情報検索演習
第2回

教科書が発行されました
まだ買ってない人は購入すること

原田智子編著「三訂情報検索演習」樹村房
2006年10月　￥1,995

2006年10月11日
後期 水曜4/5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/2006/ir/

今日のおしながき

- お知らせ
- 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

お知らせ

- 来週(10/18)は休講です
- 教科書が発行されました
  - 原田智子編著「三訂情報検索演習」樹村房　2006年10月　￥1,995
  - 次回からは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
  - 「演習課題」: 授業中に作成してその場で提出する課題
  - 「レポート課題」: 授業時間外に作成して、授業開始時に提出する課題
  - 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、「第1回講義補足資料」を参照
  - http://localhost/~yuka/lecture/2006/ir/IR01-20060927-add.ppt
  - 配布資料にもあります
  - 使うときは
    - 「スタート」→「マイコンピュータ」→「リムーバルディスク」をダブルクリック

資料訂正と補足資料

- 授業Webサイト
  - 訂正済み第1回資料
  - 補足資料
- スライド4
  - 渡辺満彦　→　原田智子
- スライド8
  - 授業のホームページ　→　授業のWebサイト
  - 「yuka@nier.go.jp」を追加
- スライド29
  - スライド29 第一回課題提出　を追加

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

(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

図1-6:情報検索の流れ –p.13

図1-7:情報検索申込書 –p.14

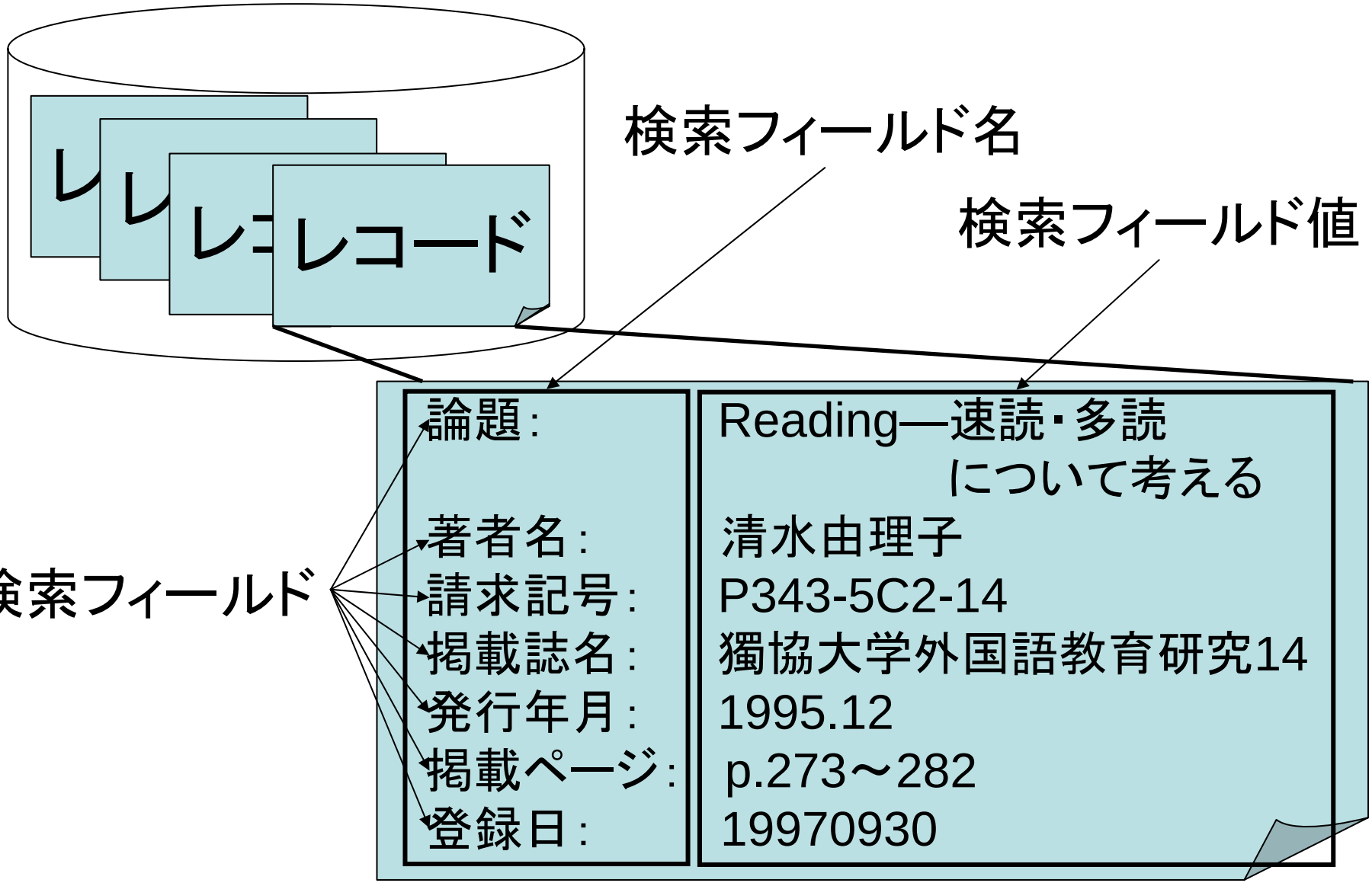
p.? は
テキストのページ数
を表す

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)

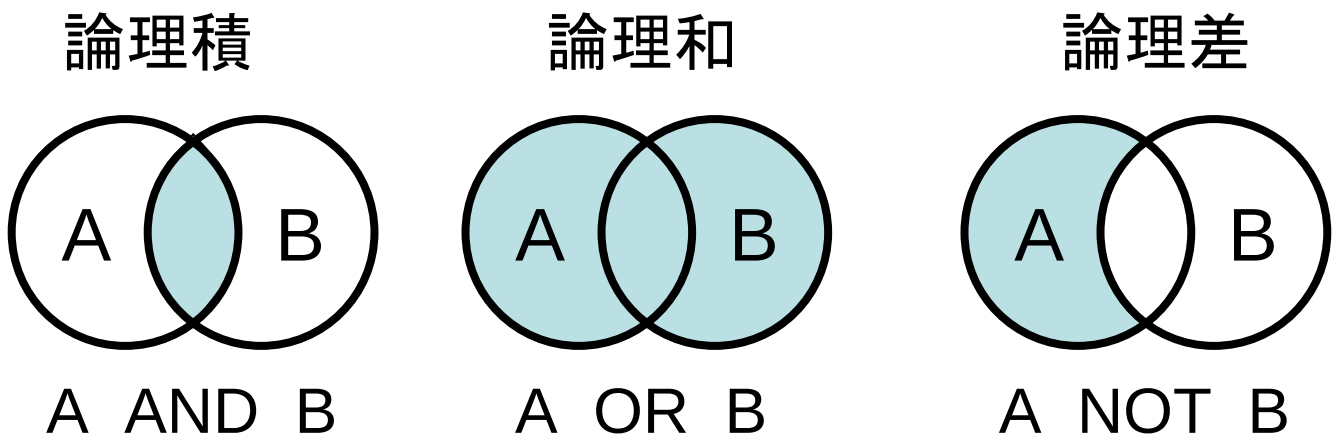


情報検索の理論(1) —p.19

- 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。

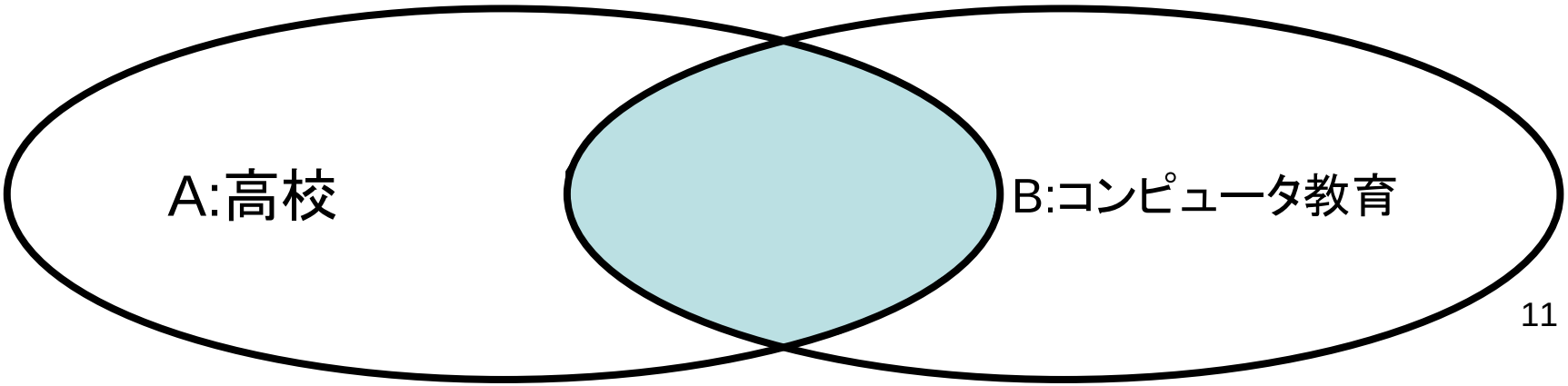


1－9図 p.19

情報検索の理論(2) —p.19
(1)論理演算子 –p.19

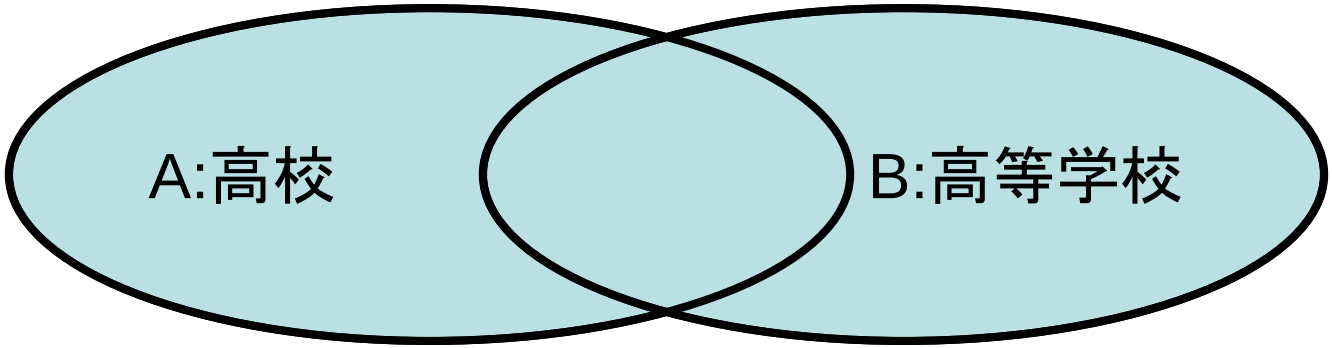
p.? は
テキストのページ数
を表す

- 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”



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”

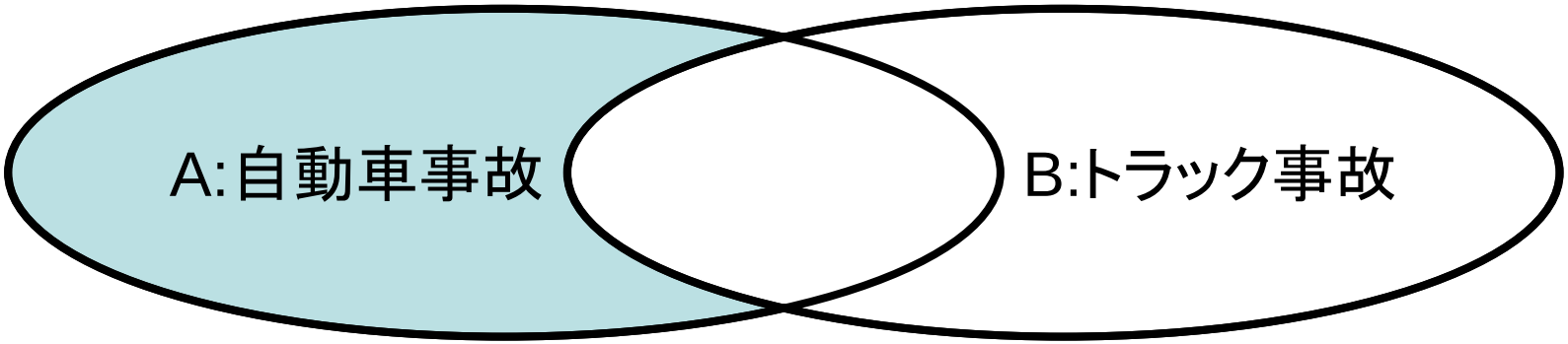


p.? は
テキストのページ数
を表す

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”



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,CATS,CATV
    - 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! ⇒ DISK,DISC
    - 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

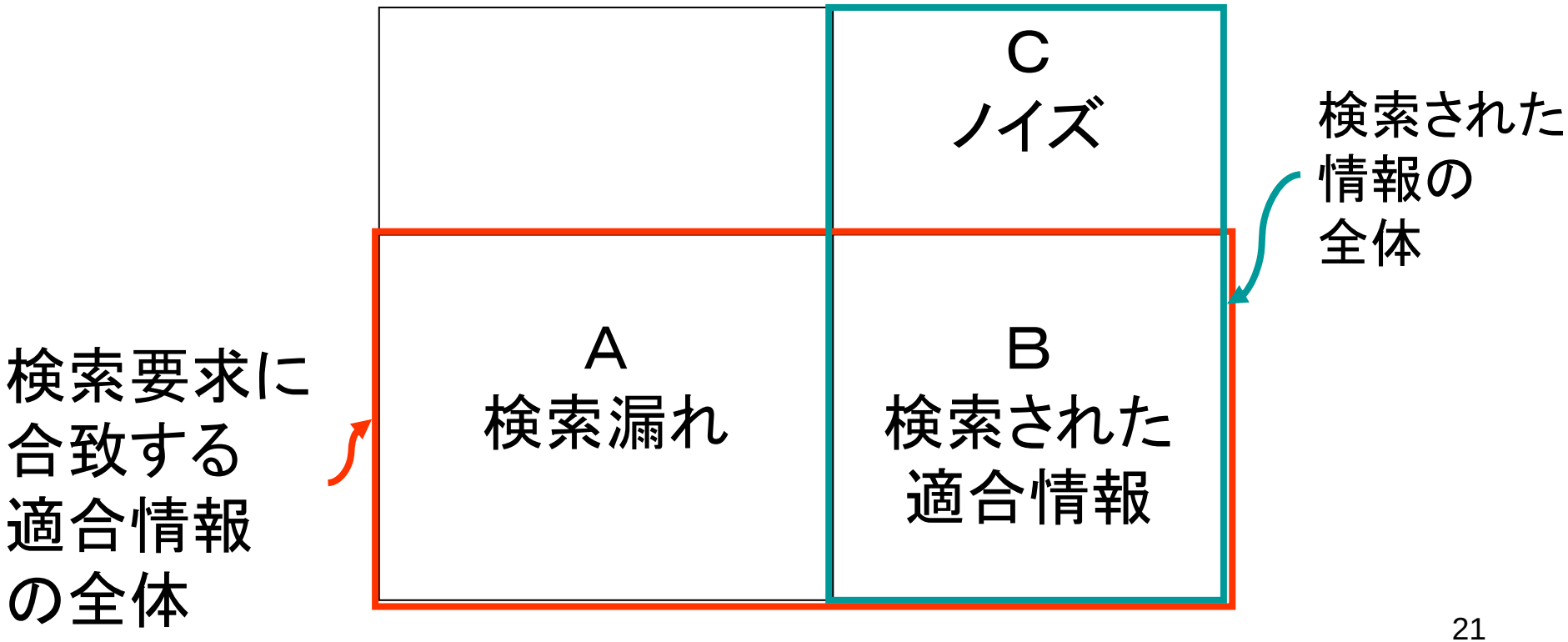
情報検索結果の評価(2) –p.23

- 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

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%

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

第1回レポート課題

- 今日までの講義をA4用紙1ページ分にまとめなさい
  - 〆切：次回の講義にUSBメモリで持参すること
  - Microsoft Office Wordで作成すること
  - ファイル名：report01-学籍番号名前.doc
  - ヒント：スライドの「今日のまとめ」、目次、章タイトル
- 以下の項目をレポートの冒頭につけること
  - レポートのタイトル：第1回レポート課題
  - 授業名：情報検索演習
  - 時限：4限　or　5限
  - 提出した日付
  - 学籍番号
  - 氏名

第2回演習課題
（レポート提出の練習）

- 今までの講義についての感想もしくは質問をなんでもよいから記述したWordファイルを作成しUSBメモリに保存しなさい
- 電子的なファイルとして提出しなさい
- ファイル名：ir2006-10-11-学籍番号名前.doc
- 以下の項目を演習課題の冒頭につけること
  - 演習課題のタイトル：第2回演習課題
  - 授業名：情報検索演習
  - 時限：4限　or　5限
  - 提出した日付
  - 学籍番号
  - 氏名