「リサーチ＆プランニング」
第二回　検索の達人になる

デジタルハリウッド大学
橋本大也

本日の内容

- 今回と次回講義は検索の基本と応用
  - 今日は基本的な事項の理解

- 検索エンジンの仕組みを考えるWG（20分）
- 検索エンジンの仕組み（30分）
- 検索エンジンを使いこなす（40分）

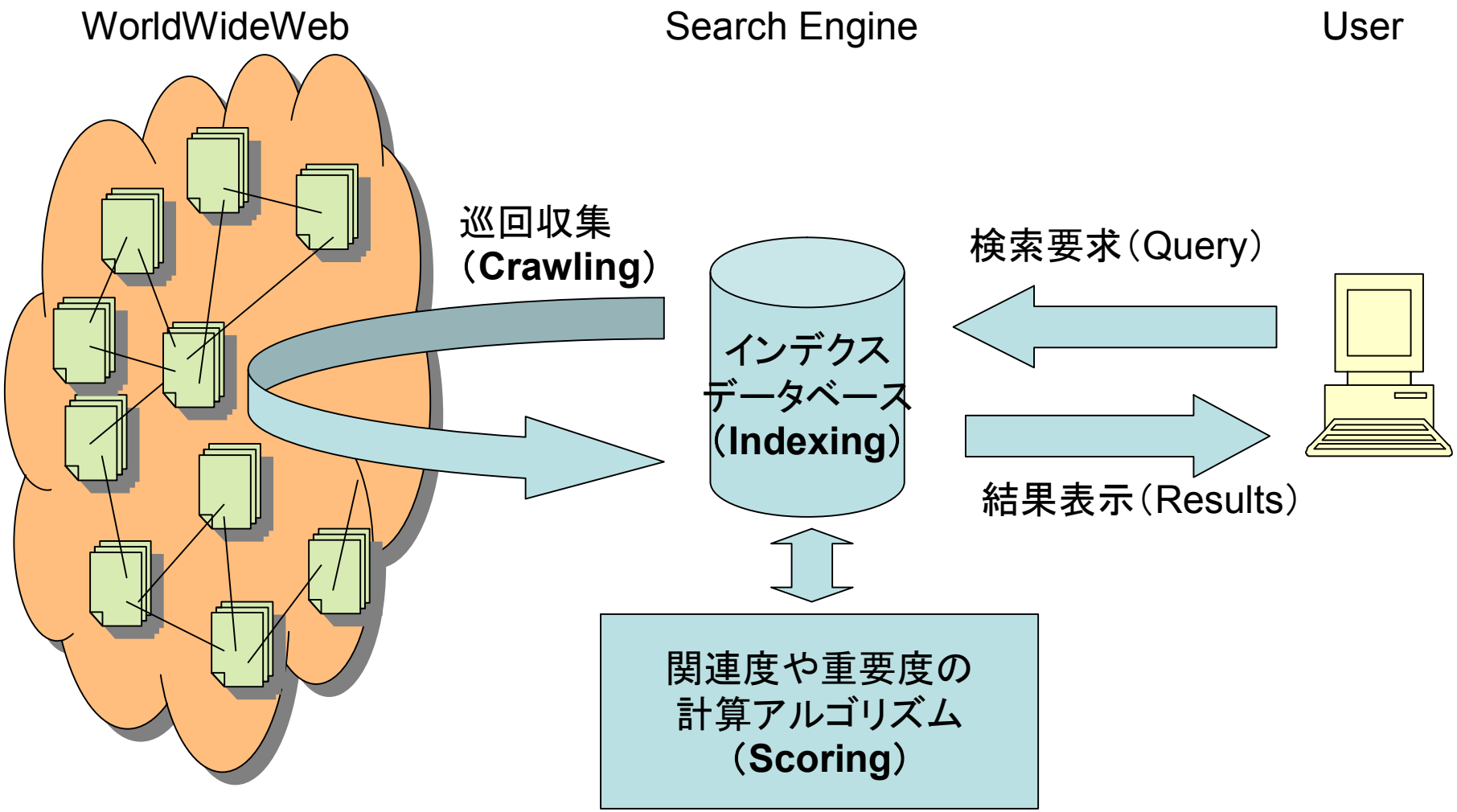
- 達成目標
  - 検索エンジンの仕組みを自分で考えてみる
  - 現実の仕組みを理解する
  - 効果的な使い方について知る

第一部　検索エンジンの仕組みを考える

- 検索エンジンの仕組みについて、自由に想像して考えてみてください。どんな部品が必要でしょうか。左端にWWW、右端にユーザ（あなた）、その真ん中に検索エンジンを配置して、全体を簡単な図に描いてください。

- 考える上でのポイント
  - なぜ検索エンジンは数十億ページを瞬時に検索できるのでしょうか？
  - 関連度が高い有名なページが結果表示で上位に表示されるのはなぜでしょうか？

第2部　検索エンジンの仕組み



考察ポイントの答え

- 数十億ページを瞬時に検索できるのは
  - あらかじめWWWのコピーとその索引を作成しておき、ユーザの要求に応じてWWWそのものではなく、小さな索引を検索しているから。
  - 人間が分厚い本の索引からキーワードを短時間で発見できるのと同じ

- 関連度の高い重要なページが上位表示されるのは
  - 検索語に対する関連性（Relevancy）や重要度を計算するアルゴリズムが組み込まれているから。

仕組みの要点をまとめると

- WWWをロボットがリンクをたどって定期巡回してコピーをサーバに持ち帰る（Crawling）
- どんな単語や文字列パターンが、どのページの何文字目にあったかの索引を作成する(Indexing)
- 関連度や重要度を計算して、結果表示の順位を決定する(Scoring)

Web検索エンジンに求められる能力

- 漏れなく最新のWWW全体を高速巡回する能力（Crawling）
- 検索要求に対して最適化された索引を作成する能力(Indexing)
- ユーザにとって関連度や重要度の高い情報を上位に表示する能力(Scoring)

索引を作る代表的な手法2例

1. 形態素解析・分かち書き法
   - 辞書にある単語で索引を作成する
     - 辞書にない単語を検索できない
     - インデクスが比較的小さいサイズになる
     - 検索時のノイズが比較的少ない
     - 類義語検索や自動分類など高度な検索への応用がしやすい

2. N-gram法
   - N文字ずつずらしたパターンで索引を作成する
     - 辞書にない単語を検索できる
     - 意図しない検索結果が含まれやすい
     - インデクスが比較的大きなサイズになる

ー長ー短がある。他にも多数、発明されている。

もっと検索エンジンについて知りたい人は
Namazu、Chasenについて調べてみよう

Googleが便利だと言われるのは？

- スコアリングで工夫があるから、と言われる。
- ・Google の秘密 - PageRank 徹底解説
- http://www.kusastro.kyoto-u.ac.jp/~baba/wais/pagerank.html

- ページの人気度をリンクの数から計算する仕組み（Link Popularity ＆ Page Rank）
  - リンクされることは人気があることだ、という仮説
  - たくさんのページからリンクされているページは重要なページだ（Link Popularity）
  - そうした重要なページからリンクされているページもまた重要だ(Link Popularity)
  - ページごとにPageRankというスコアを計算
  - PageRankスコアの高い順に検索結果を表示する

第3部　検索をつかいこなす

- Googleを代表ケースとして取り上げる
  - 他の検索エンジンでも同様機能がある
- 検索語の工夫、検索オプションの使い方
- 検索エンジンの利用知識は個人の情報収集の効率を大きく左右する

1　AND、OR、NOT検索

- 「私」と「あなた」の両方が入ったページを検索したいAND検索
  - 私　あなた
- 「私」あるいは「あなた」どちらかが入ったページを検索したいOR検索
  - 私 | あなた
- 「私」の検索結果から「あなた」の入っているページを抜きたい
  - 私 –あなた

- 応用例
- (デジタルハリウッド | デジハリ)　大学 – 大学院
- デジタルハリウッド大学についてのみ探す

２　フレーズ検索

- “長年のご愛顧ありがとうございました”
  - 盛者必衰検索、失敗例を探す
- “サービスをご利用のみなさまへ“
  - サービス約款を探す
- “発送をもって　ipod”
  - 懸賞キャンペーンを探す

- 二重引用符で囲うのがポイント。フレーズを単語で区切られない。
- 発見したいページに特徴的に現れる言葉に注目する。

３ ドメイン指定

- 「site:dhw.co.jp 杉山」

- 検索機能がないサイトを検索する
- デジハリのサイトから校長先生の情報を探す

- 「site:*.ac.jp」 学術関連を探す
- 「site:*.go.jp」 政府公式を探す

４　イメージ検索

- 画像につけられた説明文や同じページに登場するキーワードを検索する

- 変化球：　松浦亜弥 1024 768

５　ファイル形式で絞る

- filetype:pdf 検索
- PDF形式で、「検索」というキーワードが入った文書のみが検索対象になる。
- filetype:ppt filetype:doc filetype:xls

６　日付指定で絞る

- daterange:2453006-2453371
  - 2004年1月1日から2004年12月31日までの更新日付のページを対象に検索する
  - 数字はユリウス日
  - 紀元前4713年1月1日からの経過日数

- Google Daterange Checker
  日付からユリウス日を計算してくれる
  - http://google.bookstudio.com/daterange.htm

7　とは、といえば、どうよ

- デジタルハリウッドとは
- 検索エンジンとは
- XMLとは

- とは、といえば、どうよ、などをつけることで
  - 定義
  - 評価
  - 関連

    など広げて探すことができる

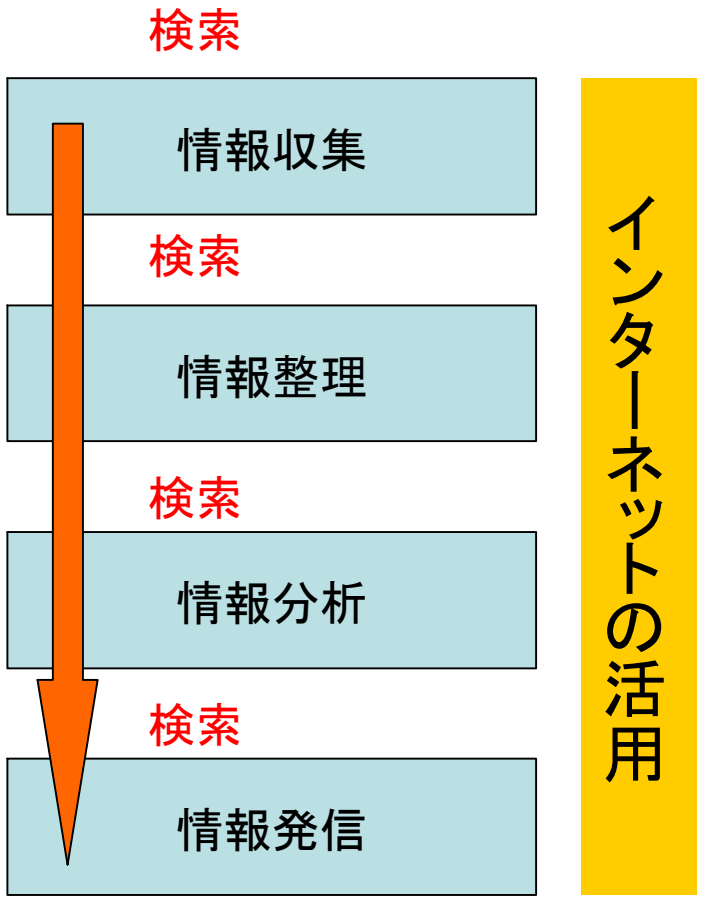
情報を探索するプロセス

- 情報探索プロセスに関する研究の紹介
- 情報探索を成功させるには？

情報探索のプロセスは検索の連続

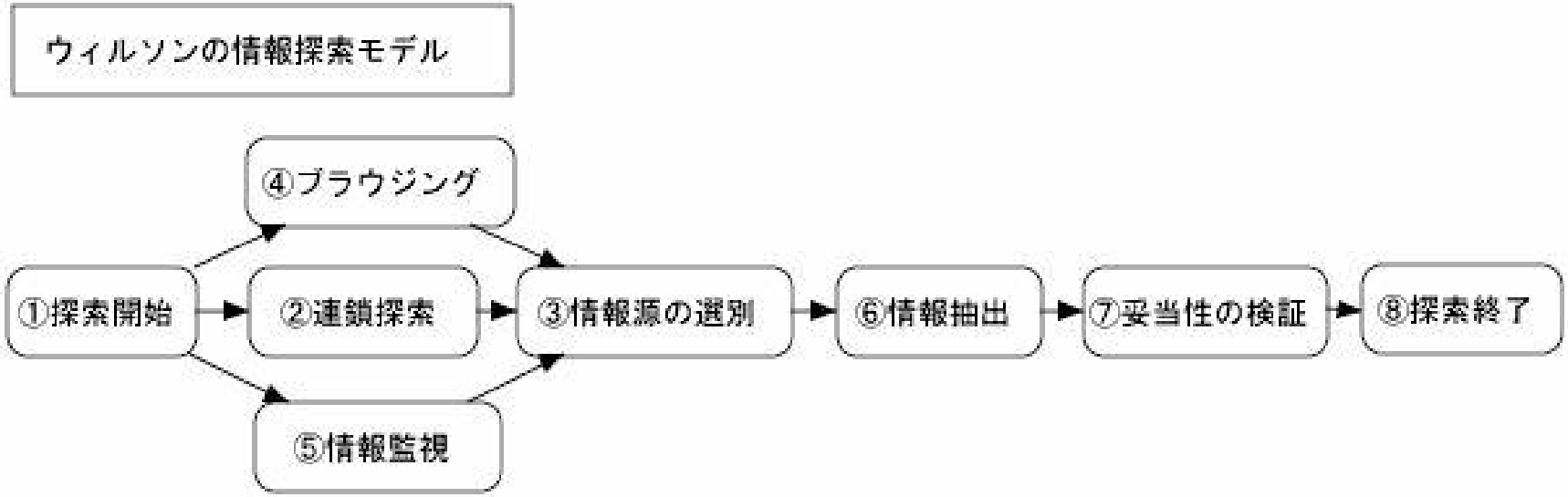
この講義が想定するワークフロー

- アツメル
  - 情報を収集する
- ナラベル
  - 情報を整理する
- ヒキダス
  - 情報を分析する
- カキダス
  - 情報を発信する



リサーチのワークフロー図

ウィルソン情報探索モデル



連鎖探索＝関連検索、情報監視＝モニタリング
のツールとノウハウの必要性

クールトー情報探索モデル（感情、自己効力感）

段階	第1段階	第2段階	第3段階	第4段階	第5段階	第6段階	
	タスク定義	トピック選択	漠然とした情報探し	フォーカス形成	情報収集	情報探し終了	執筆開始
感情	不確実・不安	漠然とした希望	恐怖・疑い、フラストレーション	明快	方向性、自信	開放感	満足、不満足
思考	漠然 →						明快
情報行動	関連情報を探す →						適合情報を集める

感情や自己効力感が情報探索の質に影響する

楽しく、面白がりながら情報を検索していくことが大切

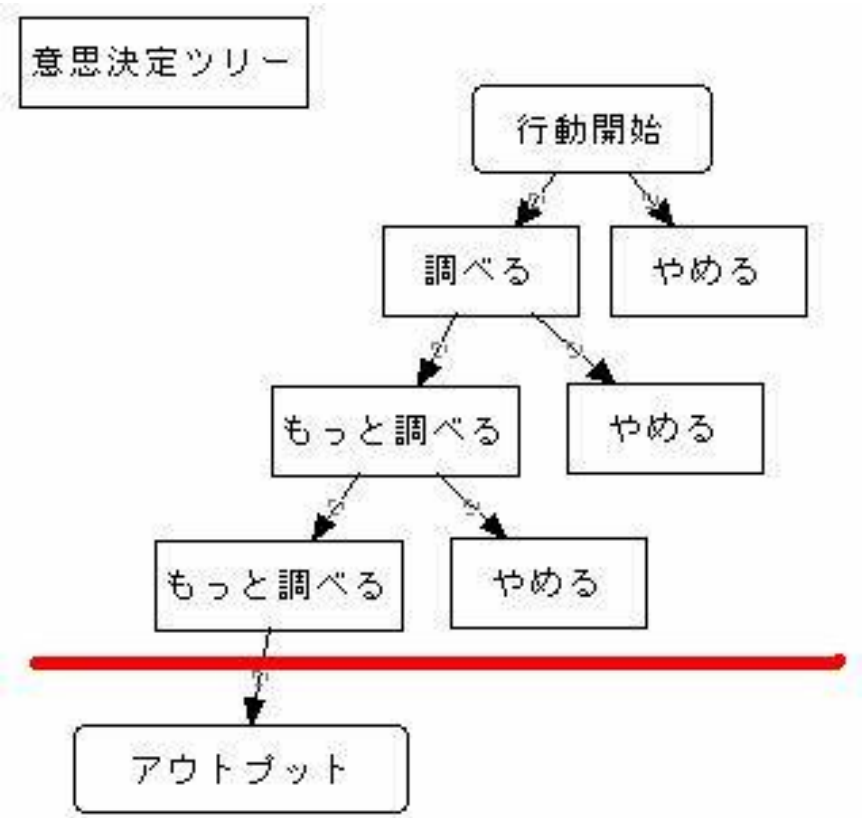
バンデューラの多重ゴールモデル（見えるゴール目指して）

「つまり人間の日常行動は、（1）未来の望ましい出来事（遠隔ゴール）を心に描き、（2）個々の行動の成果を評価する基準（直近ゴール）を設定して、それを実現させる可能性の高い行動を起こすことで生じている」

情報探索行動において直近ゴールとは何？

中間的なアウトプットを作ること＝ここまで調べらことのメモや要約を残す＝進んでいる、できてきている、楽しい

中間アウトプットが鍵を握る



・提案されなかった企画は発注されない
・発表されなかった論文は評価されない
・書かれなかったメールは反応が返ってこない

頭で考えたことがあるだけなのと、メモや文書にまで仕上げたことでは、その結果に天地の差がある。

中間アウトプットの重要性

今日のポイント

- 検索エンジンの仕組みを理解して使う
- 検索語の選び方で工夫をする
- 検索や探索自体を楽しむことが情報をうまく収集する鍵になる
- 情報を何度も検索しながら中間メモを作ることで
  - できてきた確信が持てる
  - 探索が効果的に行える
  - 方向性がぶれない