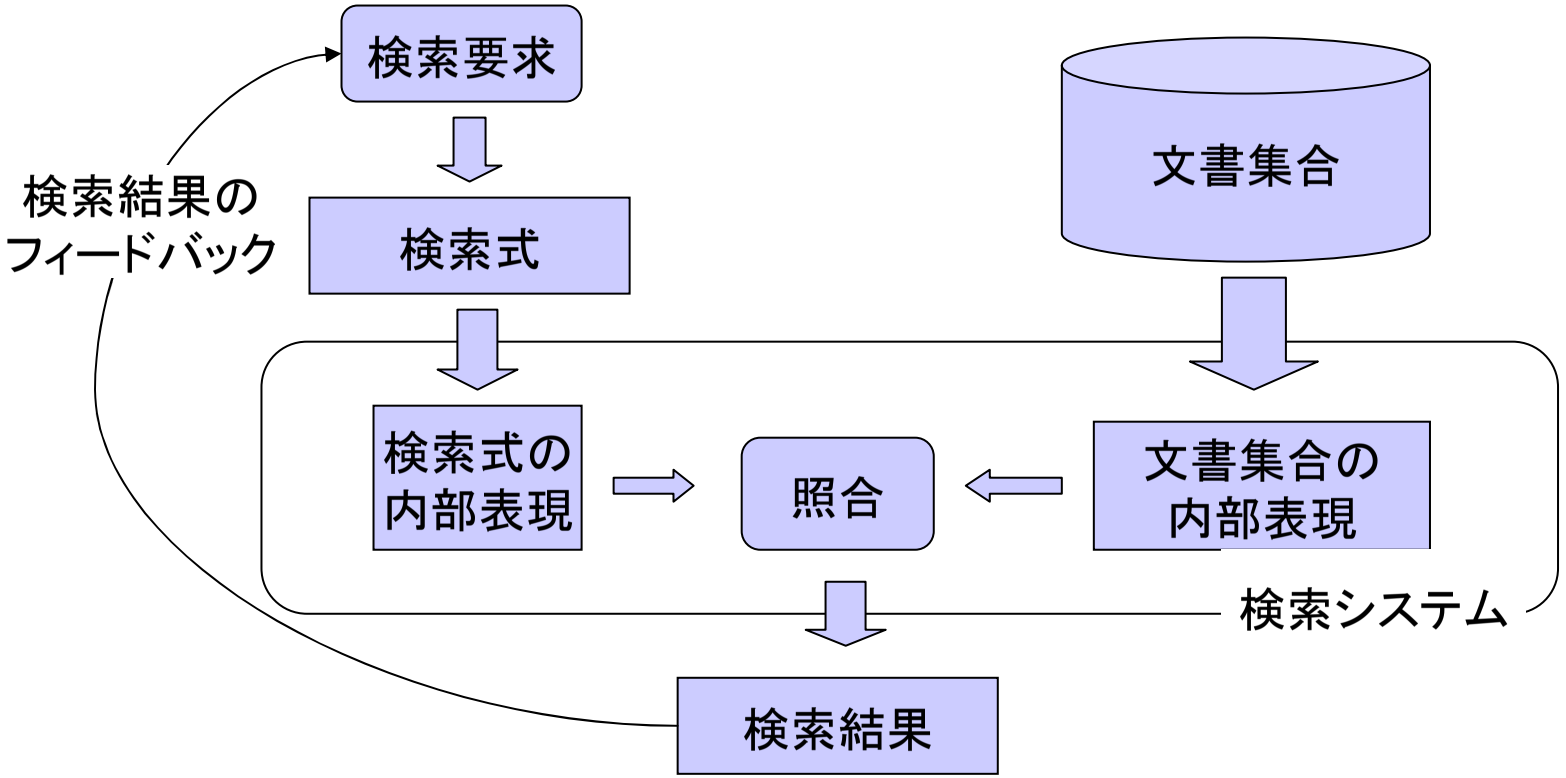
第3週ゼミ

1. 情報検索

- 情報検索とは･･･大量の情報の中からユーザの要求を満たす情報を見つけだすこと



索引付け

- 検索システムにおいて検索対象である文書集合を計算機内部で処理できる内部表現に変換する処理

- 人手による索引付け
- 自動的な索引付け

自動的な索引付け

- 文書から語を切りだすために形態素解析を行う
- 方法１
- 機能語・・・助詞「が」、「は」、「を」
  助動詞「だ」、「だろう」
- 内容語・・・自立語のように意味を持つ語
  「桜」「美しい」
- 機能語を削除し、内容語のみを索引語としてもちいることで索引語の数を減らすことができる。

自動的な索引付け

- 方法2
- Nグラム索引
- 決められた文字数の単位で文章を切り出し索引語とする

  「自然言語」→「自然」、「然言」、「言語」
- 無意味な索引が作成される
- アルゴリズムが単純、検索漏れを生じない

自動的な索引付け

- より適切な文書を検索するために・・・

  索引語が文書の内容とどれだけ密接に関係しているかを求める
- tf･idf法

$$\mathrm{tf} \cdot \mathrm{idf} = \mathrm{tf} \times (\log \frac{\mathrm{N}}{\mathrm{df}} + 1)$$

N:全文書数

tf:文書内での索引語の出現頻度

df:文書集合中で索引語が表れた文書数

2. 検索モデル

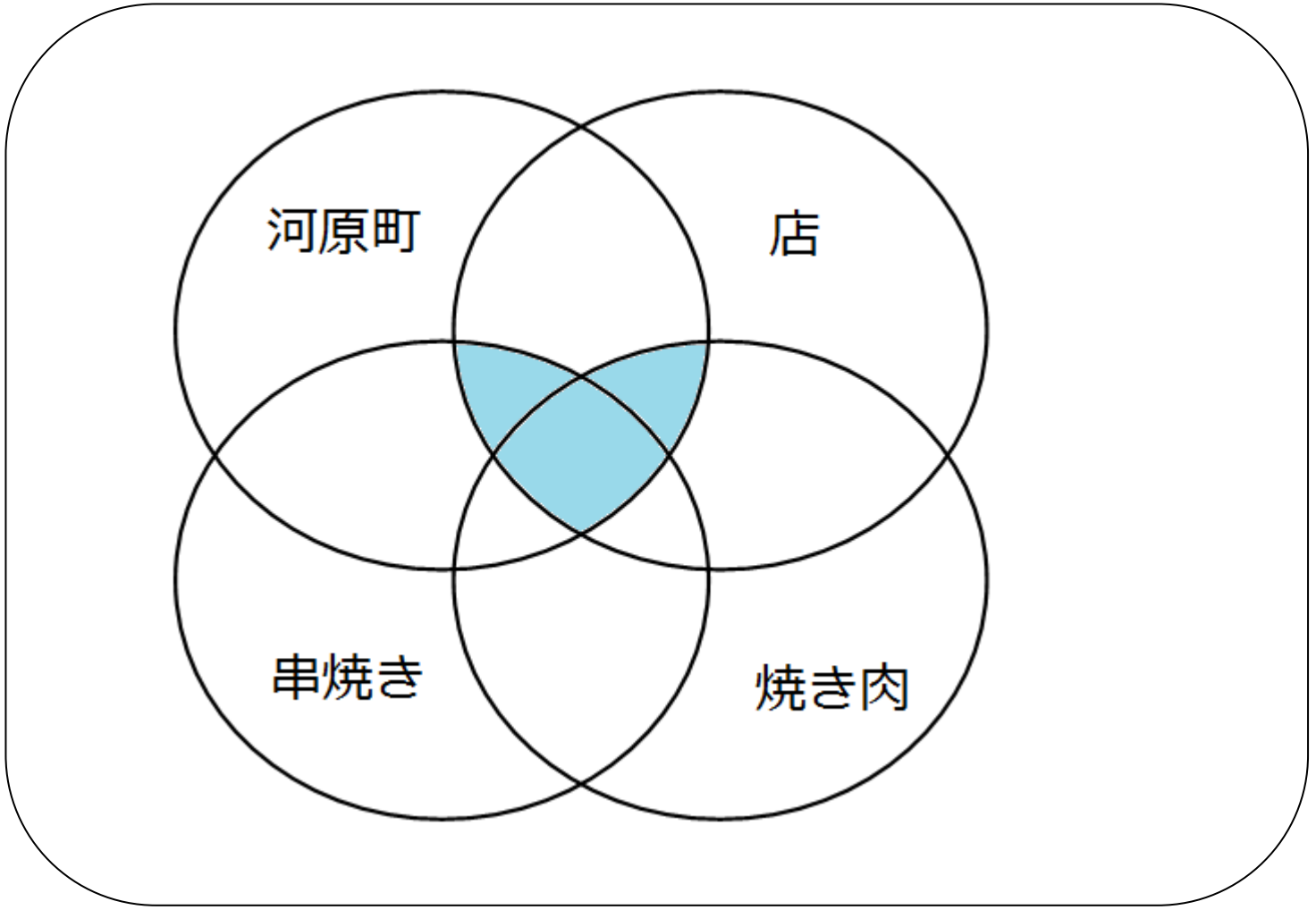
- ブーリアンモデル
- ベクトル空間モデル

ブーリアンモデル

- AND、OR、NOTの組み合わせで表現する
  「河原町にある串焼きか焼き肉のお店」
→　河原町 AND（串焼き OR 焼き肉）AND 店

- 検索語同士の関係を明示的に記述できる
- 論理式の組み合わせにより複雑な検索要求に対応できる
- 検索語の完全一致が必要
- 関連がありそうな情報の検索、結果の順位付けができない

ブーリアンモデル



ベクトル空間モデル

- 文書中の索引語の重みを要素としたベクトルで文書を表現する

- 検索対象となる文書D_1、D_2、…、D_m

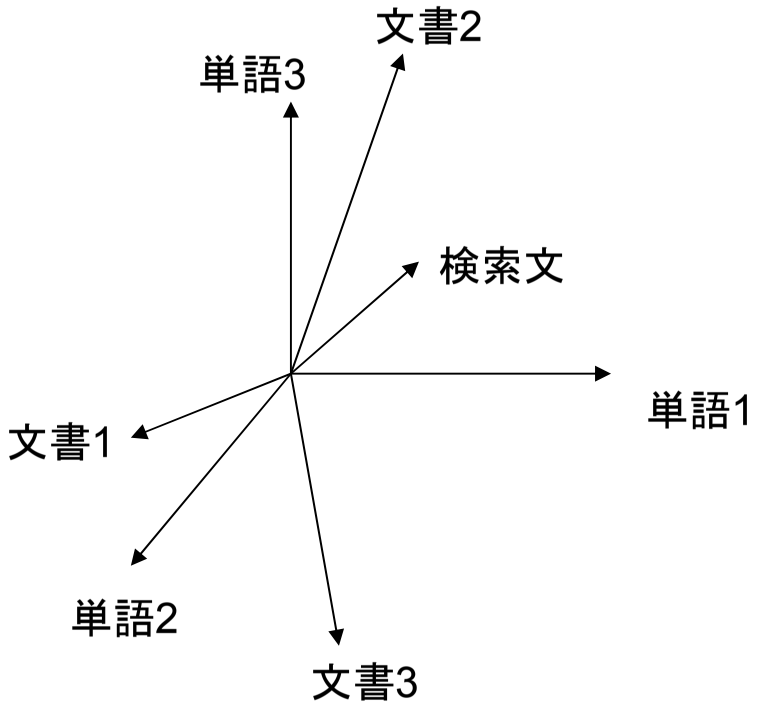
$$D_j = \begin{pmatrix} d_{1j} \\ d_{2j} \\ \vdots \\ d_{mj} \end{pmatrix}$$

- d：索引語の重み
- 文書内に存在しない索引語の重みは0

ベクトル空間モデル

- 同様に検索質問もベクトルで表現する

$$\mathbf{q} = \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{pmatrix}$$



ベクトル空間モデル

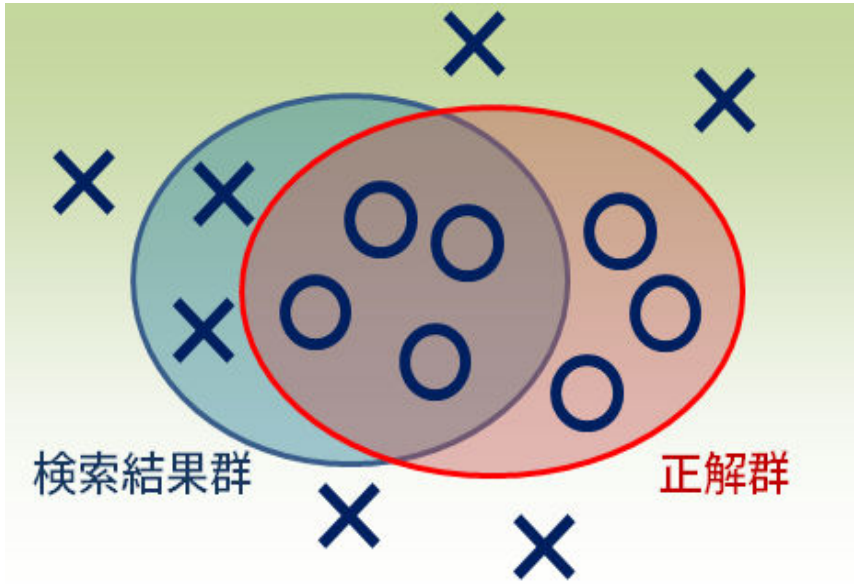
- 文書をベクトルで表すことで2文書間の類似度を求める事ができる

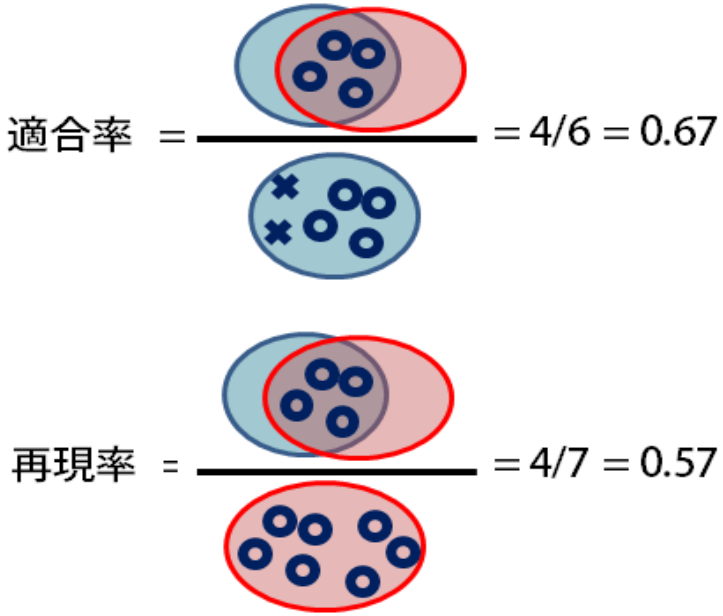
文書	類似度
文書D1	0.41
文書D2	0.21
文書D3	0.28
文書D4	0
文書D5	0.73

- 検索質問qに対して文書はD5、D1、D3、D2、D4の順に近い
- 大規模な文書集合を対象とする場合、計算量が非常に大きくなる

3. 情報検索システムの評価

- 再現率･･･検索式に適合する文書をどれだけ検索できたか
- 適合率･･･検索結果のうちどれだけが検索式に適合しているか





情報検索システムの評価

- 検索システムの総合的な評価
- F値・・・再現率Rと適合率Pの平均調和

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