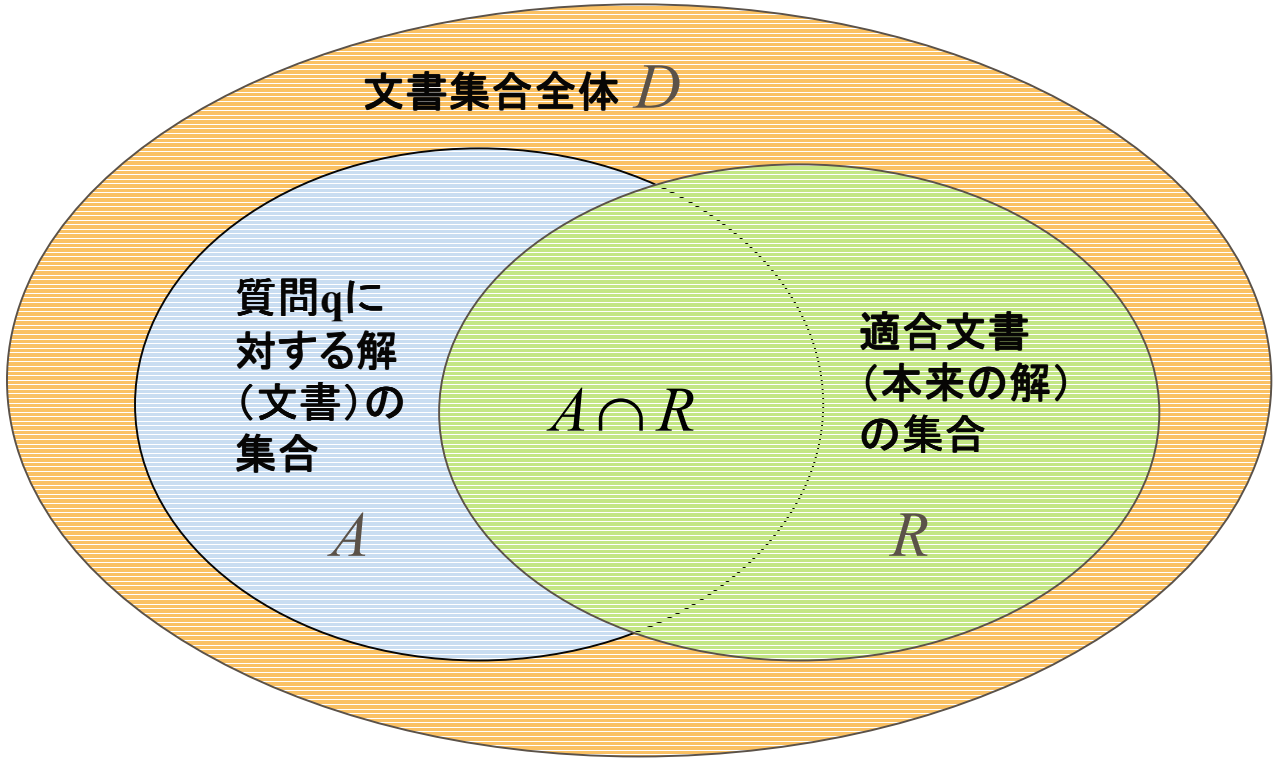
情報学科CSコース情報システム（3年後期）
講義ノート
―第2回―

適合率，再現率，ベクトル空間モデル，類似検索，tf/idf法，適合フィードバック，クラスタリング，LSI法

田中克己

角谷和俊

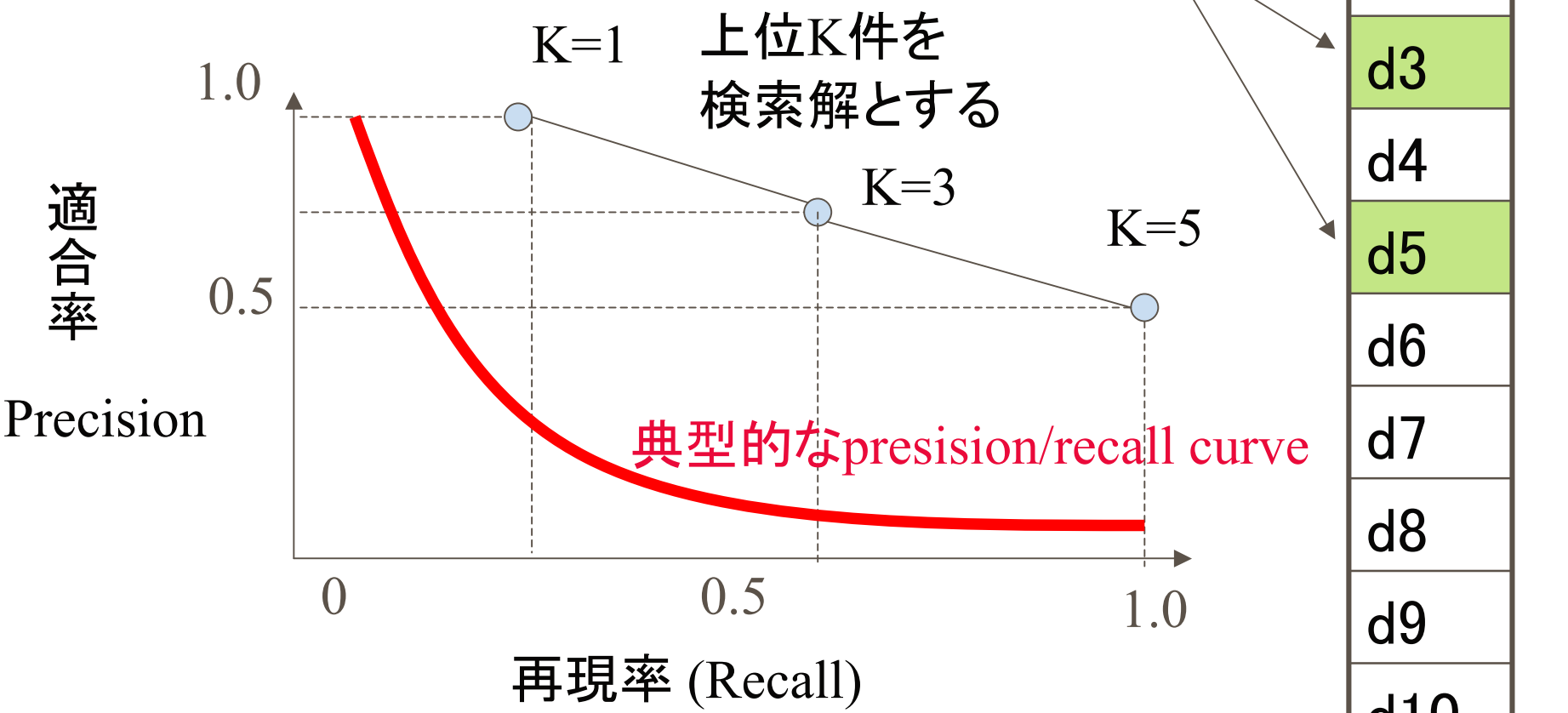
情報検索システムの評価尺度



- 再現率 $recall = \frac{|A \cap R|}{|R|}$
- 適合率（精度） $precision = \frac{|A \cap R|}{|A|}$

再現率 vs 適合率

- 一般に，互いにtrade-offの関係
- Precision/recall curve



テストコレクション

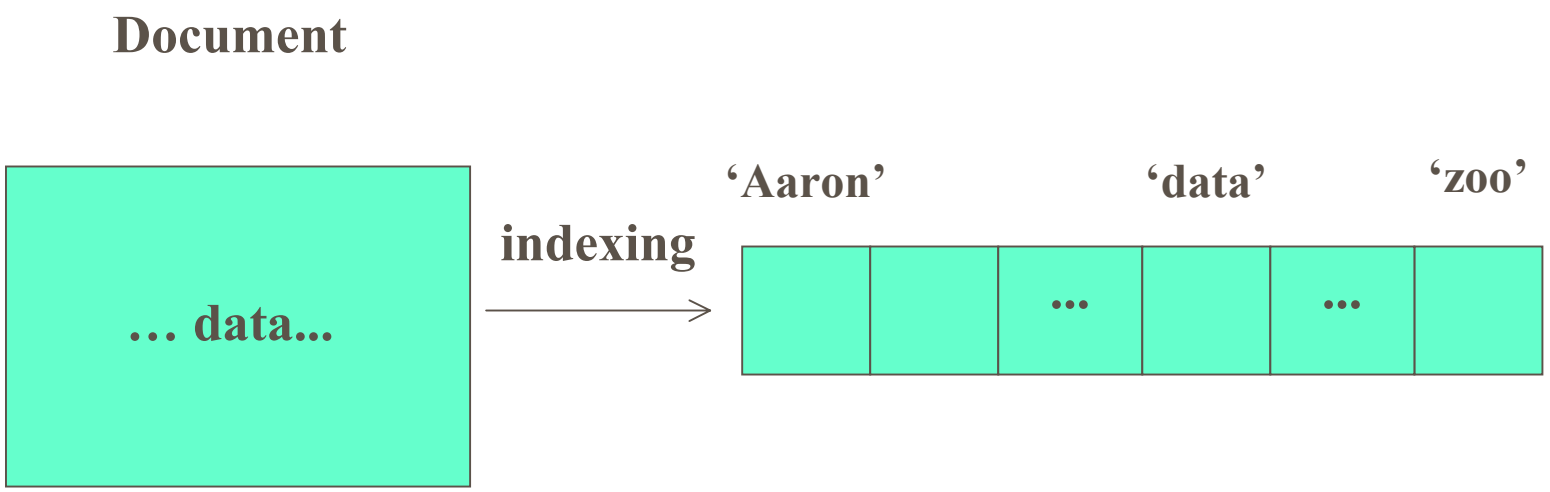
- テストコレクション
  - (a) 文書集合，(b) 多数の質問，(c)各質問に対する適合文書の集合を組にしたデータベース．情報検索システムの性能評価に重要．
- 適合文書集合の作成の困難さ
  →再現率計算の困難さ
  - 適合解の集合を作ることは大規模テストコレクションや，Web検索では困難．
  - Pooling method:同一の文書集合に対し，多数の検索エンジンで同じ質問を出し，上位N 個の検索結果を全て集める．N の値として，100 程度が多い。この結果に対してのみその適合性を人手で判断し，それを文書集合全体における適合文書集合とする．

ベクトル空間モデル（Vector Space Model）

- indexing
  各文書をV次元ベクトルで表現.
  ベクトルの各要素は{1,0}または正実数（語の重み）
  （Vは，文書群から抽出された索引語の総数）
  （ストップワードリストやシソーラスの利用）
- cluster generation
  類似ベクトル群をグループ化
- cluster search
  質問(ベクトル)にもっとも類似のクラスタを検索

ベクトル空間モデルの特徴ベクトル生成

*出現：1，非出現：0
*語の出現頻度（正規化）
* tf/idf法（term frequency/inverse document frequency）



tf/idf法(1)

- ターム頻度(term frequency: tf)

$$tf_{ij} = freq(i, j)$$ 文書D_i におけるタームt_j の出現頻度

$$tf_{ij} = K + (1 - K)\frac{freq(i, j)}{\max_{i,j}(freq(i, j))}$$

$$tf_{ij} = \frac{\log(freq(i, j) + 1)}{\log(\text{文書 j 中の総ターム種類数})}$$

tf/idf法(2)

- 文書頻度　document frequency
  df_j ＝ タームt$_j$が出現する文書数
- 実際はその逆のinverse document frequencyをを使う.
  文書総数Nによる正規化

$$idf_j = \log \frac{N}{df_j}$$

- 文書D$_i$のタームt$_j$の重み　$w_{ij} = tf_{ij} \times idf_j$

類似度(1)

- 文書D_jの特徴ベクトル

$$D_i = (w_{i1}, w_{i2}, \ldots, w_{in})$$

- 質問Qの特徴ベクトル
  - タームt_jを含めば1，含まなければ0という値からなるベクトル

$$Q = (w_{q1}, w_{q2}, \ldots, w_{qn})$$

- nは文書集合における全ての異なるターム数

類似度(2)

- 内積

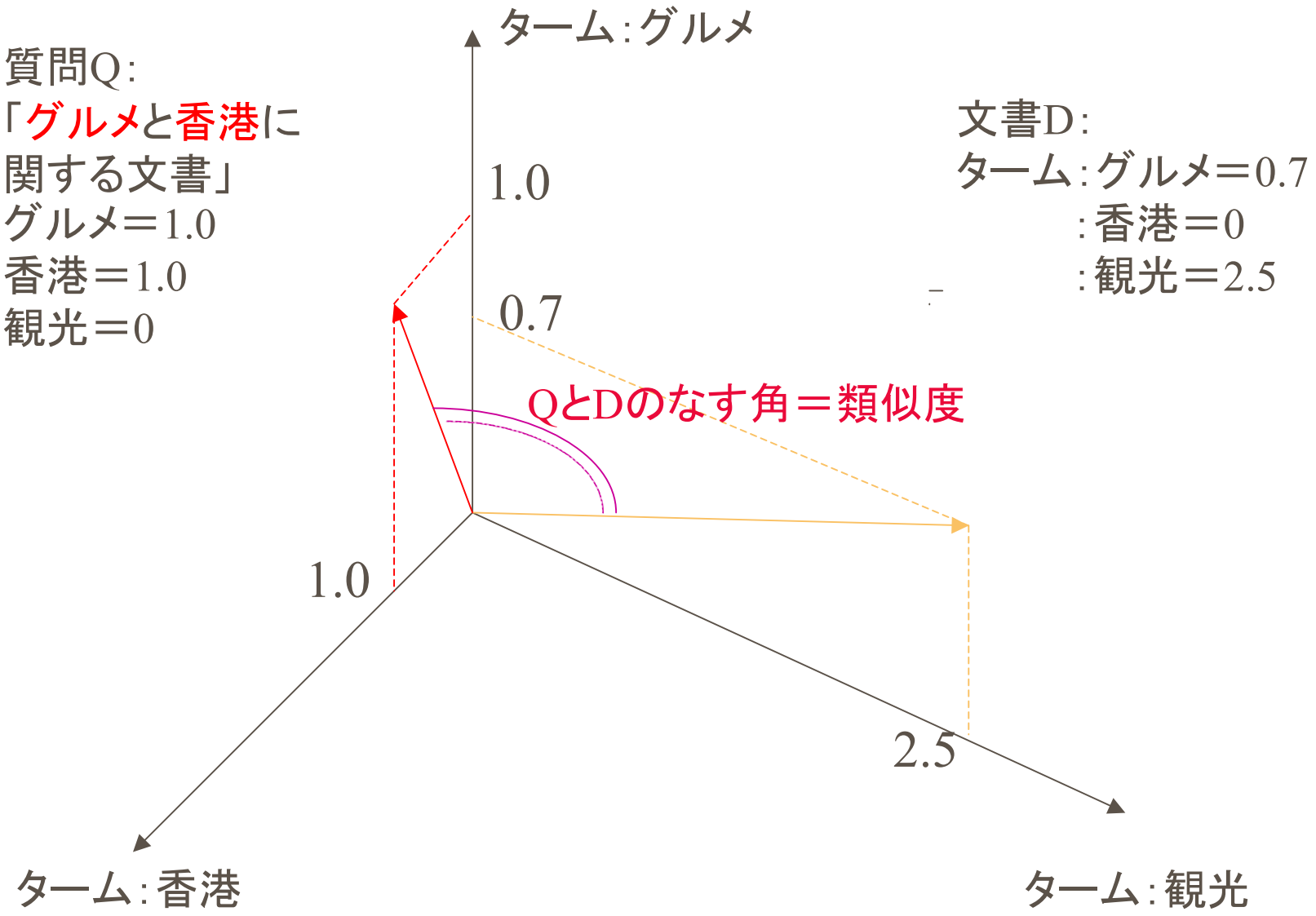
$$\mathrm{sim}(Q, D_i) = w_{q1} w_{i1} + ... + w_{qn} w_{in}$$

- コサイン相関値

$$\mathrm{sim}(Q, D_i) = \frac{w_{q1} w_{i1} + ... + w_{qn} w_{in}}{\sqrt{w_{q1}^2 + ... + w_{qn}^2} \times \sqrt{w_{i1}^2 + ... + w_{in}^2}} = \cos\theta$$

- 質問と文書の類似度
  文書と文書の類似度

ベクトル空間と類似度


ターム:グルメ
質問Q:
「グルメと香港に
関する文書」
グルメ＝1.0
香港＝1.0
観光＝0
文書D:
ターム:グルメ＝0.7
:香港＝0
:観光＝2.5
1.0
0.7
QとDのなす角＝類似度
1.0
2.5
ターム:香港
ターム:観光

練習問題

- Q: “gold silver truck”
- D1: “Shipment of gold damaged in a fire”
  D2: “Delivery of silver arrived in a silver track”
  D3: “Shipment of gold arrived in a truck”
- tf/idf法で各文書の特徴ベクトルを求めよ
  Qと各文書の類似度（内積，コサイン相関値）で求めよ.

	a	arrived	damaged	delivery	fire	gold	in	of	silver	shipment	truck
D1	0	0	.477	0	.477	.176	0	0	0	.176	0
D2	0	.176	0	.477	0	0	0	0	.954	0	.176
D3	0	.176	0	0	0	.176	0	0	0	.176	.176
Q	0	0	0	0	0	.176	0	0	.477	0	.176

内積の場合，ランキングはD2>D3>D1

パッセージ検索

- パッセージ
  - 文書の内容を特徴付けるのは文書全体よりはむしろ特定の部分（段落など）
  - 文書Dの代わりにパッセージP1, …,Pkの各特徴ベクトルと質問ベクトルとの類似度を計算しこれをマージする.
- パッセージの候補

  １　固定長に分割したテキストの部分

  ２　形式段落

  ３　形式的な節、章

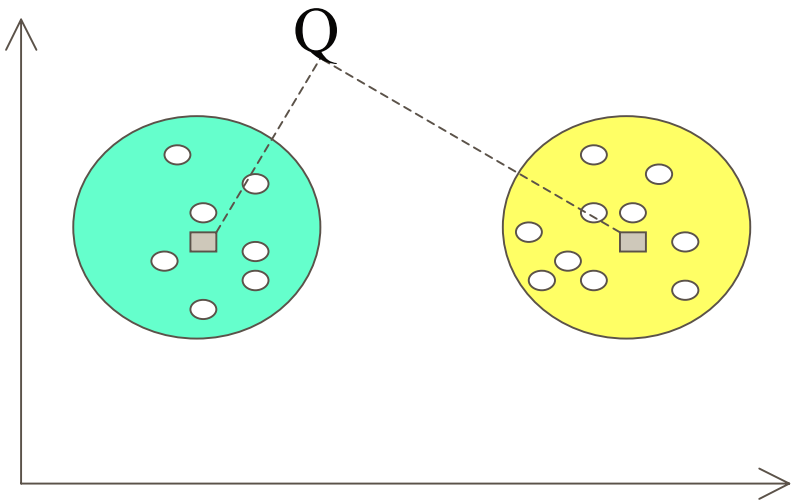
適合フィードバック

- Rocchio (1971)
  - ベクトル空間モデルでの適合フィードバック
  - Rocchio,J.J. “The SMART Retrieval System Experiments in Automatic Document Processing,” chapter Relevance Feedback in Information Retrieval, pp.313-323, Prentice Hall.
  - Qは元の質問．Q’はQとユーザの反応から修正された質問．Qの検索結果集合の内，$R_1,..,R_{n1}$はユーザが適合と判断したもの，$S_1,\ldots,S_{n2}$は不適合と判断したもの．このプロセスを繰り返す．（各項に適当な重みα，β，γを付加）

$$Q' = Q + \frac{1}{n_1}\sum_{i=1}^{n1} \mathrm{R_i} - \frac{1}{n_2}\sum_{i=1}^{n2} Si$$

クラスタリング（clustering）

- 文書–クラスタ間の類似度（例えばコサイン相関値）
  - 文書とクラスタの中央値（centroid）
  - クラスタ内の文書との距離のうち最小のもの
  - クラスタ内の文書との距離のうち最大のもの



クラスタ生成

- 健全なクラスタ生成の方法
  - グラフ理論的アプローチ
  - 文書間の類似度がある閾値を超えたものを枝（edge）で結ぶ　→　無向グラフ
  - 無向グラフ中の連結成分（connected component）または極大クリーク（clique，部分完全グラフ）を1つのクラスタとする．文書数Nに対してO(N×N)以上の計算量必要
- 反復法
  - サンプルから適当なクラスタ（seeds）作成．ある文書をそれに最も近いクラスタに追加．クラスタのセントロイドを修正，これを繰り返す．高速．