横浜国立大学 大学院 環境情報学府
情報メディア環境学専攻(前期)

言語情報処理原論(9)

Foundation of Natural Language Processing (9)

森 辰則
mori@forest.eis.ynu.ac.jp

文書情報の組織化

多数の情報の中から利用者が要求する情報を捜し出す．これを効率良く行なうための各法

- □選別
  - ○情報検索
    - ▹情報が必要とされている時点での選別
    - ▹例えば，図書館での所蔵検索のように，すでにある文書群から知りたい事柄が記述されている文書を特定する
    - ▹検索要求は短期的
  - ○情報フィルタリング
    - ▹情報が得られた時点での選別
    - ▹例えば，新聞記事の中から興味のある記事だけを拾い読みするように，時々刻々と配信される情報から必要なものを取り上げる．

文書情報の組織化 (cont.)

- □分類
  - ○二つより多いグループに文書群を分類
    - ▹カテゴリ付与: あらかじめ与えられた分類体系に沿って文書を分類
    - ▹文書クラスタリング: 類似する文書をグループ化することによって分類．グループに対する適切な名前付け．

文書情報の組織化 (cont.)

- □抽出
  - ○情報抽出，主題情報の抽出
    - ▹中心的な情報だけを抽出
    - ▹特定の記事カテゴリに対して抽出すべき情報がわかっていると仮定
    - ▹例:製品発表記事から「メーカ名」「発表年月日」「製品名」「価格」をぬき出せ
    - ▹より一般的な「あらゆる文書の中心的な情報を抽出せよ」は難しい
- □要約
  - ○文書の表す意味内容を非常に短いテキストで簡潔に表現する
  - ○抽出した情報を文章の形で表現することに相当

文書のモデル

- □分解不可能なオブジェクトとして扱う
  - ○文書の中身をシステムの処理対象としない，あるいは，文書自身の情報が得られない場合
  - ○書誌情報をつけ，それにより扱う(国大の所蔵検索など)
  - ○内容に関する検索はできない
- □長い文字列として扱う
  - ○最も基本的な扱い
  - ○文字列検索(完全一致，正規表現など)
- □一定の長さの文字列の列/集合/統計情報として扱う
  - ○n-gram モデル
- □人間にとって意味のある言語単位の列/集合/統計情報として扱う
  - ○「単位」=語(形態素)，句，文，談話,...

情報検索

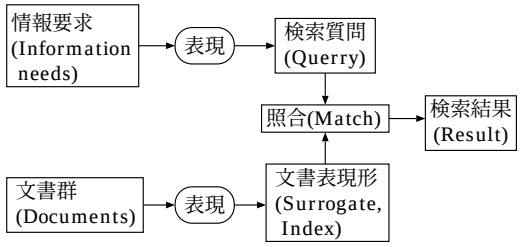
情報検索とは

- □広義
  - ○文書群において「知りたい」ことが記述されている文書を特定する
    - ▹利用者の未解決の問題を解決できる文書をみつけること
- □狭義
  - ○利用者の与えた検索質問文(query)に適合する(relevant)文書を見つける

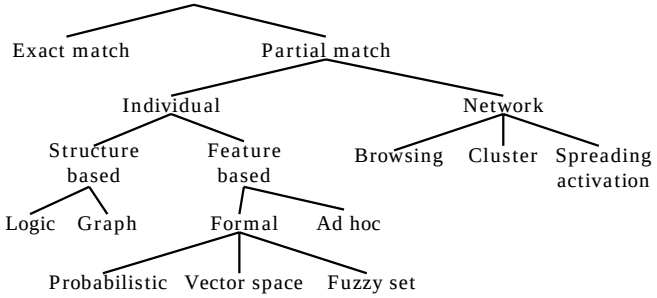
利用者の要求(Information needs)

- □直観的要求
  - ○存在するがはっきりした情報要求の形にまではいたっていないもの
- □意識された要求
  - ○意識されており頭の中で記述できるもの
- □具体化された要求
  - ○要求の具体化された記述
- □調整された要求
  - ○情報システムへの質問

情報検索の枠組



情報検索の技術の分類



伝統的なキーワード検索(Exact match)

- □一つの語，あるいは，複数の語を論理演算子で結合した論理式を検索質問として受け付ける．
- □論理式が真となる文書(の集合)を結果として返す．
- □主要論理演算子: ∧(and), ∨(or), ¬(not) (これらと括弧'()')
  - ○T (ある語): その語が索引語リストに含まれる文書で真，それ以外で偽
  - ○Q1∧Q2: 論理式Q1ならびにQ2が共に真となる文書で真，それ以外で偽
  - ○Q1∨Q2: 論理式Q1とQ2のうち少なくとも一方が真となる文書で真，それ以外で偽
  - ○ ¬Q: 論理式Qが偽となる文書で真，それ以外で偽
- □転置(インデクス)ファイルにより容易に適合文書を取り出せる

伝統的なキーワード検索(Exact match) (cont.)

欠点

- □質問に部分一致する文書をとりこぼす
- □検索した文書が順位づけされていない
- □語の重要性を扱えない
- □表現の語彙に影響されやすい

望ましい(最低限の)検索方式

「検索質問が与えられた時に語の重要性を考慮しつつ，文書の候補を順位をつけて表示する」

- □語の重要性を表現できる
  - ○索引語の抽出ならびに重みづけ
  - ○TF・IDF 法
- □検索された文書の順位づけ
  - ○類似度を定義できるモデルで文書と質問を表現
  - ○ベクトル空間モデル