問題となると思われる点をいくつか挙げよ.
- 複合語が複数の索引語に分かれてしまう場合がある
  - 「締約」「国」, 「プロ」「野球」, 「北」「区」, etc.
- 形態素解析の結果は必ずしも正しいとは限らない
  - 「東京」「都」 or 「東」「京都」?
- 助詞, 代名詞, 連体詞などを索引語として用いない場合, それらを含む検索ができない
  - 「君の名は」のうち, 「名」しか索引に含まれない

# 教科書 第2章の演習問題の 解説(2/3)

- 2.2 図2.4(下図)に示した二つの文書が検索システムで索引付けされていて、これに対して「気候変動」という問合せを行うとする。文書1と文書2のどちらを上位にランキングするのが適切かについて検討し、その理由を述べよ。



# 教科書 第2章の演習問題の解説(3/3)

「文書 1」  
(ID=1)

この条約の締約国は、地球の気候の変動及びその悪影響が人類の共通の関心事であることを確認し...

「文書 2」  
(ID=2)

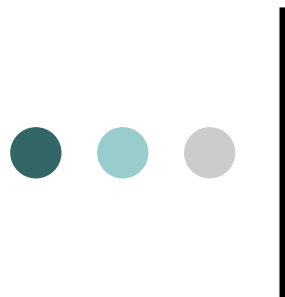
「気候変動」の悪影響とは、気候変動に起因する自然環境又は生物相の変化であって...

➡

転置索引

索引語	文書 ID
条約	1
締約	1
国	1
気候	1, 2
変動	1, 2

- 文書1では「気候」「変動」が1回ずつ、文書2では2回ずつ出現している
- このことから、文書2の方が「気候変動」について述べている可能性が高いと考えられる



# まとめ

- 情報検索システムの各構成要素について概略を説明した
- 情報検索システムは、大きく分けて**索引付け**の処理と**検索**の処理の二つから構成されている
- 各構成要素の詳細は次回以降で説明する