

# 情報検索演習

## 第2回

2007年10月3日

後期 水曜5限

江草由佳

国立教育政策研究所

yuka@nier.go.jp

講義資料:

<http://momiji.mimoza.jp/lecture/2007/ir/>

# 今日のおしながき

- USBメモリの使い方
- 講義内容
  - 情報検索の流れ
  - レコードと検索フィールド
  - 情報検索の理論
    - 論理演算子(AND,OR,NOT)
    - トランケーション(前方一致、後方一致、中間任意、中間一致)
  - 情報検索結果の評価
    - 検索漏れ、ノイズ
    - 再現率
    - 精度
- 課題1:論理演算とトランケーション

# お知らせ

- 教科書
  - 原田智子編著「三訂情報検索演習」樹村房  
2006年10月 ¥1,995
  - 必ず持ってきてください
- 提出課題（演習課題とレポート課題）について
  - 提出課題は必ず**提出すること**
  - **締切厳守**。締切を**過ぎて提出したものは大幅減点**
  - 欠席して提出できなかったもの→次回に提出する
    - 当然、**減点**はします。

# 演習：講義資料の保存と利用 (USBメモリの使い方)

- 講義資料をWebから取得しUSBに保存
  - やりかたは、「講義・演習 Q and A」の「講義資料等をUSBメモリに保存する方法」を参照
  - <http://momiji.mimoza.jp/lecture/2007/QA/#download-usb>
- USBの基本的な使い方は
  - 「講義・演習 Q and A」の「USBメモリの使い方」を参照
  - <http://momiji.mimoza.jp/lecture/2007/QA/#usb>

# 予告：第1回レポート課題

- 次回に以下のようなレポート課題を出します。
- 第3回までの講義の内容をA4用紙1ページ分にまとめなさい
  - ヒント：スライドの「今日のまとめ」、目次、章タイトル
- まとめるとしたらどういう風にまとめられるだろう？と意識しながら講義を受けましょう。

## 2.情報検索の流れ

### (1)情報検索の受付と検索準備 -p.12

1)検索の受付 -p.12

2)インタビュー -p.15

3)検索テーマの主題分析 -p.15

4)検索対象の決定 -p.15

5)検索語の決定 -p.16

6)検索式の作成 -p.16

図1-6:情報検索の流れ -p.13

図1-7:情報検索申込書 -p.14

### (2)検索の実行 -p.17

### (3)検索結果の整理と情報提供 -p.17

### (4)検索結果の保存と管理 -p.17

p.? は  
テキストのページ数  
を表す

# レコードと検索フィールド(1) –18p.

- レコード

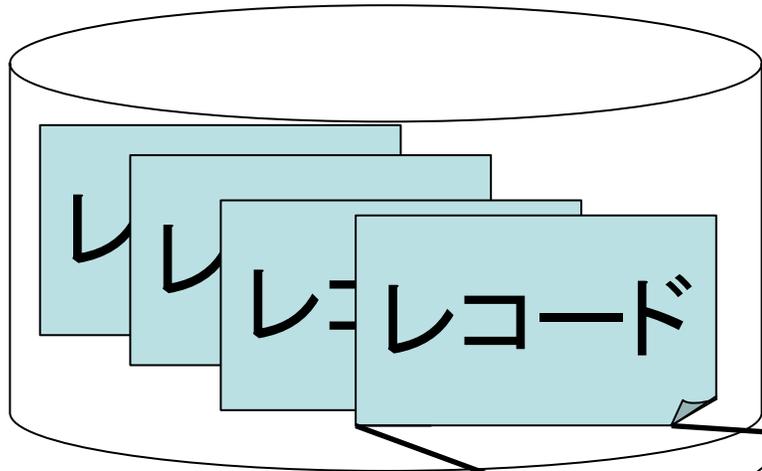
- データベースに収録されている個々の論文や新聞記事などの情報単位
- レコードの内容はデータベースの種類や内容によって異なる

- 検索フィールド

- レコードは複数の検索フィールドからなる
- 例)レコード番号、論題、著者名

p.? は  
テキストのページ数  
を表す

# レコードと検索フィールド(2)



検索フィールド名

検索フィールド値

検索フィールド

論題:	Reading—速読・多読 について考える
著者名:	清水由理子
請求記号:	P343-5C2-14
掲載誌名:	獨協大学外国語教育研究14
発行年月:	1995.12
掲載ページ:	p.273~282
登録日:	19970930

# 情報検索の理論(1) —p.19

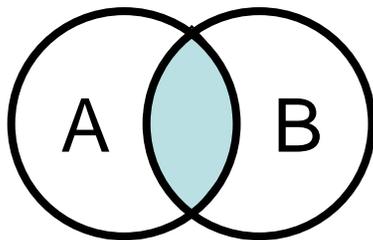
- コンピュータ検索では論理演算の概念が基本
  1. データベース全体から合致するものを検索し
  2. 論理積、論理和、論理差の集合の概念をもちいて、広げたり、狭めたりして検索
- 論理演算
- トランケーション

# 情報検索の理論(1) —p.19

## (1)論理演算子 —p.19

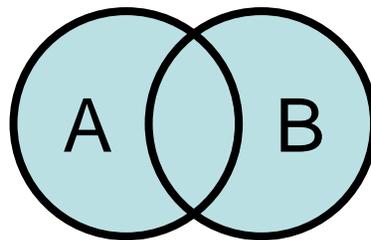
- 論理演算
  - 論理積 (AND)、論理和 (OR)、論理差 (NOT)
- 論理演算子
  - 説明ではAND, OR, NOTを使うが、演算子の書き方はシステムによって様々である。

論理積



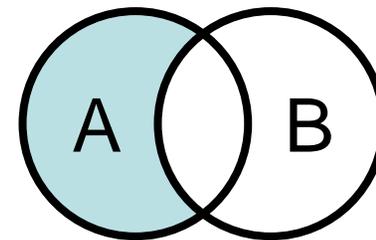
A AND B

論理和



A OR B

論理差



A NOT B

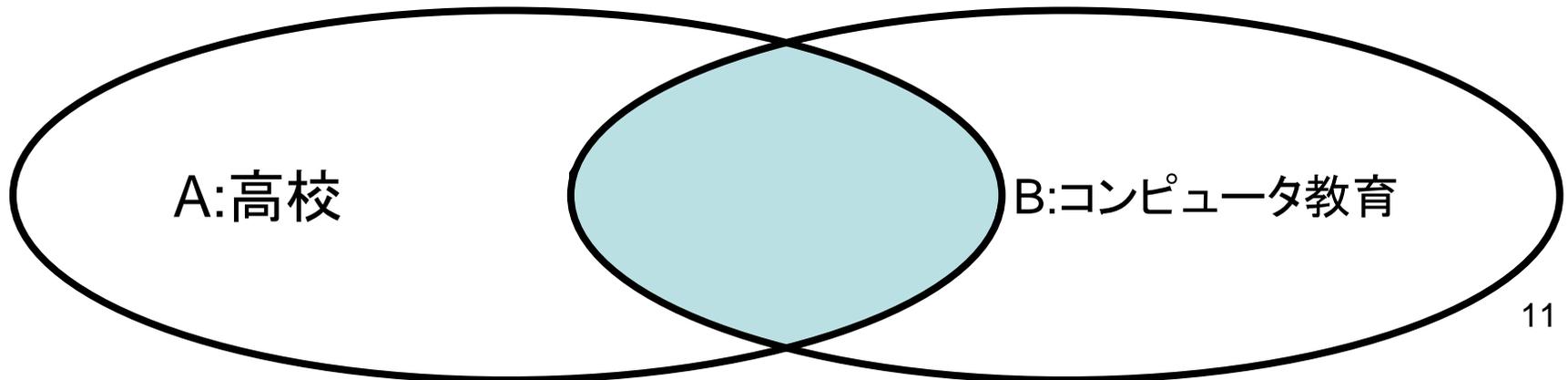
# 情報検索の理論(2) —p.19

## (1)論理演算子 —p.19

p.? は  
テキストのページ数  
を表す

- 論理積 (AND) —p.20

- Aという検索語を持つ情報の集合とBという検索語をもつ集合の両方を含む部分を検索すること
- 情報を絞り込んでいくときに有用
- 例:「高校におけるコンピュータ教育」
- 検索式:”高校 AND コンピュータ教育”



# 論理積の例：教育 and 情報

「教育」と「情報」の両方を含む

ヒット 高等学校「情報」教員養成におけるメディア・リテラシー教育

ヒット 情報教育に関わる問題—高等学校の情報関係科目の履修の経験による検討

3. 高校生のコンピュータに対する意識調査

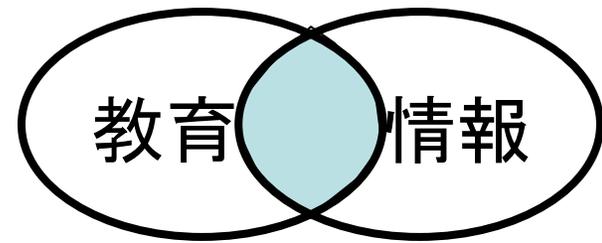
4. 高等学校教員免許状取得の現状と課題—教科「情報」の免許状の取得

ヒット 高校普通教科「情報」とスキル教育

6. 高校生の情報活用に関する日中比較

7. 高校教育の多様化-高等学校現場からの報告

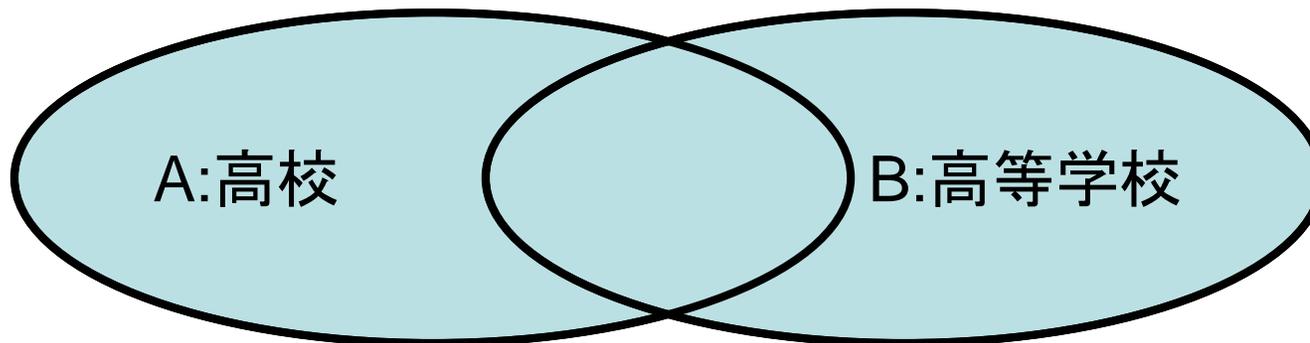
8. デジタルメディア利用教授不安の減少と高校生のコンピュータ不安



# 情報検索の理論(3) —p.19

## (1)論理演算子 —p.19

- 論理和 (OR) —p.20
  - Aという検索語を持つ情報の集合とBという検索語のいずれか一方の検索語をもつ集合部分と、両方をもつ集合部分全てを検索すること
  - 同義語などの検索に有用
  - 例:「高校におけるコンピュータ教育」
  - 検索式:”高校 OR 高等学校”

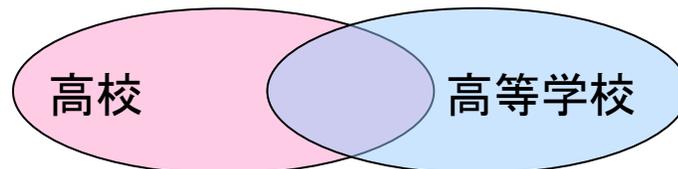


p.? は  
テキストのページ数  
を表す

# 論理和の例：高校 or 高等学校

「高校」か「高等学校」のどちらかを含む

- ヒット 高等学校「情報」教員養成におけるメディア・リテラシー教育
- ヒット 情報教育に関わる問題—高等学校の情報関係科目の履修の経験による検討
- ヒット 高校生のコンピュータに対する意識調査
- ヒット 高等学校教員免許状取得の現状と課題—教科「情報」の免許状の取得
- ヒット 高校普通教科「情報」とスキル教育
- ヒット 高校生の情報活用に関する日中比較
- ヒット 高校教育の多様化-高等学校現場からの報告
- ヒット デジタルメディア利用教授不安の減少と高校生のコンピュータ不安



# 情報検索の理論(4) —p.20

## (1)論理演算子 —p.20

p.? は  
テキストのページ数  
を表す

- 論理差 (NOT) —p.20
  - Aという検索語を持つ情報の集合からBという検索語をもつ集合部分を除いた部分を検索すること
  - 複数の言語で書かれていることはすくないので、ロシア語文献を抜くといった検索に有効
  - 例:「トラック事故以外の自動車事故」
  - 検索式:”自動車事故 NOTトラック事故”

A:自動車事故

B:トラック事故

# 論理差の例：自動車 not トラック

「自動車」を含みかつ「トラック」を含まない

ヒット

アジアの自動車産業

ヒット

新しい自動車の構造と運転取扱法

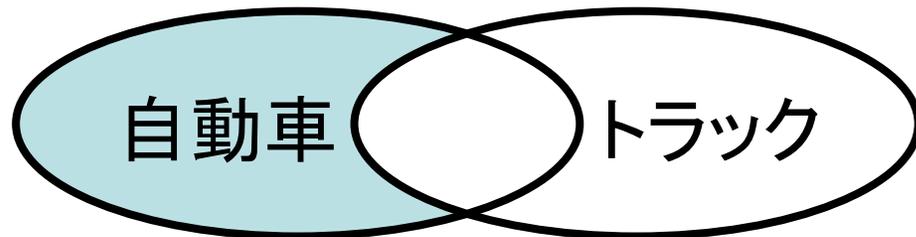
3. 自動3輪トラックの構造と運転

ヒット

アメリカの自動車会社ビッグ3の復活

5. はたらく自動車：トラック・工事用車両

6. 大型トラック・トレーラの安全対策の研究



# 情報検索の理論(8) —p.20

## (3) トランケーション —p.21

### • トランケーション

– 検索語を入力する場合に、語の一部を任意文字に指定して検索すること

– マスク文字(ワイルドカード)

- 任意文字とする部分に使用する入力文字

- マスク(mask): 覆い隠すから由来



– マスク文字: システムによって異なる

以下は説明で使用

- ? : 0文字以上、何文字でもよい

- # : 0文字または一文字

- ! : ちょうど一文字

p.? は  
テキストのページ数  
を表す

# 情報検索の理論(9) —p.20

## (3) トランケーション —p.21

### 1) 前方一致検索 —p.22

- 検索語の**前方**が**一致**する検索
- 検索語の**末尾**を**任意文字**に指定する検索
- 大抵どのシステムにもある

#### • 例)

- 情報**?** ⇒ 情報、情報**検索**、情報**検索システム**
- CAT**#** ⇒ CAT, CAT**S**, CAT**V**
  - 猫を検索したいときによいが、CATVまで検索されてしまう
- DIS**!** ⇒ DIS**K**, DIS**C**
  - 英米綴りの違いを同時に検索できる

p.? は  
テキストのページ数  
を表す

# 情報検索の理論(10) —p.20

## (3)トランケーション —p.21

### 2)後方一致検索 —p.22

- 検索語の**後方**を**一致**させる検索
- 検索語の**始まり**を**任意文字**に指定する方検索

#### ● 例)

- **?**情報 ⇒ 情報、**安全**情報、**特許**情報

p.? は  
テキストのページ数  
を表す

# 情報検索の理論(11) —p.20

## (3) トランケーション —p.21

### 3) 中間任意検索 —p.23

– 検索語の途中を任意文字に指定する検索

#### • 例)

– 情報 ? システム ⇒ 情報システム、情報管理システム、情報検索システム

– ログ!ン ⇒ ログイン、ログオン

- 同義語が同時に検索できる

– WOM!N ⇒ WOMAN, WOMEN

- 単数形、複数形が同時に検索できる

– GR!Y ⇒ GREY, GRAY

- 英米綴りの違いを同時に検索できる

p.? は  
テキストのページ数  
を表す

# 情報検索の理論(12) —p.20

## (3) トランケーション —p.21

### 4) 中間一致検索 —p.23

- 検索語の**中間**が**一致**する検索
- 検索語の**両端**を**任意文字**に指定する検索
- インターネットの検索エンジンでは中間一致していることが多い
- 一般に3文字以下の略字ではノイズを招くので、トランケーションを使わず、完全一致させたほうがよい

#### ● 例)

- **?情報?** ⇒ 情報、**交通**情報、情報**システム**、**交通**情報**システム**

p.? は  
テキストのページ数  
を表す

# 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  
テキストのページ数  
を表す

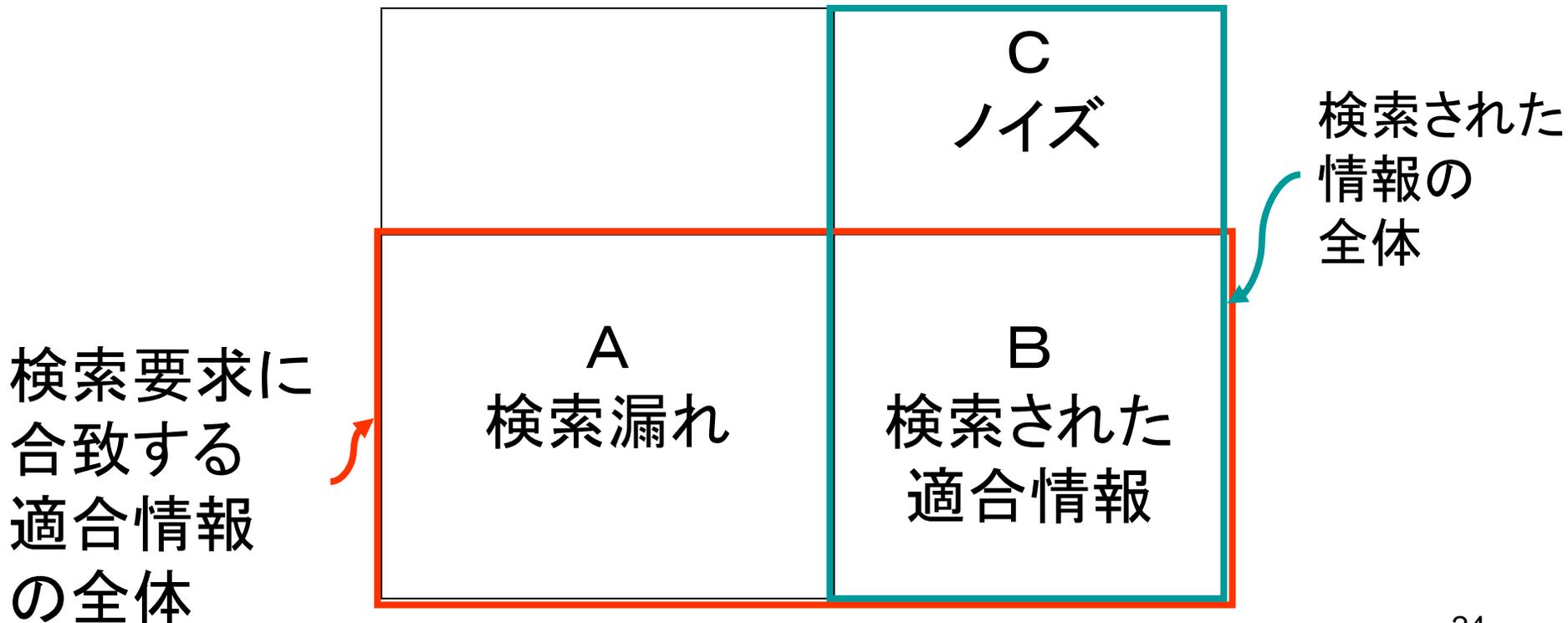
# 情報検索結果の評価(2) –p.23

- 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

再現率  $R = B \div (A + B) \times 100\%$   
精度  $P = B \div (C + B) \times 100\%$

図1-10 -p.33



# 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:  $10/10 \times 100 = 100\%$
  - 再現率:  $10/100 \times 100 = 10\%$   
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:  $30 \div 50 \times 100 = 60\%$ 
    - ノイズは40%
  - 再現率:  $30 \div 70 \times 100 = 43\%$ 
    - 検索漏れは57%

# 課題1:論理演算とトランケーション

- 課題1を行い、印刷して提出しなさい
  - 課題1は講義Webサイトにあります。
  - 保存方法は、「[講義資料等をUSBメモリに保存する方法](#)」を参照
  - 2ページにわたらないように気をつけること
  - 学籍番号、名前、提出日を忘れないこと
  - 課題1はUSBに保存すること
- この課題のねらい:
  - 論理演算やトランケーション機能を使った検索式を組み立てることができるようになる。
  - 論理演算子の書き方やマスク文字が、システムによって異なることを体験する。

# 今日のまとめ

- 講義内容
  - 情報検索の流れ
  - レコードと検索フィールド
  - 情報検索の理論
    - 論理演算子 (AND, OR, NOT)
    - トランケーション (前方一致、後方一致、中間任意、中間一致)
  - 情報検索結果の評価
    - 検索漏れ、ノイズ
    - 再現率
    - 精度