情報検索演習
第2回資料
講義の続き、
CD-ROM検索の基本

鶴見大学
2009年10月7日
江草由佳
国立教育政策研究所
yuka@nier.go.jp

本日のお品書き

- 課題の提出
- 情報検索の評価
- データベースとは（起源、定義）
- 第1回レポート課題出題
- CD-ROM版情報検索の演習
- CD-ROM検索の基本
  - 第2回演習課題：CD-ROM検索の準備と検索のおおまかな流れ
  - 第3回演習課題：CD-ROMブラウズ

情報検索結果の評価(1) –p.23

- 検索結果の評価
  - 検索終了後、求める情報が適切に検索できているか、検索漏れはやノイズがないかどうかをチェックする
- 検索漏れ
  - 本来必要な情報でデータベースに存在するにもかかわらず検索されなかった情報
- ノイズ
  - そのテーマに不要な情報が入り込んで検索された情報

p.? は
参考書のページ数
を表す

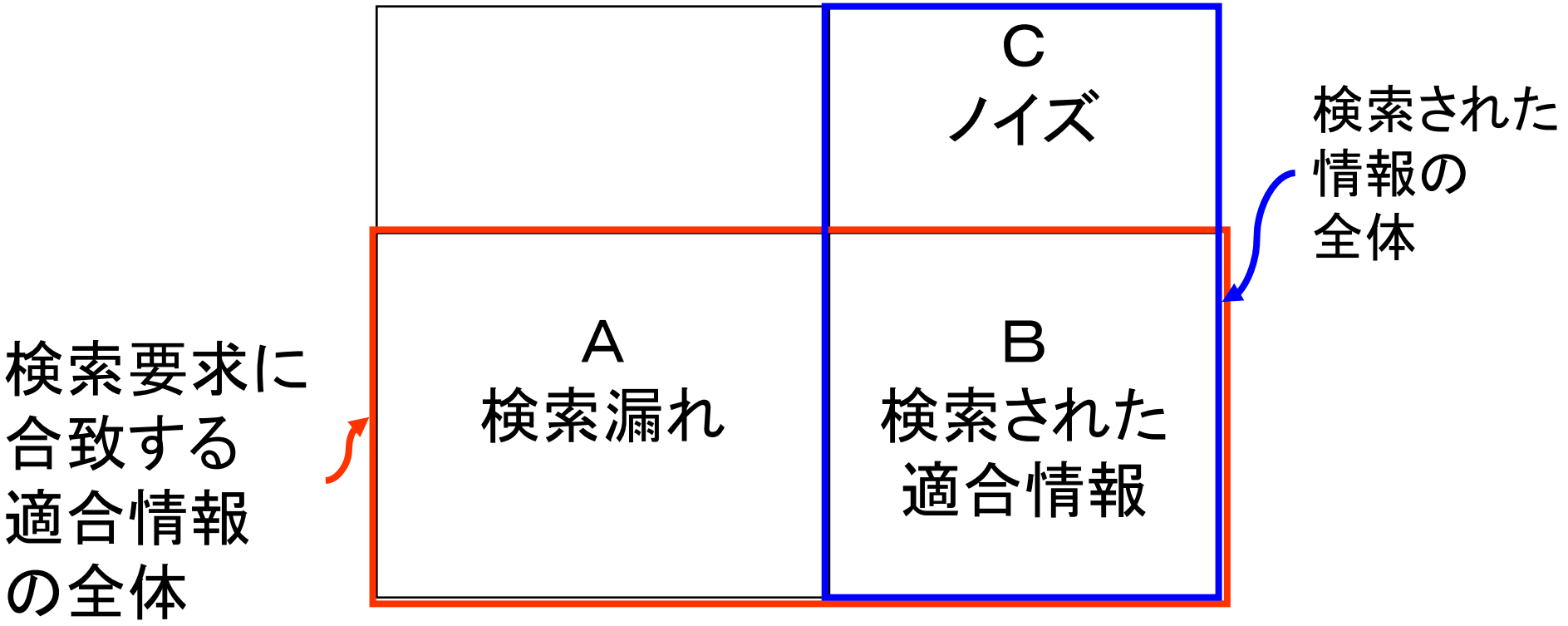
情報検索結果の評価(2) –p.23

- 再現率（recall ratio）
  - データベースに合致する適合情報のうち、どれだけ適合情報が検索されたかという割合を表す
  - 検索漏れの程度を表す指標となる
  - データベース全体の適合情報を実際に調べるのは不可能なため、普通、正確にはわからない
- 精度(precision ratio)（適合率(relevance ratio)）
  - 実際に得られた検索結果の情報全体のうち、どれだけ適合情報が検索されたかという割合を表す。
  - ノイズの程度を表す指標となる
  - 検索結果から容易に計算可能
- 両方とも高いのが理想だが、どちらかを上げるとどちらかが下がるという関係にある

情報検索結果の評価(3) –p.23

図1-10 –p.33

再現率　R　=　B ÷ (A + B) ×100%
精度　　P　=　B ÷ (C + B) ×100%

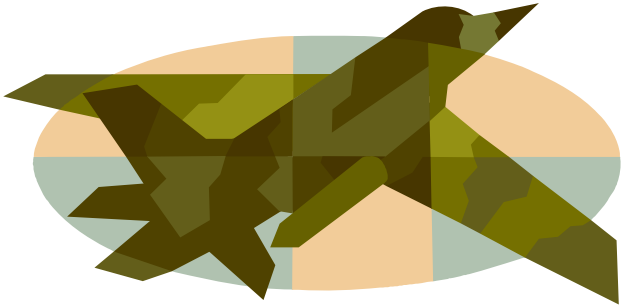


情報検索結果の評価(4) –p.23

- 例）10件の文献を出力、１０件全部が適合文献、データベース全体には100件の適合文書が存在
  - 精度：10/10×100=100%
  - 再現率：10/100×100=10%
    残りの９０%が検索漏れ
- 例）50件の文献を出力、30件が適合文献、データベース全体には70件適合文書が存在
  - 精度：30÷50×100 = 60%
    - ノイズは40%
  - 再現率：30÷70×100 = 43%
    - 検索漏れは57%

データベースの起源　–27p.

- 1950年代
- 米国国防総省が戦力に関する情報を保管、集中管理するためコンピュータを使ったライブラリーを開発
- データの基地(data base)から由来


データベースの定義(1)

- 著作権法二条十の三 –p.28
  - 論文、数値、図形その他の情報の集合物であって、それらの情報を電子計算機を用いて検索することができるように体系的に構成したもの
- 日本工業規格(JIS) –p.28
  - 適用業務分野で使用するデータの集まりであって、データの特性とそれに対応する実態の間の関係とを記述した概念的な構造によって編成されたもの(X0017)
  - 特定の規則に従って電子的な形式で、一か所に蓄積されたデータの集合であって、コンピュータでアクセス可能なもの(X0807)

データベースの定義(2) –28p.
--日本のデータベースの特徴--

- データベースとは”コンピュータを用いて検索できる”ことが重要である。情報が電子メディアに蓄積され、コンピュータ、携帯情報端末(PDA)、地上波テレビ端末などを使用して検索できる状態になっている。
- データや情報がコンピュータ処理できるように体系的に整理され、統合化・構造化されて蓄積・保存されており、必要な情報だけを部分的に取り出せる。
- 蓄積情報の検索や更新が容易に行えるよう、効率化を図ったものである

一方、ヨーロッパにおけるデータベースの定義では、
コンピュータを使用するかしないか、電子的であるかどうかについては特に限定していない

第1回レポート課題:第1,2章のまとめ

- 教科書の第1章、第2章(p.1～p.36)の内容をWordで**2ページピッタリ**にまとめなさい。
  - **10/21の授業開始時**に提出しなさい。
  - **講義名、第？回レポート課題、学籍番号、名前、提出日**を最初に記述すること
  - 2分割印刷で1枚に印刷して提出しなさい

第1回レポート課題：
この課題のねらい

- 情報検索の理論について理解する
  - 情報検索の定義(何の訳語？誰が言ったの？何年ごろ？など)
  - 情報検索の種類（3種類）
  - データベースの定義・・・どの法律？どんなもの？
  - データベースの分類
  - データベースのファイル構造
    - どんなファイルがあった？（大きくわけて2種類）
  - 情報検索の理論
    - 論理演算（論理積、論理和、論理差、図も描けるように）
    - トランケーション（前方一致、後方一致、中間一致、中間任意）
  - 検索結果の評価（再現率、精度）

CD-ROM版情報検索の演習

1. 検索プログラムのインストール：資料p.1～
   - 次回以降も毎回する必要がある
   - 个再起動するとアンインストールされてしまうため
2. 検索の大まかな流れ：資料p.8
3. 検索の基本
   1. ブラウズ
   2. 論理検索（検索フィールド内/間）
   3. トランケーション
   4. 範囲検索

人物略歴情報

- 人物・人材データベース「WHO」（日外アソシエーツ）から、現在活躍中の俳優・女優・歌手・タレントなど7,273人を収録
- 検索項目
  - **姓名**：芸名、本名、ヨミもある
    - 例：和田アキ子、ワダアキコ、飯塚現子、イイズカアキコ
  - **職業**：ヨミもある
    - 例：作曲家、サッキョクカ
  - **出身地**：都道府県しかない場合もある
    - 例：東京、東京・永田町、東京市、東京市王子区、東京都、東京都・お茶の水、東京都千代田区、東京都千代田区永田町
  - **生年月日**：YYYYMMDD(年4桁、月2桁、日2桁)
    - 例：19031216,19891115,19911208
  - **キーワード(件名)**：ヨミもある
    - 例：文化座、ブンカザ

第2回演習課題：CD-ROM検索の準備と検索のおおまかな流れ

1) 自分の生まれた年(_____年)と同じ年に生まれた人物を検索しなさい。

何人いましたか(何件ヒットしましたか)？：_____人

2) 5 番目のレコードの人物の生年月日を答えなさい：_____

3) 2)で答えた生年月日を YYYYMMDD の形式で答えなさい：_____

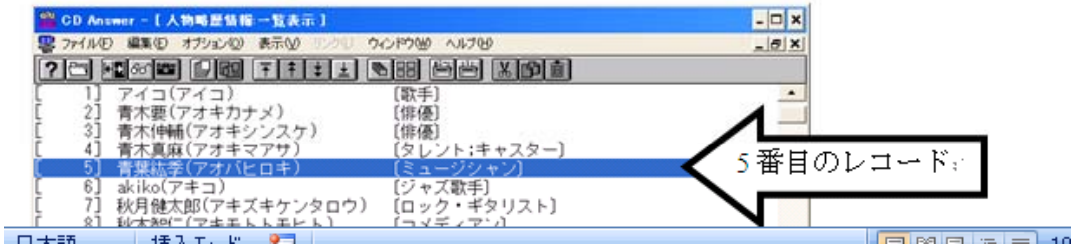
4) 自分の同じ都道府県(_____)出身の人物を検索しなさい。

ヒット件数：_____

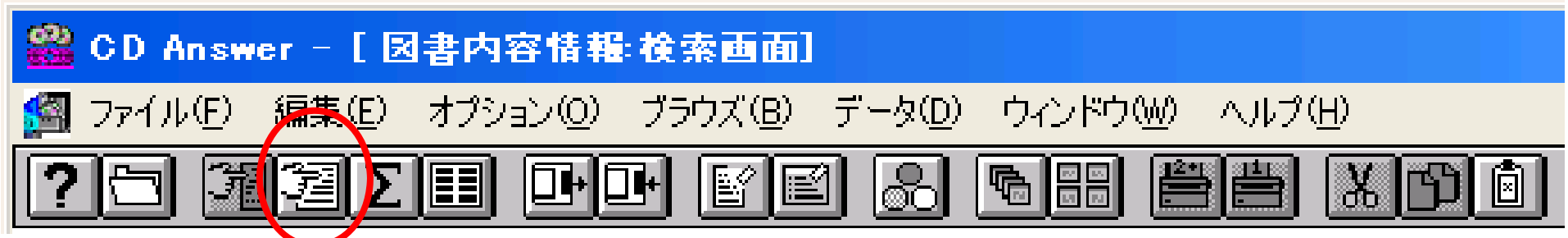
5) 3 番目のレコードの出生地(出身地など)と生年月日を答えなさい。

出生地：_____　　生年月日：_____

補足説明：5 番目のレコード



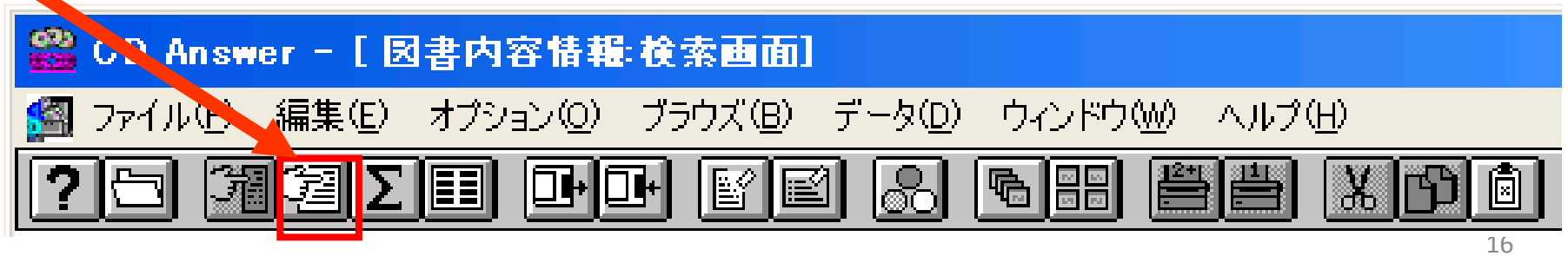
CD-ROM検索基本機能
--ブラウズ機能--



- ブラウズ機能
  - 登録されている検索語の一覧を見る
- 注:日本語入力モードにしてからブラウズボタンを押すこと

実演・演習：ブラウズ機能

- 最初に、必ずブラウズ機能を最初に使ってどんな検索語があるか確かめてみるようにしましょう！
- 演習：すべての検索項目でブラウズ機能を使ってみて、ざっと眺めてみましょう
- 演習：「人物略歴情報」で「職業」が「司会」をブラウズ機能で調べてみましょう



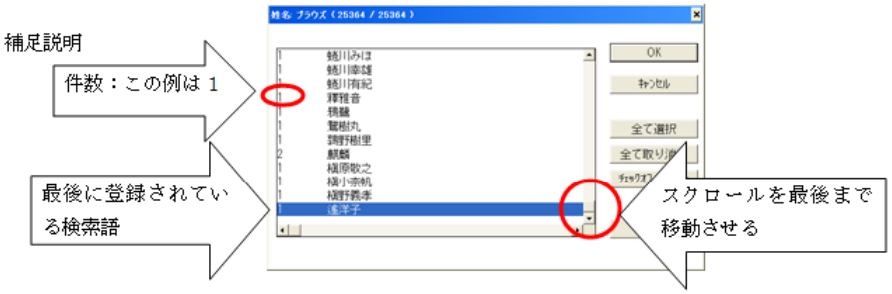
第3回演習課題

--CD-ROM検索・ブラウズ機能--

1. 人物略歴情報：ブラウズ機能を使って最後に登録されている検索語と件数を調べ空欄を埋めなさい

検索項目	件数	最後に登録されている検索語
姓名		
職業		
出生地		
生年月日		
キーワード(件名)		

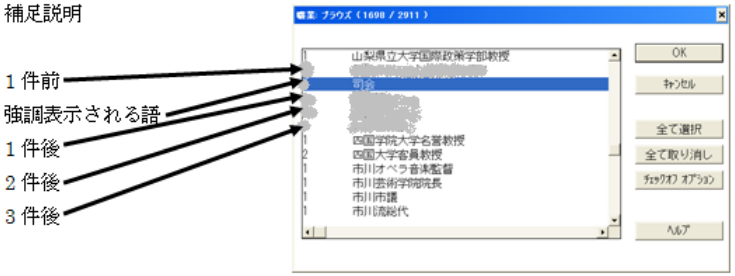
補足説明



「職業」でブラウズ機能を使って「司会」と入力して、以下の空欄を埋めなさい。

検索語の順序	件数	検索語
1 件前		
強調表示される語	5	司会
1 件後		
2 件後		
3 件後		

補足説明



本日のまとめ

- 課題の提出
- 情報検索の評価
- データベースとは（起源、定義）
- 第1回レポート課題出題
- CD-ROM版情報検索の演習
- CD-ROM検索の基本
  - 第2回演習課題：CD-ROM検索の準備と検索のおおまかな流れ
  - 第3回演習課題：CD-ROMブラウズ