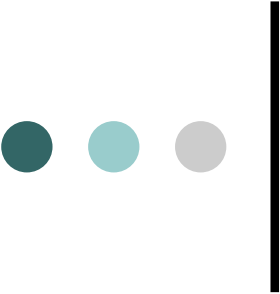




情報アクセス論 第1回

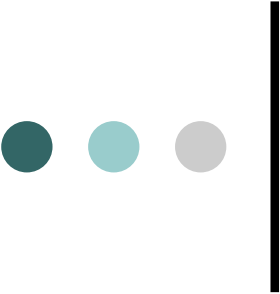
「情報アクセス技術の概要」

情報理工学部
前田 亮



講義内容(シラバス)

- 大量の情報資源の中から必要な情報を効率的に見つけ出すための**情報アクセス技術**が重要になってきている
- 情報アクセスの基本となる**情報検索**の概念と実現手法について理解することを目標とする
- また、文書の分類・クラスタリング, ソーシャル検索, 文書以外の各種メディアの検索, 多言語情報アクセス, テキストマイニング, 情報の可視化など, 情報アクセスに関わる最新の話題についても学ぶ



講義内容(シラバス)

○ 到達目標

- 情報検索の各種理論と, それを実現するための技術について理解する
- 文書以外の各種メディアに対する検索手法を理解する
- 情報検索に関する最新の技術動向を把握する



講義スケジュール

前半(担当:前田)

第1回 情報アクセス技術の概要

第2回 情報検索システムの構成

第3回 文書の収集・加工

第4回 索引付け

第5回 検索モデル

第6回 問合せ処理・ユーザインタ
フェース

第7回 情報検索システムの性能
評価

後半(担当:西原)

第8回 分類・クラスタリング

第9回 ソーシャル検索

第10回 各種メディアの検索

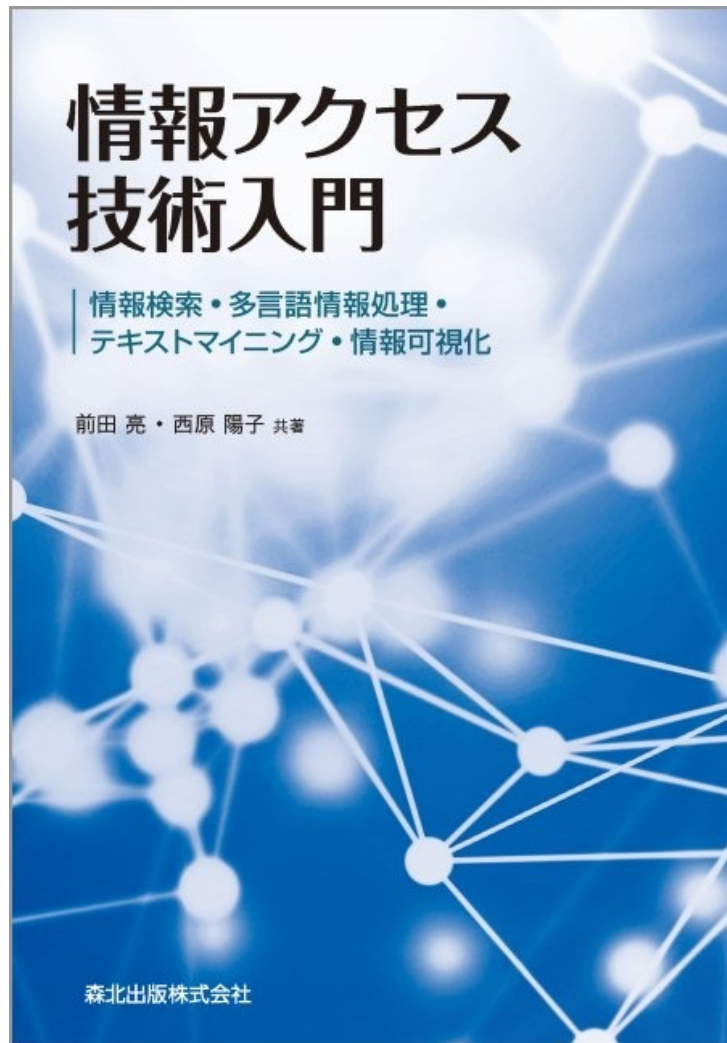
第11回 多言語情報アクセス

第12回 テキストマイニング (1)

第13回 テキストマイニング (2)

第14回 テキスト, データの可視化

教科書



○ 前田 亮, 西原 陽子 共著

「情報アクセス技術入門 -
情報検索・多言語情報処理・テキ
ストマイニング・情報可視化」

森北出版, 2017(2,592円)

○ 生協などで購入しておくこと



評価方法・基準

- **定期試験** (80%), **小テスト** (20%) で評価
 - 簡単な小テストを複数回行います
- 出席について
 - 出席は取りません



情報検索とは

○ 情報検索 (Information Retrieval: **IR**)

● 「情報 (Information)」

- ある決められた表現方法によって伝えられる, 意味を伴ったデータ(『[計算機科学の基礎](#)』)

● 「検索 (Retrieval)」

- 調べて探しだすこと。特に、文献・カード・ファイル・データベース・インターネットなどの中から必要な情報を探すこと。(『[大辞泉](#)』)
- 文書やデータの中から、必要な事項をさがし出すこと。「索引で語を一する」(『[広辞苑](#)』)
- 動詞retrieve(「取り戻す」)の名詞形

「情報検索」の定義

- 「大量の情報の中からユーザの要求を満たす情報を見つけ出すこと」(『情報検索アルゴリズム』)
- 「ユーザの持つ問題(**情報要求**)を解決できる情報を見つけだすこと」(**広義**)
- 「ユーザの検索質問に適合する文書を文書集合の中からみつけだすこと」(**狭義**)
(『情報検索と言語処理』)
- 情報検索の対象は文書に限らない
 - 画像, 映像, 音声, 音楽, etc.



Retrieval vs. Search

- 検索 (**retrieval**) と探索 (**search**) の違い
- ゴールデンレトリバー
(Golden **Retriever**)
 - 狩りのとき、獲物を取って戻ってくる
- **retrieval** は、既にあると判っているものを持ってくる (取り戻す, 引き出す)
- **search** は、対象が見つかるかどうか判らないが、とにかく探す



Retrieval vs. Search

- 図書館の本を検索

- 目的の本があるとあらかじめ判っている
場合が多い

→ **retrieval**



- Webを検索

- 探したいものが見つかるかどうかは
判らない

→ **search**



情報検索 vs. データベース検索

- **データベース検索**の場合は、検索結果は一意に決まる
 - 例:「情報理工学部4回生で取得単位数が96単位未満の学生の氏名と学生証番号を知りたい」
- 一般に**情報検索**では、検索結果は一意に決まらない
 - 例:「立命館大学(Google)」「立命館大学(Bing)」
 - より適切な文書をランキング上位に持ってくることが重要
- 「適切」な文書は利用者によって異なる
 - 「クリエーションコア」



情報アクセスとは

○「情報検索」は古い？

- 情報分野では1950年代から使われている
（「データベース」(1960年代)より古い）

○ 新しい用語「情報アクセス」

- 情報源の**探索**，情報の**分類・検索・フィルタリング**，検索結果の**提示**手法（たとえば**可視化**，**要約**），**マイニング**などを含む，情報検索より広い概念

情報検索の歴史

- 計算機以前
- 図書館における資料の検索
 - カード目録
 - 書名目録, 著者名目録
 - 分類法
 - 日本十進分類



	Murakami, Haruki	見出し
913.6 M972	世界の終りとハードボイルド・ワン derland / 村上春樹著. -- 東京 : 新潮社, 1985.6. 618p ; 20cm 付:参考文献 ISBN: セカイノオワリトハートボイルドワン derland A1:ムラカミ, ハルキ NDC8:913.6	書名
		著者名
19850909		分類番号
8500-03235-9		所蔵場所
図開架	<BN00841804>	
	8500-03235-9	



分類法

○ 日本十進分類

- 概念を階層的に分類する
 - 類・綱・目の3桁で表わす
 - より細かい区分は, ピリオドの後で分類
(例: **007.58**: 情報検索・機械検索)
- 欠点
 - 新しい概念への対応が難しい
(例: **007**: 情報科学, **548**: 情報工学)
 - 複数の概念にまたがる場合

類目表

- 0**: 総記
- 1**: 哲学
- 2**: 歴史
- 3**: 社会科学
- 4**: 自然科学
- 5**: 技術・工学
- 6**: 産業
- 7**: 芸術・美術
- 8**: 言語
- 9**: 文学

Yahoo! カテゴリ

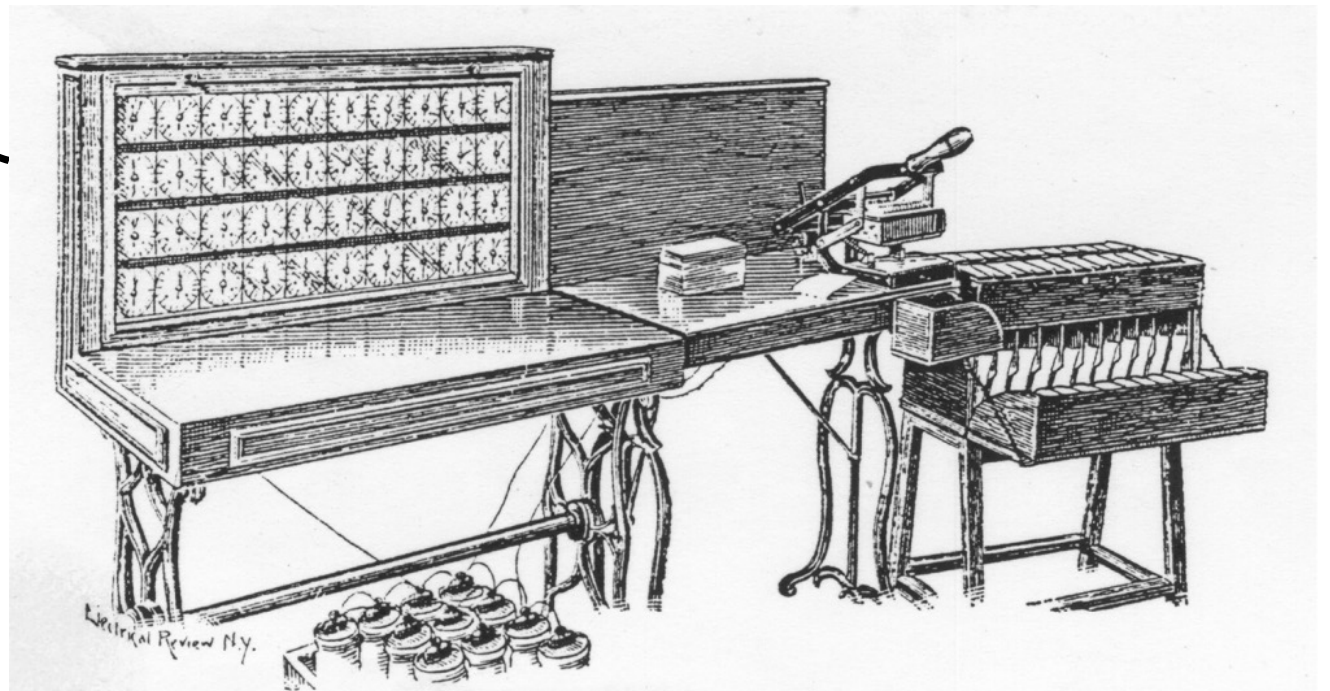
- ディレクトリ型検索エンジン
- カテゴリ(概念)が階層構造になっている
- 人手でサイト情報を収集・分類
- 2018年3月29日にサービス終了



機械による情報検索

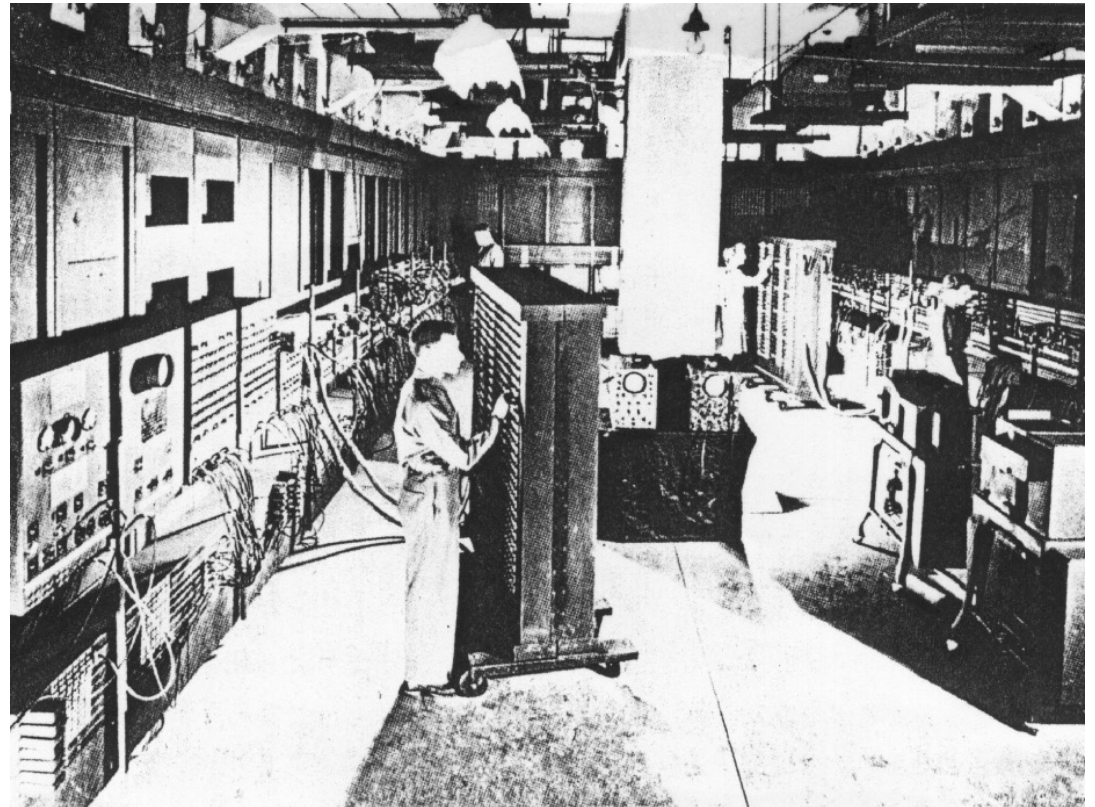
○ パンチカードによる索引 (1940)

- 50 × 20孔のパンチカード
- カードが主題概念を, 孔が文献番号を表す
- 複数の主題を一度に検索できる
- 管理できる
文献数に限界



● ● ● | 計算機による情報検索

- 世界初の電子計算機**ENIAC**(1946)
 - 5年後には計算機を用いた情報検索の研究が始まった
 - 逐次探索方式のため検索に時間がかかる

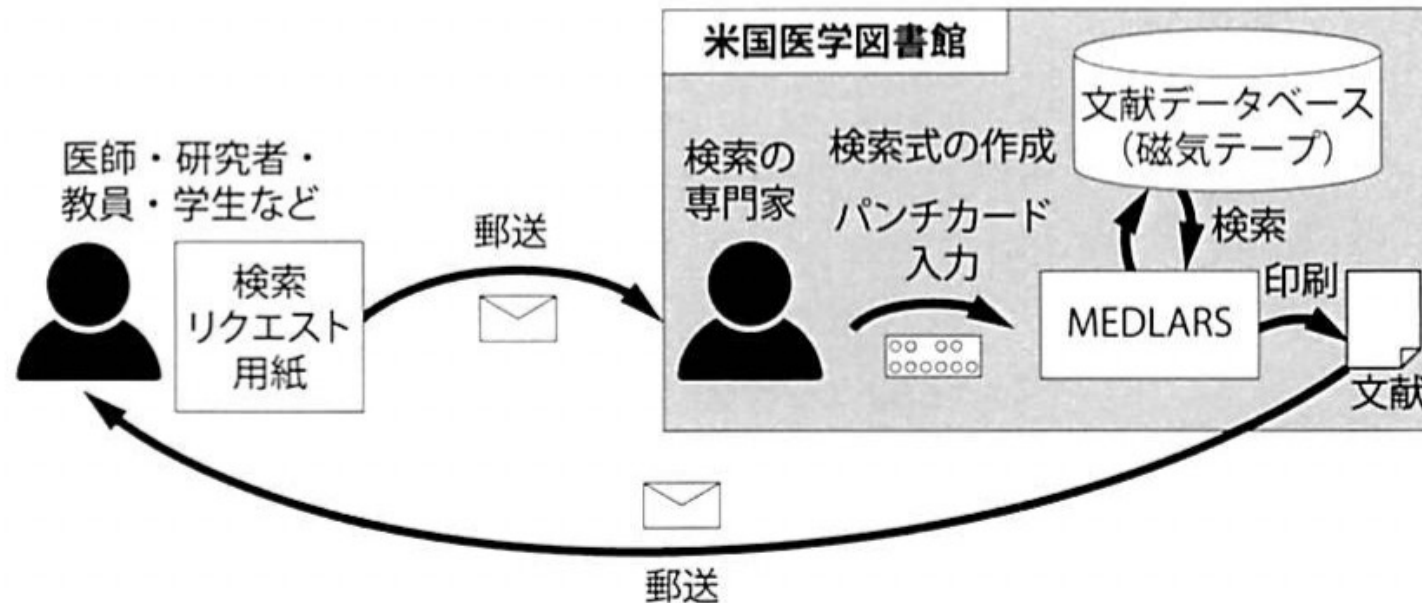


情報検索システムの発展(1)

○ オフラインバッチ型情報検索システム(1964)

● 米国医学図書館のMEDLARS

- 医学文献を対象とした検索システム
- パンチカードで検索式を入力し、一括処理
- 検索結果が返ってくるまで3～6週間かかっていた



情報検索システムの発展(2)

- **オンライン会話型情報検索システム**(1960年代後半～)
 - 直接計算機を操作する対話型
 - 磁気ディスク装置によるランダムアクセス
 - コンピュータネットワークによる遠隔利用
- 現在の情報検索の原型は、1970年代頃にはすでに確立されていた





全文検索

- 過去の検索システムは**二次資料**（タイトル，著者，索引，抄録など）の検索が中心
 - 図書館の蔵書検索（OPAC）などが該当
 - 今で言う**メタデータ**の検索
- **全文検索**とは，**一次資料**（文献の本文）の検索
 - Web検索，新聞記事検索，etc.

● ● ● | Web検索エンジン(1993～)

- 従来からの情報検索の研究成果が利用されている
- それに加えて, Web特有の情報を利用
 - タグ情報, ハイパーリンク構造, 言語情報など
- 最近の検索技術
 - 個人化(パーソナライズド)検索
 - ・ 過去の検索履歴, 過去に読んだページ, 現在地など
 - エンティティ検索
 - ・ 人名・地名・組織名などのエンティティ(実体)から, 構造化された知識ベースを検索(例: アップル)

Web検索エンジンの具体例(1)

○ Google (1998～)



- スタンフォード大学の大学院生の研究プロジェクトとして開始
- ハイパーリンク構造を利用してWebページの重要度を測るPageRankを考案
 - 「Web情報技術概論」で詳しく説明
- 200以上の要素をランキングに利用
 - [Google's 200 ranking factors](#)
- 地図・メール・動画・ニュース・機械翻訳・Webブラウザなど, さまざまなサービスを展開
 - 本講義で扱うのは検索機能のみ

● ● ● | Web検索エンジンの具体例(2)

○ **Microsoft Bing** (1998～) Microsoft Bing

- マイクロソフトにより, MSNサーチとして開始
- 主にWindowsから利用される
- 検索エンジンのシェア争いでは苦戦
 - Google: 71.61%, Bing: 21.11%, Yahoo!: 2.53%
 - (Search Engine Market Share, NetApplications.com, 2022/4～2024/3)
- Google同様, 地図・動画・機械翻訳・ショッピングなど, さまざまなサービスを展開
 - 本講義で扱うのは検索機能のみ

● ● ● | Web検索エンジンの具体例(3)

○ **Yahoo!** (1994～)



- ディレクトリ型検索エンジン([p.15](#))として開始
- 以前は検索エンジンを自社開発していたが、現在はMicrosoft Bingの検索エンジンを利用

○ **Yahoo! JAPAN** (1996～)



- Yahoo!とソフトバンクの合併で設立されたが、現在はYahoo!とはほぼ無関係
- 現在はGoogleの検索エンジンを利用



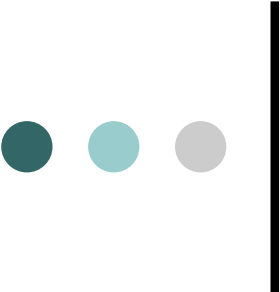
「情報要求」とは

- 情報検索の定義(再掲)

- 「ユーザの持つ問題(**情報要求**)を解決できる情報をみつけたすこと」(**広義**)

- 情報要求(**information need**)

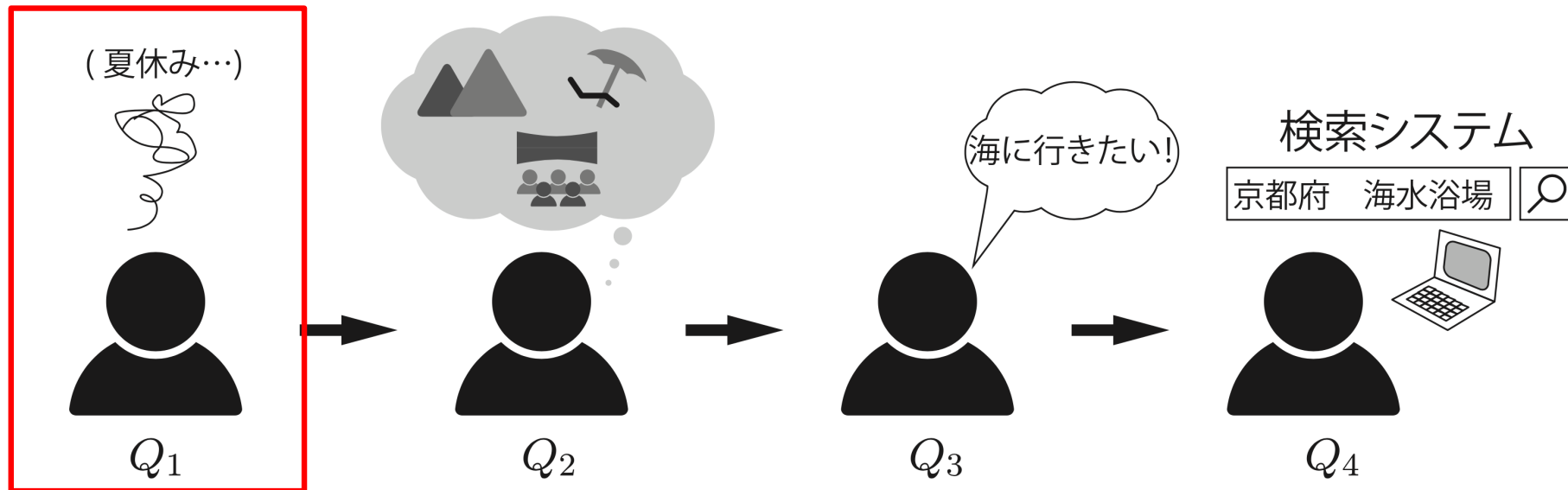
- 「ユーザがある目的を達成するために現在持っている知識では不十分であると感じている状態」
(『**情報検索と言語処理**』)



情報要求の分類 (Taylor, 1968)

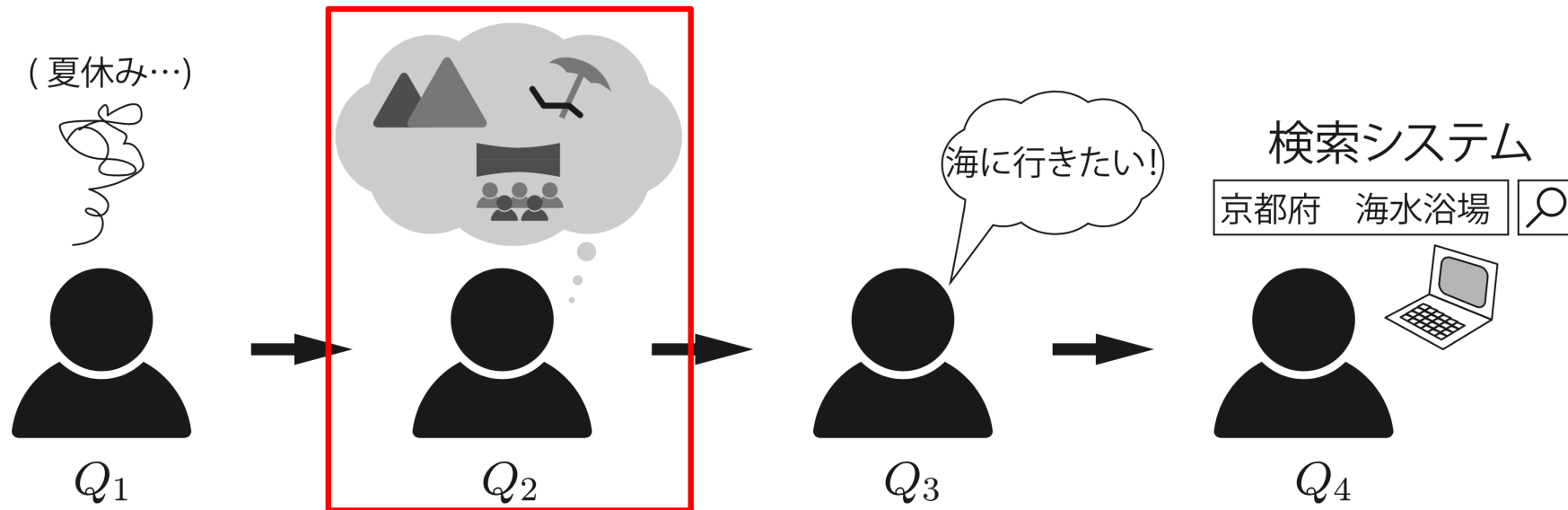
- Q_1 : **直感的** (visceral) 要求
 - 現状に満足していないことは認識しているが、具体的に言語化して説明できない
- Q_2 : **意識された** (conscious) 要求
 - 頭の中では問題を意識できるが、あいまいでまとまりのない表現でしか言語化できない
- Q_3 : **形式化された** (formalized) 要求
 - 問題を具体的な言語表現で言語化できる
- Q_4 : **調整済みの** (compromised) 要求
 - 問題を解決するために必要な情報の情報源が同定できるくらい具体化された状態

情報要求の分類(Q_1)



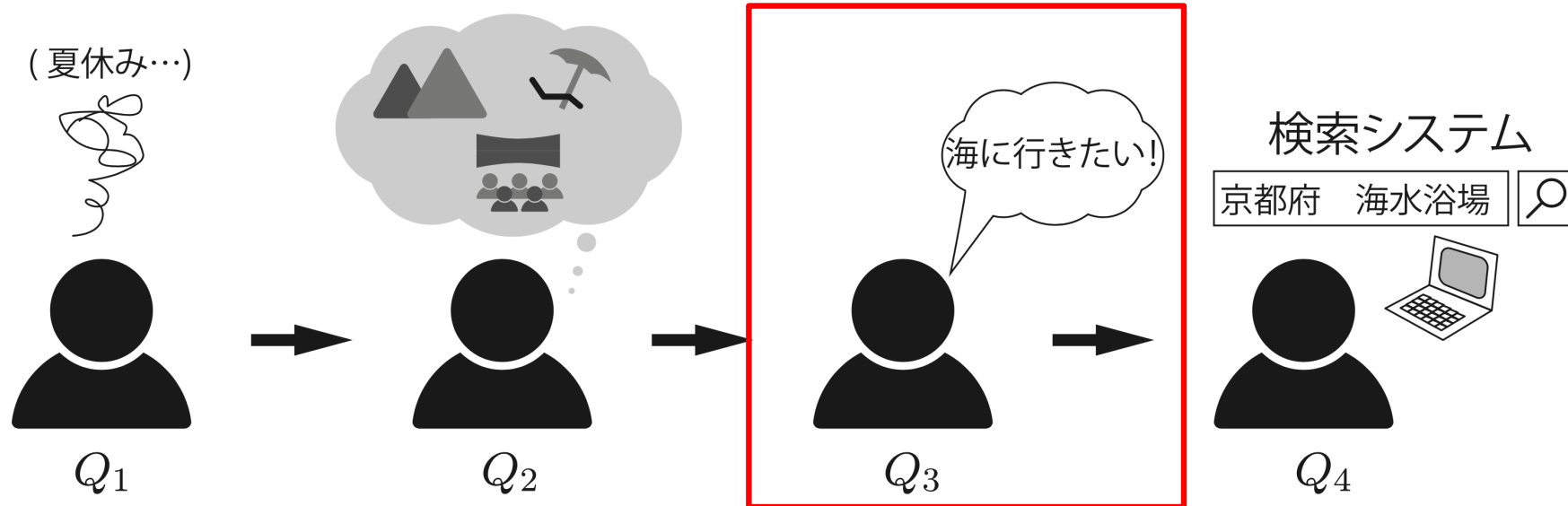
- Q_1 の段階は、とにかく何か情報が足りない、という非常に漠然と困った状態
 - 問題があることは認識できるが、それを言語化できない

情報要求の分類(Q_2)



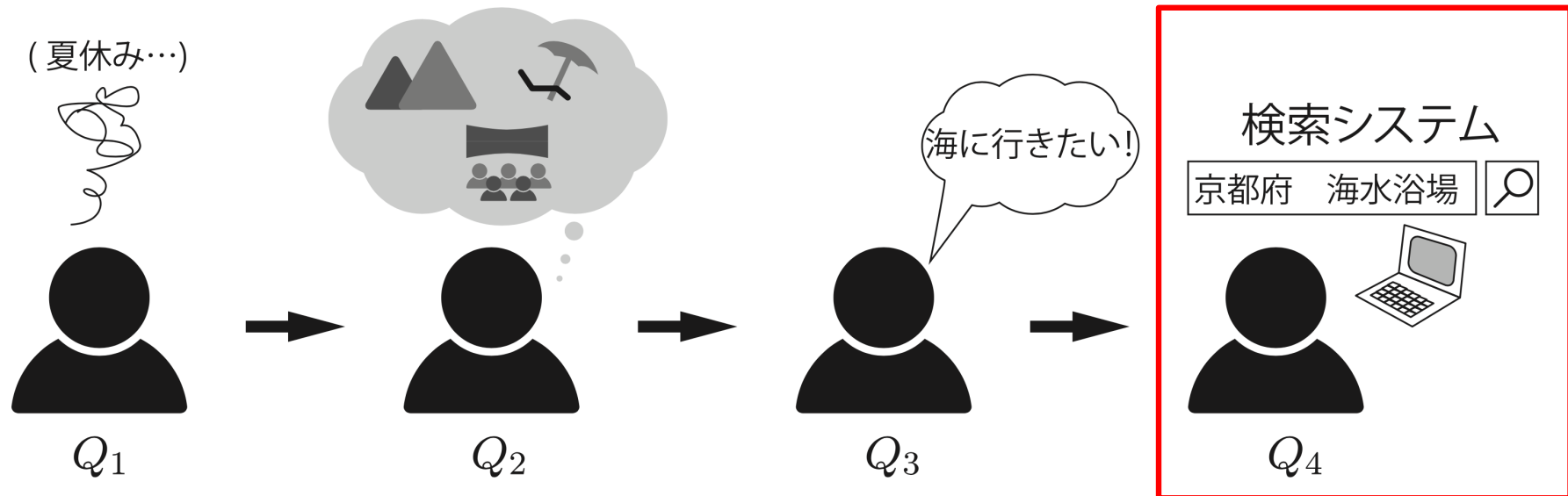
- Q_2 の段階は、何が問題かは何となくわかるが、まだ明確には言語化できない状態
 - 言語化する努力によって問題が明確に

情報要求の分類(Q_3)



- Q_3 の段階は、問題が整理できて、明確に言語化できる状態
 - どうやって問題が解けるかが分かっているわけではなく、何が問題かが明確になっただけ

情報要求の分類(Q_4)



- Q_4 の段階は、問題解決のために具体的にどう
いう手段があるかが見えてきた状態
 - たとえば、身近な情報検索システムでどういう
キーワードで検索すれば必要な情報が得られ
るかがわかる

問題解決の戦略

- $Q_1 \sim Q_4$ のどの状態にあるかによって問題解決のための戦略は異なる
- 必要な書籍の名前が分かっている(Q_4)
 - 図書館の蔵書検索システムを使う
 - 司書に相談する
- 問題そのものがあいまいな場合($Q_1 \sim Q_2$)
 - 同僚に相談する
 - その分野の専門家に相談する
 - 自分で文献を調査する

情報要求を具体化
($Q_1 \rightarrow Q_2 \rightarrow Q_3 \rightarrow Q_4$)



情報要求と情報検索システム

- 本来は、情報要求を動機として必要な情報を見つける行為を支援するのが情報検索システム
 - 広義の「情報検索」
- 実際には、ユーザの問合せに対して適切と思われる文書を提示するのが情報検索システム
 - 狭義の「情報検索」
- 現在の検索システムのほとんどは、 Q_4 の段階にあるユーザを前提としたシステム
- 一般に**問合せ**と呼ばれるものは、通常は Q_4 の段階のものを指す



情報検索の用語の定義

○ 文書集合 (**document collection**)

- 検索対象となる文書群

○ 問合せ (**query**)

- 利用者の情報要求を情報検索システムに入力できる形で表現したもの

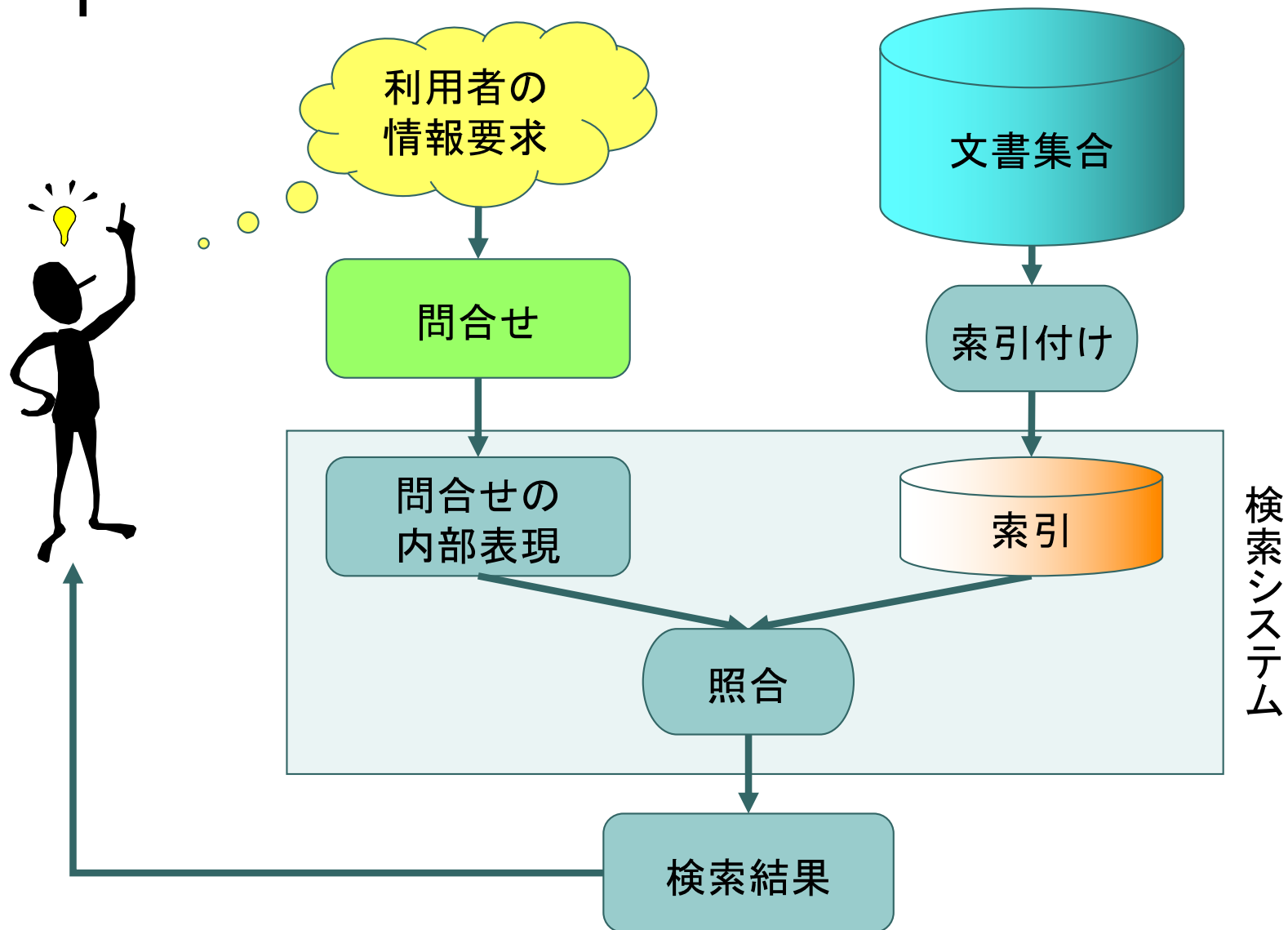
○ 索引 (**index**)

- 検索を高速化するために、あらかじめ文書集合から作っておくデータ構造

○ 索引付け (**indexing**)

- 文書中から索引語を抽出し、索引を作成する処理

情報検索の流れ





まとめ

- 情報検索の定義
 - データベース検索との違い
- 計算機の出現以前から、情報検索は行われていた
 - 図書館のカード目録, 分類法, パンチカード
- 計算機の進化とともに、情報検索の技術も発展
- インターネットの普及により、情報検索の重要性が再認識
 - 現在の情報化社会を支える基盤技術の一つ