自然言語処理論 I

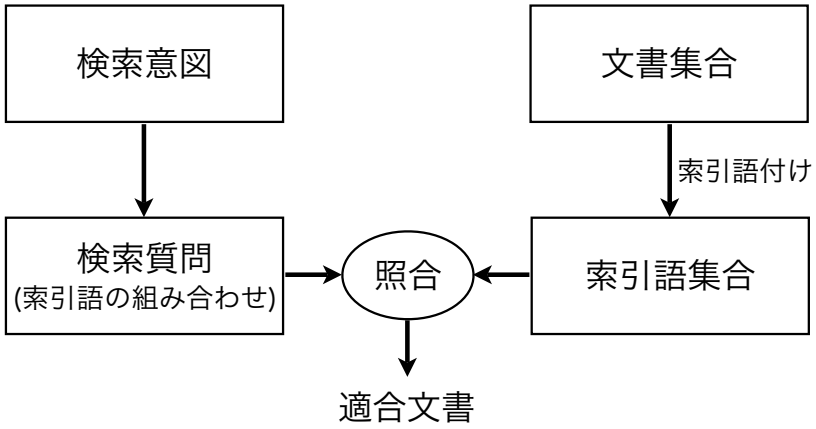
12. テキスト処理(文字列照合と検索)

情報検索

- information retrieval (IR)
- 広義の情報検索
  - 情報源からユーザの持つ問題(情報要求)を解決できる情報を見つけ出すこと
- 狭義の情報検索
  - 文書集合の中から、ユーザの検索質問に適合する文書を見つけ出すこと
    - ✦ 適合文書: 検索質問の答えが書いてある文書
  - テキスト検索(text retrieval)

情報検索

- 索引語(index term)による照合



検索質問(query)

- 索引語またはその組み合わせ
- 検索質問の与え方
  - 索引語を直接利用する
    - ✦ 論理式の利用
    - ✦ ex. $(t_a \text{ and not } t_b) \text{ or } t_c$
  - 自然言語で記述する
    - ✦ 索引語に自動的に変換する
    - ✦ ex. 「チーズの作り方が知りたい」
      → チーズ and 作り方

索引語付け(indexing)

- 文書から索引語を取り出すこと
  - 自動索引語付け
    - ✦ テキスト検索の対象文書数が多いため
  - 形態素解析などの処理が必要
- 索引語の単位
  - 単語 (チーズ、作り方、材料)
  - 句 (チーズの作り方、チーズの材料)
  - 適切な単位を決めることは難しい
    - ✦ 単語を索引語とすることが一般的

ストップワード

- ストップワード (stop word) とは?
  - 索引語に加えるべきでない単語
- 具体的には...
  - 機能語 (function word)
    - ✦ 日本語: 助詞, 助動詞など
    - ✦ 英語: 冠詞, 前置詞 など
    - ✦ (参考) 内容語 (content word)
      名詞, 動詞など、意味のある単語
  - be動詞、have、ピリオドなどの記号
  - どの文書にもよく出現し、情報検索の手がかりとはならないため

照合

- 転置インデックス法
  - inverted indexing
- ベクトル空間モデル
  - vector space model (VSM)

転置インデックス法

- 文書毎に索引語のリストを作る

	小説	あらすじ	書評	推理
文書1	1	0	0	0
文書2	1	0	1	1
文書3	1	1	0	0
文書4	0	0	1	0

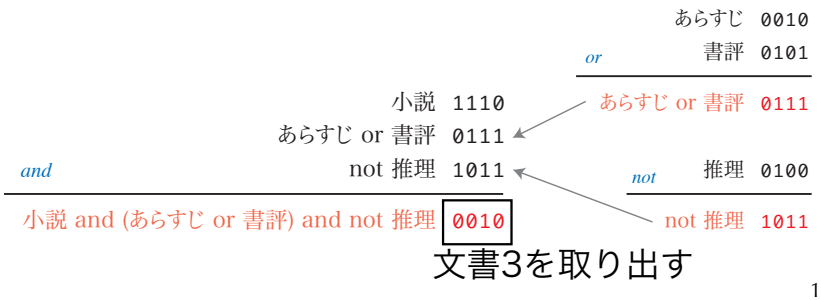
転置インデックス法

- 行列を転置する
  - 索引語を含む文書のリストがすぐに得られる

	文書1	文書2	文書3	文書4
小説	1	1	1	0
あらすじ	0	0	1	0
書評	0	1	0	1
推理	0	1	0	0

転置インデックス法

- 検索質問を論理式で与える場合
  - 転置インデックスの行をベクトルとみなす
  - ベクトルのビット演算で計算可能
- 小説 and (あらすじ or 書評) and not 推理



文書3を取り出す

ベクトル空間法

- 文書と検索質問をベクトルで表現
  - 文書ベクトルD_i, 検索質問ベクトルQ
  - ベクトル間の類似度を計算
    - ✦ 最大の類似度を持つ文書D_iを取り出す
- ベクトル
  - w_j^iは索引語の重み

$$D_i = \begin{pmatrix} w_1^i \\ \vdots \\ w_j^i \\ \vdots \\ w_n^i \end{pmatrix} \begin{matrix} \leftarrow 索引語_1 \\ \vdots \\ \leftarrow 索引語_j \\ \vdots \\ \leftarrow 索引語_n \end{matrix}$$

索引語の重み付け

- 単純な重み付け
  - 文書に存在すれば1、それ以外は0
  - (検索質問ベクトルQの重み付け)
- TF・IDF法
  - TF (term frequency)
    - ✦ tf_j^i: 文書 i における索引語 j の頻度
    - ✦ 同じ文書に何回も現われる単語ほど、検索の有力な手がかりとなる

索引語の重み付け

- TF・IDF法(つづき)
  - IDF (inverse document frequency)
    - $idf_j = \log \frac{N}{df_j}$
    - df_j: 文書頻度 (索引語jを含む文書数)
    - 色々な文書に現われる単語は、検索の有力な手がかりとはならない
  - 索引語の重み

$$w_j^i = {tf_j}^i \cdot idf_j = {tf_j}^i \cdot \log \frac{N}{df_j}$$

2ベクトルの類似度計算

- 類似度: sim(D_i,Q)
  - 類似度の大きい上位n個の文書を取り出す
- 類似度の例
  - ベクトルの内積

$$D_i \cdot Q = \begin{pmatrix} w_1^i \\ \vdots \\ w_n^i \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ \vdots \\ q_n \end{pmatrix} = \sum_j w_j^i q_j$$

    - 特にq_jが1または0、w_j^iの要素がTF・IDFのとき
      内積=検索質問に含まれる索引語のTF・IDFの和

テキスト検索の評価

- 一般的なテキスト検索システム
  - 検索質問Qを入力
  - Qに適合すると思われる文書をn個出力
    - ex. sim(D_i,Q)の値の大きい順に文書を出力
  - 出力文章数は容易に調整可能

テキスト検索の評価

- 評価基準
  - precision (適合率、精度)

$$\frac{\text{システムが出力した適合文書数}}{\text{システムが出力した文書数}}$$

  - recall (再現率)

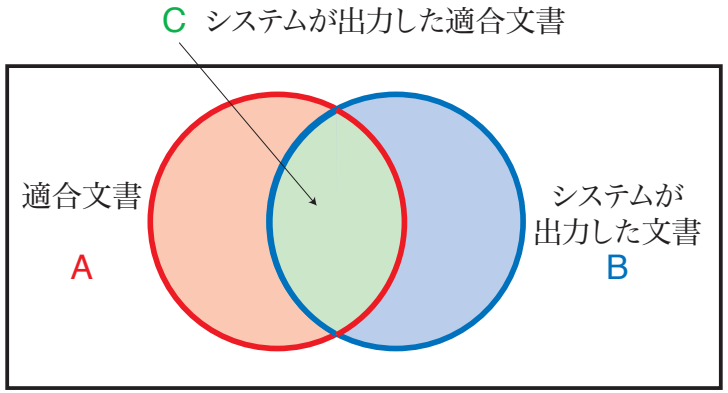
$$\frac{\text{システムが出力した適合文書数}}{\text{文書集合に含まれる適合文書数}}$$

  - F値 (F-measure)

$$F = \frac{2PR}{P+R} \quad (P\text{= precision},\ R\text{= recall})$$

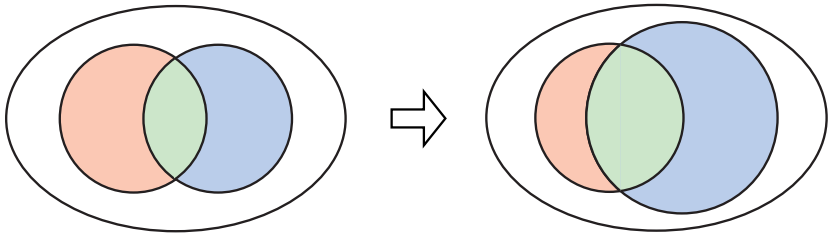
precisionとrecall

- precision = C / B
- recall = C / A



precisionとrecall

- 両者は一般にトレードオフの関係
- システムが多くの文書を取り出せば...
  - precision 小、recall 大



precisionとrecall

- precisionが重視されるとき
  - ユーザに適合文書のみを提示したいとき
  - ウェブの検索エンジン
- recallが重視されるとき
  - 検索漏れを少なくしたいとき
  - 特許文書の検索
- precisionとrecallの両方を評価するとき
  - F-値による評価

テキスト検索の工夫

- より正確なテキスト検索を目指す
- 関連フィードバック
  - relevance feedback
- 質問拡張
  - query expansion

関連フィードバック

- 1回の検索で良い結果が得られることは稀
- ユーザとインタラクティブに検索を行う
- 全体の流れ
  - システムがテキスト検索を行う
    - n個の文書をユーザに提示する
  - ユーザは、個々の文書が適合文書であるかどうかを判定する

(例)

文書1	文書2	文書3	文書4	文書5
○	×	○	○	×

関連フィードバック

- 全体の流れ(続き)
  - 検索質問ベクトルQを修正する
    - $Q' = Q + \frac{1}{|R|}\sum_{D_i \in R} D_i - \frac{1}{|N|}\sum_{D_i \in N} D_i$
    - R: ユーザが適合文書と判定した文書集合
    - N: ユーザが不適合文書と判定した文書集合
  - Q'で検索をやり直す
  - 以上を繰り返す

関連フィードバック

- 関連フィードバックの効果
  - 適合文書と似た文書が新たに検索される
  - 非適合文書と似た文書は検索されなくなる
  - precision, recallの向上が期待できる
- 擬似関連フィードバック
  - 人間による適合文書の判定は行わない
  - 検索結果の上位の文書を適合文書とみなして適合フィードバックを行う
  - 完全な自動処理

質問拡張

- 自然言語には様々な表現がある
  - 検索質問が「自動車」のとき
  - 車、乗用車、自家用車を含む文書を取り出すことはできない
- 質問拡張とは?
  - 検索質問中の単語と関連のある単語を検索質問に自動的に追加する処理
  - Q=(自動車)
    → Q=(自動車、車、乗用車、自家用車)
  - recallの向上が期待できる

質問拡張

- 検索質問に加えるべき単語は?
  - 異表記の単語
    - ✦ 林檎 → りんご
    - ✦ 言い換える → 言い替える、いいかえる
  - 同義語
    - ✦ 映画 → ムービー、シネマ、キネマ、フィルム
  - 上位語
    - ✦ ビール → 酒
  - 下位語
    - ✦ 酒 → 日本酒、ビール、ワイン、ウィスキー...
- 辞書、シソーラスを利用する

まとめ

- テキスト検索の手法
  - 索引語付けによるテキスト表現
  - 転置インデックス法
  - ベクトル空間モデル
    - ✦ TF・IDF法による重み付け
- 評価基準
  - precision, recall, F値
- テキスト検索の工夫
  - 関連フィードバック
  - 質問拡張