情報アクセス論 第7回
「情報検索システムの性能評価」

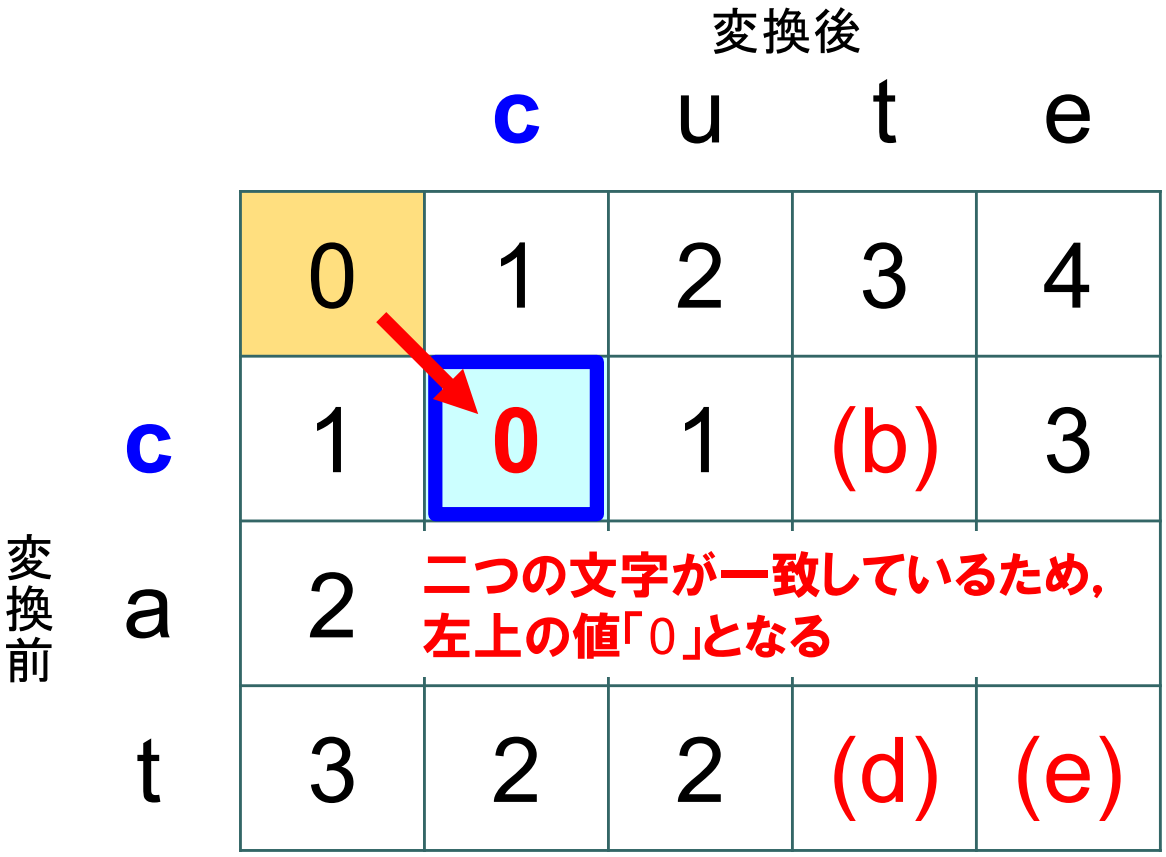
情報理工学部
前田 亮

前回の小テストの解説(1/6)

- 「cat」と「cute」の編集距離を求める以下の表の空欄に当てはまる値を選択肢から選べ

		変換後				
			c	u	t	e
変換前		0	1	2	3	4
	c	1	(a)	1	(b)	3
	a	2	(c)	1	2	3
	t	3	2	2	(d)	(e)

前回の小テストの解説(2/6)



文字列aのi文字目と
文字列bのj文字目が
一致している場合は
$C[i,j]=C[i-1,j-1]$

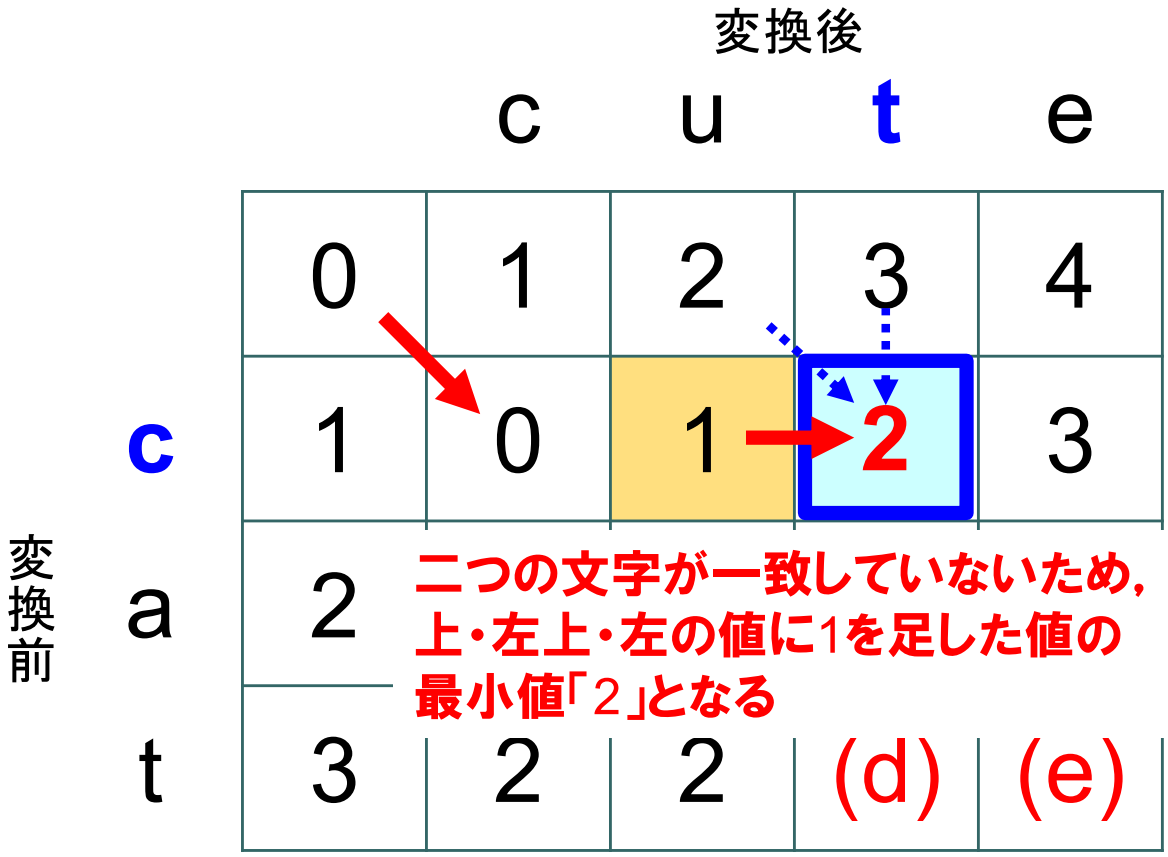
そうでない場合は以
下のうちの最小値:

① $C[i-1,j]+1$

② $C[i,j-1]+1$

③ $C[i-1,j-1]+1$

前回の小テストの解説(3/6)



文字列aのi文字目と
文字列bのj文字目が
一致している場合は
$C[i,j]=C[i-1,j-1]$

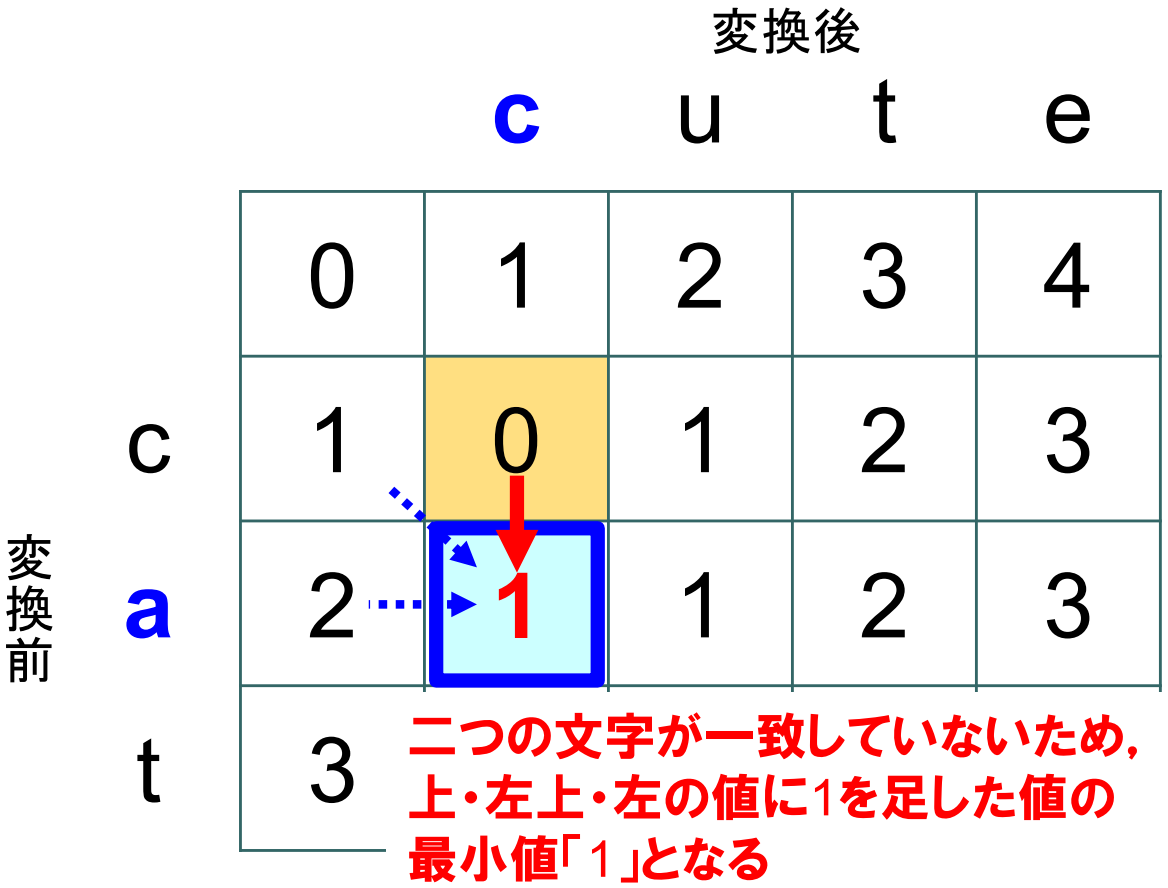
そうでない場合は以
下のうちの最小値:

① $C[i-1,j]+1$

② $C[i,j-1]+1$

③ $C[i-1,j-1]+1$

前回の小テストの解説(4/6)



文字列aのi文字目と
文字列bのj文字目が
一致している場合は
$C[i,j]=C[i-1,j-1]$

そうでない場合は以
下のうちの最小値:

① $C[i-1,j]+1$

② $C[i,j-1]+1$

③ $C[i-1,j-1]+1$

前回の小テストの解説(5/6)

変換前 \ 変換後		c	u	t	e
	0	1	2	3	4
c	1	0	1	2	3
a	2	1	1	2	3
t	3	2	2	1	(e)

二つの文字が一致しているため，左上の値「1」となる

文字列aのi文字目と
文字列bのj文字目が
一致している場合は
$C[i,j]=C[i-1,j-1]$

そうでない場合は以下のうちの最小値:

① $C[i-1,j]+1$

② $C[i,j-1]+1$

③ $C[i-1,j-1]+1$

前回の小テストの解説(6/6)

文字列aのi文字目と
文字列bのj文字目が
一致している場合は
$C[i,j]=C[i-1,j-1]$

そうでない場合は以
下のうちの最小値:

① $C[i-1,j]+1$

② $C[i,j-1]+1$

③ $C[i-1,j-1]+1$

性能評価（ベンチマーク）とは?

- 情報システムのハードウェアやソフトウェアの性能をさまざまな観点から客観的に評価
  - **ハードウェア**：CPU，メモリ，ハードディスク，ネットワーク，グラフィックボード，etc.
  - **ソフトウェア**：トランザクション速度，etc.
  - **プログラミング**：実行時間，メモリ消費量，ディスクアクセス頻度，etc.
- システム間の性能（主に効率）を比較する
- 客観的，定量的に比較できるものが対象

情報検索システムの性能評価の観点

- **効率性**（efficiency）
  - 問合せを入力してから検索結果が出るまでどれだけ時間がかかるか（応答時間）
- **ユーザインタフェース**
  - システムの機能を利用者が簡単に使いこなせるか
- **情報の網羅性**
  - 利用者が必要とする情報がシステムの文書集合に存在するか
- **有効性**（effectiveness）
  - 必要な情報をどれだけ**漏れなく正確に**検索できるか

なぜ**有効性**が重要か?

- データベース管理システム(DBMS)との違い
  - DBMSでは，同じデータに対する同じSQLの処理結果は，どのシステムを用いても同一である
  - 情報検索システムでは，同じ文書集合に対する同じ問合せでも，検索結果はシステムによって必ずしも同一ではない
- **効率性**も重要だが，計算機の性能に依存
- **ユーザインタフェース**は客観的な評価が困難
- **情報の網羅性**は文書集合の問題であり，システム自体とは無関係

有効性の評価尺度(1)

- **適合性**（relevance）
  - **問合せ**に対して，**客観的に**見て適合する文書を検索できたか?
- **適切性**（pertinence）
  - **情報要求**に対して，**利用者が**求めている文書を検索できたか?
- **有用性**（usefulness）
  - (情報要求にかかわらず)**利用者に役立つ**文書を検索できたか?

有効性の評価尺度(2)

情報要求

- **例**：**新型コロナウイルス**について調べるために，検索エンジンに「**ウイルス**」と入力して検索

問合せ

  - 「コンピュータウイルス」に関する文書は，この利用者にとってはゴミなので，**適切性**(*情報要求を満たす*)という点では✕
  - 「ウイルス」という問合せに対して「コンピュータウイルス」に関する文書が検索されることは，客観的に見て誤りではないので，**適合性**(*客観的に正しい*)は○
  - 検索された「コンピュータウイルス」に関する文書中に，たまたま利用者のコンピュータの感染に関する情報があれば，**有用性**(*利用者に役立つ*)は○

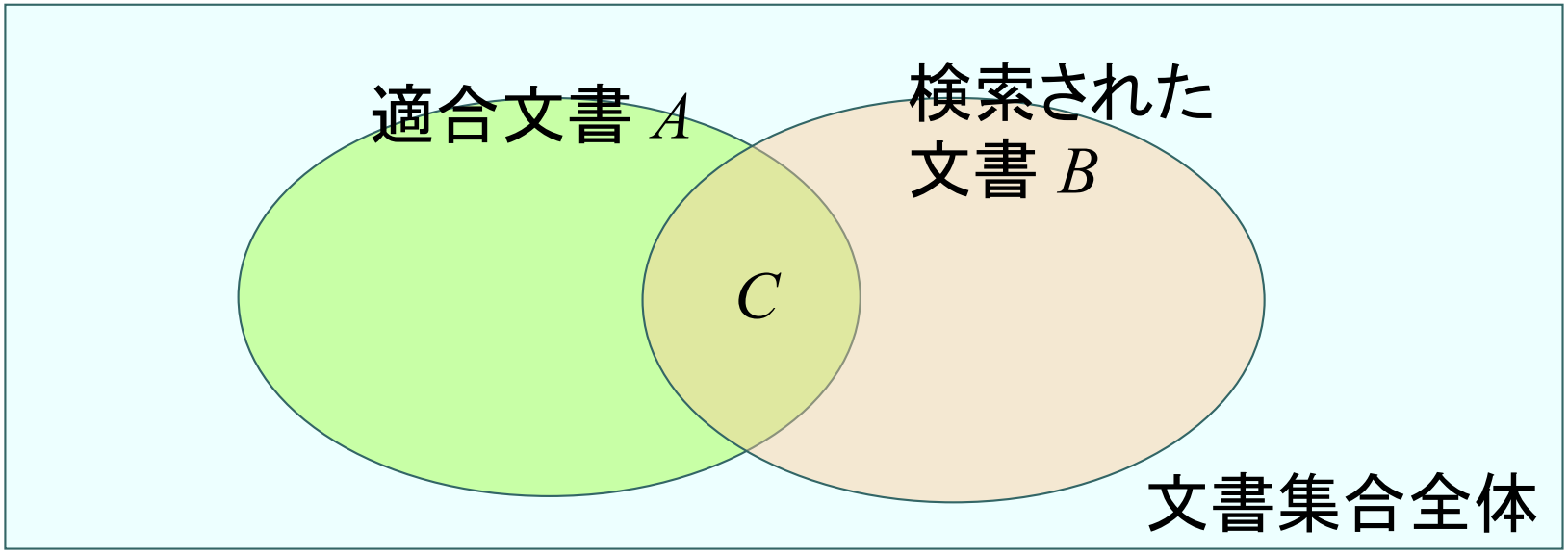
有効性の評価尺度(3)

- **適切性**や**有用性**の客観的な定義は難しい
  - 利用者個人の知識に関わるため
    - *利用者が何について調べたかったのか?*
    - *検索結果が利用者の役に立ったかどうか?*
- 実際に情報検索システムの評価に通常用いられるのは**適合性**
- ではどうやって適合性を客観的に評価するのか?

適合性の尺度：**再現率と適合率**

- **再現率**（recall）
  - 問合せに適合する文書をどれだけ検索できたか
  - 「検索漏れ」の少なさ
- **適合率**（precision）
  - 検索結果のうち，どれだけが問合せに適合しているか
  - 「検索ゴミ」の少なさ
  - 適合率は「**精度**」とも呼ばれる

再現率と適合率の式



$$\text{再現率} = \frac{\text{検索された文書中の適合文書の数}}{\text{文書集合中の適合文書の数}} = \frac{|C|}{|A|}$$

$$\text{適合率} = \frac{\text{検索された文書中の適合文書の数}}{\text{検索された文書の数}} = \frac{|C|}{|B|}$$

再現率・適合率の計算の例（ランキングなし）

- ある問合せに対する適合文書が，文書集合全体で20件であるとする
- この問合せを用いて検索した結果，10件の検索結果が得られ，うち5件が適合文書であった

$$\text{再現率}=\frac{\text{検索された文書中の適合文書の数}}{\text{文書集合中の適合文書の数}}=\frac{5}{20}=0.25$$

$$\text{適合率}=\frac{\text{検索された文書中の適合文書の数}}{\text{検索された文書の数}}=\frac{5}{10}=0.5$$

再現率と適合率の関係(1)

- 再現率・適合率は**トレードオフ**の関係にある
  - 再現率を上げるには，できるだけ多くの文書を検索結果として返せばよい
    - どんな問合せに対しても文書集合全体を返せば，再現率は常に100%
  - 適合率を上げるには，適合すると思われる文書一件だけ返せばよい?
    - その文書が適合していれば適合率100%
      - Googleの「I'm Feeling Lucky」
    - 他にも適合文書があれば再現率は低くなる

再現率と適合率の関係(2)

- 再現率・適合率のどちらが重要かは検索の目的によって異なる
  - **ある事実について知りたい**場合は，**適合率**が重要
    - 答えを含むページが一つでも見つかれば良い
    - 「立命館大学が創立したのはいつ?」
  - **網羅的に検索したい**場合は，**再現率**が重要
    - 特許出願の際に，過去の類似特許を網羅的に検索
    - ある研究を始める際に，過去の関連研究をサーベイ
- 通常は，どちらの目的にも対応する必要がある
  - 再現率・適合率のバランスが重要

再現率－適合率曲線

- ブーリアンモデルでは，再現率・適合率はそれぞれ一つの値（0～100%）
  - 検索結果がランキングされていない単なる集合のため
- ランキング可能な検索モデルでは，再現率・適合率は上位何件まで見るかによって変化
  - 各再現率レベルにおける適合率を求め，グラフで表すことができる

再現率・適合率の計算の例（ランキングあり）

- ある問合せに対する検索結果（全10件）
- 文書集合全体のうち適合文書数は5件

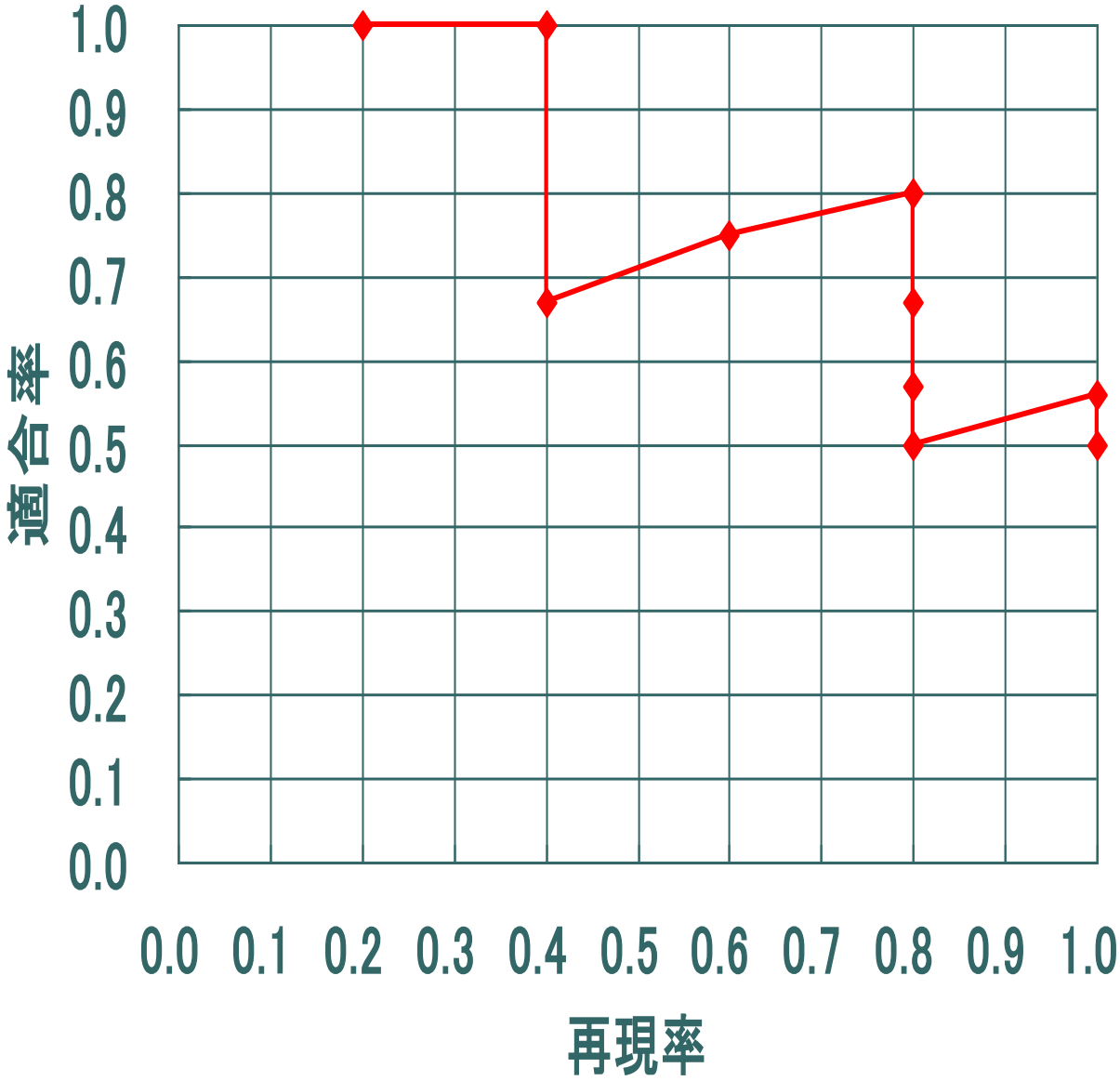
ランキングを考慮しなければ…

- 再現率：5/5=1.00
- 適合率：5/10=0.50

順位	適合性	再現率	適合率
1	○	0.20	1.00
2	○	0.40	1.00
3	×	0.40	0.67
4	○	0.60	0.75
5	○	0.80	0.80
6	×	0.80	0.67
7	×	0.80	0.57
8	×	0.80	0.50
9	○	1.00	0.56
10	×	1.00	0.50

再現率－適合率曲線(1)

- 適合率を再現率の関数とみなし，2次元座標上にプロット
- 再現率をx軸
  適合率をy軸
- 右図のように，再現率の増加に対して適合率が上下し，階段状になる
- システム間での比較が難しい



再現率－適合率曲線(2)

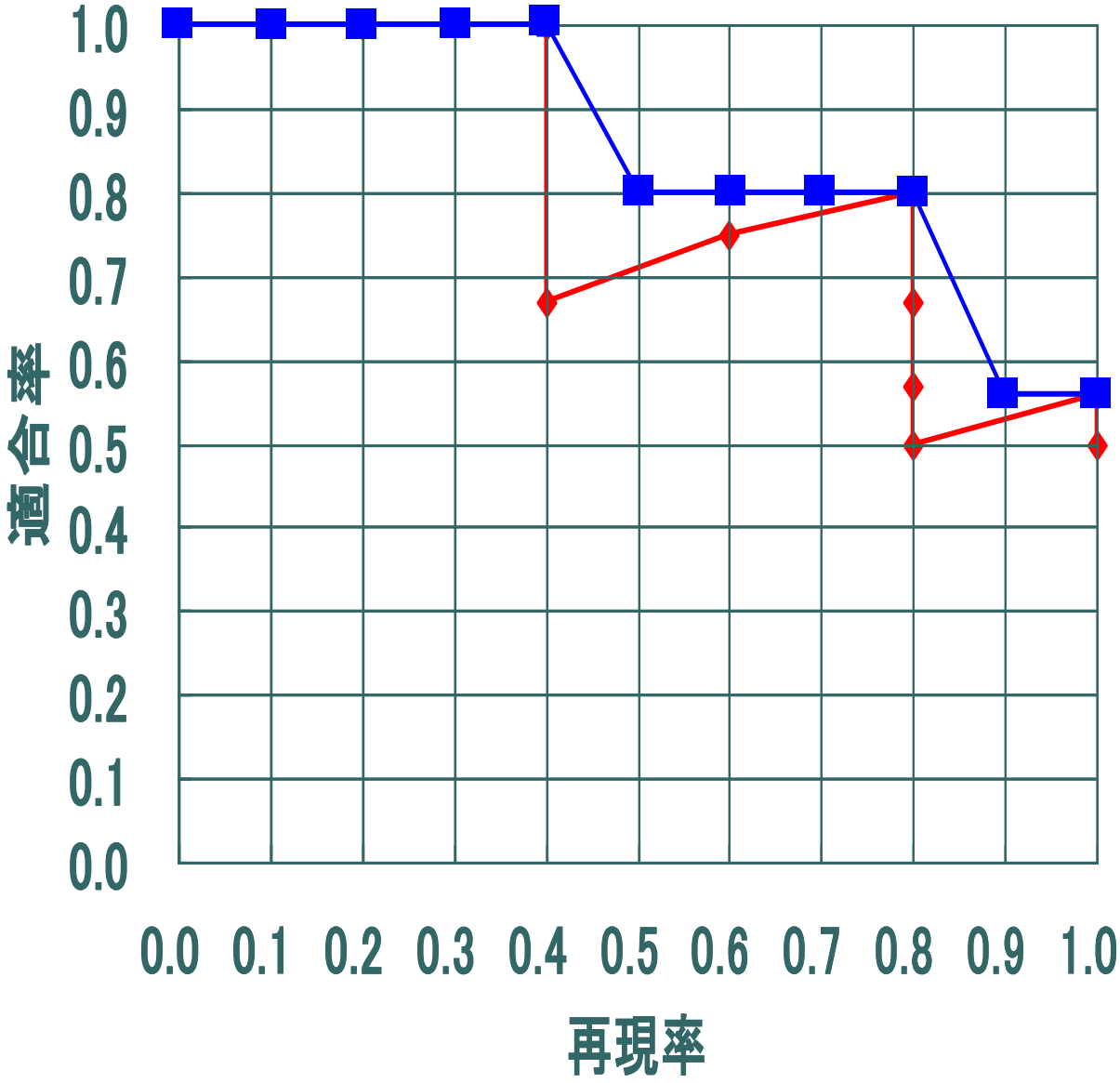
再現率を0.0～1.0の
11点で求める

(11点)補間適合率:

再現率 x での補間適合率 $P(x)$ は，上位 i 件での再現率を R_i，適合率を P_i とすると:

$$P(x) = \max_{x \le R_i} P_i$$

(x 以上のすべての再現率における適合率の最大値)



再現率－適合率曲線の見方

- 理想は，すべての再現率で適合率100%
- 通常は右下がりの曲線になる
  - 一般に，再現率が上がれば適合率は下がるため
- 曲線が上にあるほど検索性能が良いと考えられる
  - 曲線が交差している場合は，どちらが良いか一概には言えない
    - 再現率と適合率のどちらを重視するかに依存

再現率･適合率の要約(1)

- 再現率･適合率曲線によってシステムの検索性能を比較できるが，一つの数値で比較できない
- **平均適合率**:0.0～1.0の11点の各再現率レベルでの補完適合率の平均値

$$\overline{P} = \frac{1}{11}\sum_{i=0}^{10} P\left(\frac{i}{10}\right)$$

  - 2ページ前の例では，

$$\overline{P} = \frac{1}{11}(1.0\times 5 + 0.8\times 4 + \frac{5}{9}\times 2) \cong 0.846$$

再現率・適合率の要約(2)

- **F値**: 再現率と適合率の調和平均

$$F = \frac{2PR}{P+R}$$

P: 適合率
R: 再現率

  - 再現率と適合率のバランスを重視
    - たとえば P=0.9, R=0.1 の場合，算術平均は 0.5 だが，F値は以下の通り 0.18 となる

$$F = \frac{2 \times 0.9 \times 0.1}{0.9 + 0.1} = 0.18$$

  - 平均適合率と異なり，ランキングは考慮されない

実際の再現率・適合率の計算

- 再現率・適合率を求めるには，問合せにどの文書が適合するかを調べる必要
  - 適合率は，検索結果に含まれる文書のみの適合性が分かれば計算できる
    - 検索結果が100件であれば，100件の適合性を見る
  - 再現率は，「再現率と適合率の式」のページにおける A の集合が分からなければ計算できない
    - A：文書集合全体における適合文書
    - 文書集合が大きい場合，この数を求めるのは困難

テストコレクション

- 情報検索システムの評価用に作成された人工的なデータセット
  - 文書集合
    - 新聞記事，学術論文，Webページなどの文書の集合
  - 検索課題集合
    - 文書集合に応じて人手で作成された検索課題の集合
    - 検索課題のキーワード，情報要求の説明文など
  - 適合情報
    - 検索課題集合中の各問合せに対する適合文書

テストコレクションの例

- **NTCIR**
  - 国立情報学研究所が1998年から行っている情報検索の評価型ワークショップ
- **NTCIR-3 Web**
  - jpドメインから収集したWebページを対象
    - 100GB（約1500万文書）
- NTCIRと同様の評価型ワークショップとして，**TREC**（アメリカ），**CLEF**（欧州）などがある

NTCIR-3 Web：検索課題の例

検索課題

検索課題ID

キーワード（3語以内）

情報要求の記述（1文）

```
<TOPIC>
<NUM>0008</NUM>
<TITLE CASE="b">サルサ, 学ぶ, 方法</TITLE>
<DESC>サルサを踊れるようになる方法が知りたい</DESC>
<NARR><BACK>最近はやっているサルサという踊りを学ぶためにどう
    すればよいか具体的な方法が知りたい。例えば教室に通うという場合
    には、その場所や授業形態など、具体的な内容を必要とする。
    </BACK><RELE>具体的な方法の表記のない、流行であることのみを
    扱った文書は不適合とする。</RELE></NARR>
<CONC>サルサ, 習う, 方法, 場所, カリキュラム</CONC>
<RDOC>NW011992774, NW011992731, NW011992734</RDOC>
<USER>大学院修士1年, 女性, 検索歴2.5年</USER>
</TOPIC>
```

情報要求の説明

適合文書ID

検索課題作成者

NTCIR-3 Web：検索課題の例（2）

検索課題

検索課題ID

キーワード（3語以内）

情報要求の記述（1文）

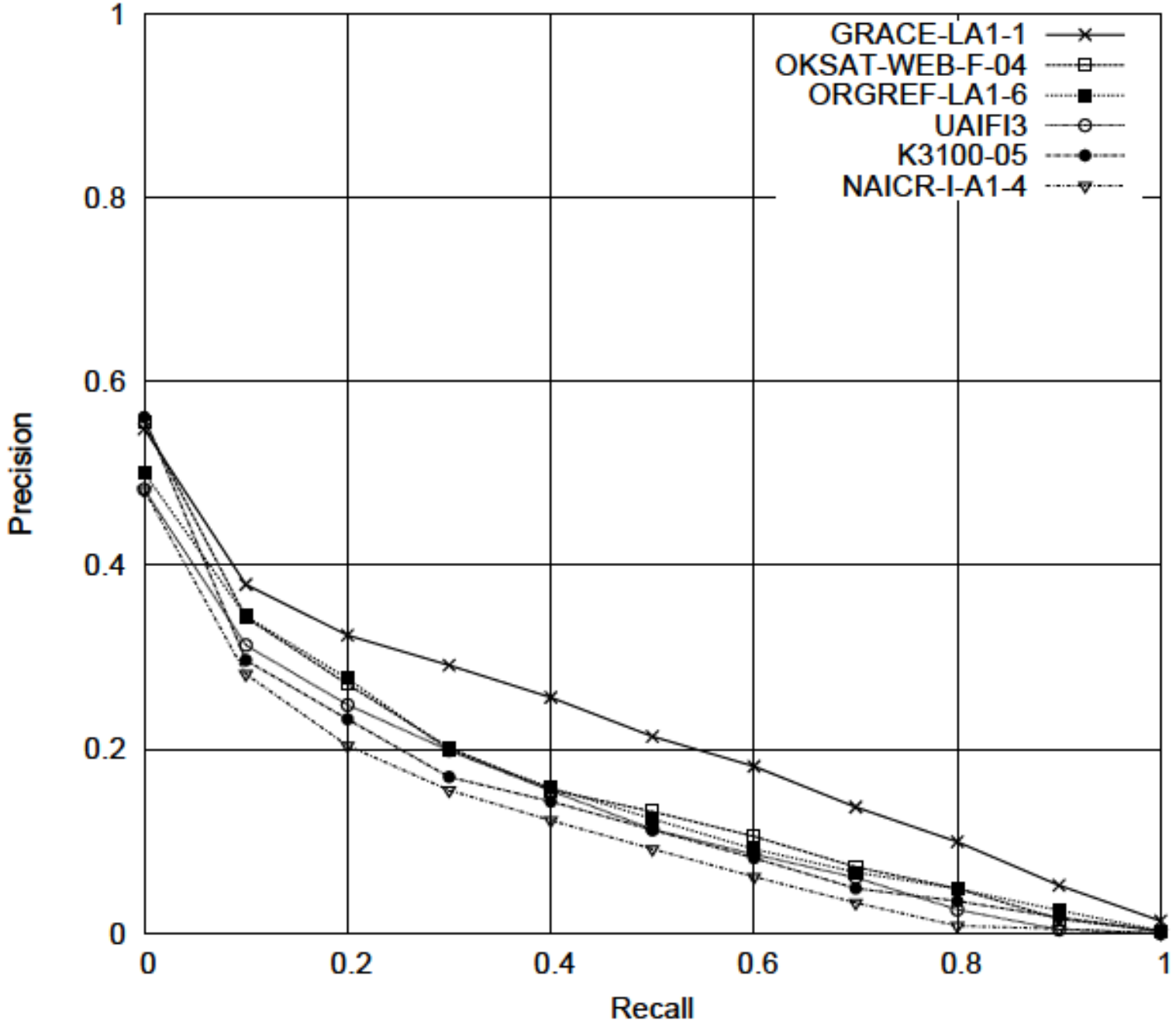
情報要求の説明

適合文書ID

検索課題作成者

```
<TOPIC>
<NUM>0014</NUM>
<TITLE CASE="b">夢, 将来, 努力</TITLE>
<DESC>将来の夢として、どのようなものを人は抱き、それに向かってどのような努力をしているのだろうか</DESC>
<NARR><RELE>適合文書は、夢について語り、そのための努力、あるいはしようとしていることについて記述している文章とする。この際、あまりにも空想的な、実現不可能な夢は対象ではない。その判断基準としては、そういった場合、夢とそれに対する努力の方向性の根拠が曖昧になるであろうから、夢とその実現に向けた努力の結び付きに注目し、それが弱いものは不適合とする。</RELE><NARR>
<CONC>夢, 将来, 努力, 目標, 予定, 自分, 私</CONC>
<RDOC>NW002695723, NW006763670, NW012273176</RDOC>
<USER>大学院修士1年, 男性, 検索歴5年</USER>
</TOPIC>
```

再現率-適合率曲線の例



- NTCIR-3 Web
- 100GB

テストコレクションの利点・欠点

- **利点**
  - ある情報検索システムの検索性能の評価が可能
  - 異なる情報検索システム間で客観的な検索性能の比較が可能
- **欠点**
  - 人手ですべての適合文書を調べるため，作成に手間がかかる
  - 同じ条件で比較するという点では客観的だが，適合情報にどうしても判定者の主観が入るのは避けられない

まとめ

- 情報検索システムの有効性の客観的な評価尺度として，再現率・適合率が良く使われる
- 利用目的によって，どちらが重要かは異なる
  - 事実検索と特許検索の違い
- 複数の検索システムの性能の比較，あるいは一つのシステムの性能の改良のためにテストコレクションが使われる