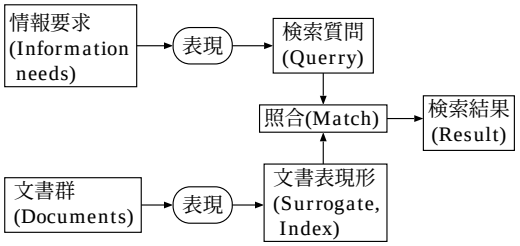
横浜国立大学 大学院 環境情報学府
情報メディア環境学専攻(前期)

言語情報処理原論(9)

Foundation of Natural Language Processing (9)

森 辰則
mori@forest.eis.ynu.ac.jp

情報検索の枠組



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

索引語の抽出

- それぞれの文書に対してどのような「語」を索引語として採用するか
  - 「語」: Term = 語，語基，句，概念 など
  - 文書 → Termの並び
- 情報検索における重要語
  - 例えば，Termに何らかの方法で重要度を付与し，その値がある閾値を越えた場合に重要語とする．観点は，
  - 包含性: ある文書においてその語が中心的話題であるか?
  - 弁別性: 他の文書にはない事柄(語)であるか?
- 不要語(stop word)
  - どの文書にも現れ，文書検索の役に立たない語．
  - これらは，前もって排除されることが多い．
  - 英語なら，冠詞,前置詞など．日本語なら，助詞，助動詞など．

TF・IDF法

- ヒューリスティックな方法(多くの場合に有効だが，常にうまくいくという保証はない)
- 文書Djにおける Term Tiの重要度w(i,j)

$$w(i,j) = TF(i,j) \cdot IDF(i)$$
$$IDF(i,j) = log(\frac{N}{DF(i)})(+1)$$

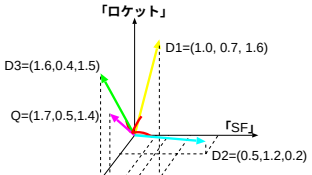
- TF(i,j): Term Frequency
  - 文書Djにおける Term Tiの出現回数
  - 包含性
- DF(i): Document Frequency
  - Term Ti を含む文書数
- IDF(i): Inverse Document Frequency(上式)
  - 弁別性

素性に基づく文書表現

- 素性: 何らかの特徴量
- 文書と質問を素性の組として表現
  - (f1,f2, ... ,fn)
- 素性の組の間の「類似度」を定義，その値により順位づけ
- 代表的な方法
  - ベクトル空間法(Vector Space Model)
- 類似度は実数値となるために，「与えられた検索要求文」と「複数の文書」の間の類似度を計算し，順序づけすることができる．

ベクトル空間法(Vector Space Model)

- 文書と検索質問の両者を同一空間上のベクトルとして表現
- ベクトル間に類似度を定義し，類似文書の順位づけをする
- 全文書の索引語Ti(i=1〜t)に線形独立なt個のベクトルViを対応させる．



ベクトル空間法(cont.)

- このベクトル空間において文書Djを以下のように文書ベクトルで表現

$$D_j = \sum_i w(i,j) V_i$$

  - w(i,j): 文書Djにおける Term Tiの重み
- 検索質問についても同様に

$$Q_s = \sum_i q(i,s) V_i$$

  - q(i,s): 検索質問Qsにおける Term Tiの重み
- 例えば最も簡単なベクトルは

語の重み

文書ベクトル

- □**最も簡単なもの**
  - ○w(i,j) = 1 (**文書j**にTerm Ti**が現れる場合**)
  - ○w(i,j) = 0 (**それ以外**)
- □**より高度なもの**
  - ○w(i,j) = TF(i,j)・IDF(i)

検索質問ベクトル

- □**最も簡単なもの**
  - ○q(i,s) = 1 (**文書j**にTerm Ti**が現れる場合**)
  - ○q(i,s) = 0 (**それ以外**)
- □**より高度なもの**

$$q(i,s) = (0.5 + 0.5 \cdot \frac{TF(i,s)}{\max_l TF(l,s)}) \cdot IDF(i)$$

ベクトル間の類似度

cosine**相関度**(correlation)

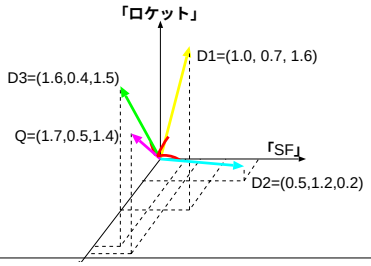
$$sim(D_j, Q_s) = \cos\theta_{j,s} = \frac{D_j \cdot Q_s}{|D_j||Q_s|}$$

- □**二つのベクトルのなす角の**cosine**値**
  - ○**二つのベクトルが同じ方向を向いている**
    - ▸cosine**値が** 1 **に近くなる**
  - ○**二つのベクトルが異なる方向を向いている**
    - ▸cosine**値が** 0 **に近くなる**(w(i,j),q(i,s)**が全て同じ符合の場合**)
- □**類似度にベクトルの長さは反映されていない**
  - ○**語の強さの相対的な分布により類似度がきまる**

ベクトル空間法 (cont.)

例の場合の類似度

- □D3**と**Q: 0.99
- □D1**と**Q: 0.94
- □D2**と**Q: 0.58



ベクトル空間法 (cont.)

もう一つの例

文書1	文書2	文書3	文書4
. . 自動車 . .	ロケット . .	. . 鉄道 . .	. . 宇宙 . .
. . 鉄道 . .	. . 宇宙 . .	. . 自動車 . .	. . ロケット . .
. . ロケット . .	. . ロケット .		

(宇宙, ロケット, 自動車, 鉄道)

	頻度ベクトル	TF・IDFベクトル
文書1	(0, 1, 1, 1)	(?,?,?,?)
文書2	(1, 2, 0, 0)	(?,?,?,?)
文書3	(0, 0, 1, 1)	(?,?,?,?)
文書4	(1, 1, 0, 0)	(?,?,?,?)

- □tfidf**法により文書ベクトルを求めよ．類似度の高い文書はどれ**?
  - ○**ただし** log2(4/3) = 0.4 **とする．**

類似度計算に基づく検索システム

検索の流れ(基本形)

- □0) **あらかじめ各文書に対する文書ベクトルを計算**
- □1) **検索質問を検索質問ベクトルに変換**
- □2) **検索質問ベクトルとすべての文書ベクトルとの類似度を計算する**
- □3) **全文書を類似度の大きい順に整列**
- □4) **上位**M**位までの文書を出力**

すべての文書と類似度計算を行なうのは効率が悪い...

転置ファイル(Inverted file) による実装

転置ファイル

- □**辞書ファイルの一種**
- □**語→ その語が出現している文書(あるいは出現位置)**
- □**同時にその語に関する情報(語の重み**(TF・IDF**値)等)も保存**

Index file

語	文書数	ポインタ
宇宙	4	
ロケット	2	
SF	3	

Posting file

文書番号	語の重み
3	0.5
5	0.2
20	0.6
53	0.1
10	0.4
37	0.3
1	0.9
16	0.5

転置ファイルによる実装 (cont.)

- □0) **あらかじめ転置ファイルを作成．文書ベクトルの長さが** 1**になるように語の重みを正規化．**Inverted file**の各項目は以下の対応．**
  - ○Ti → w(i,n),...,w(i,m),...,w(i,l)
  - ○**ただし，**w(i,m)**は，文書**m**における語**Ti**の重みで，**0**でないもの**
- □1) **検索質問をベクトル**Qs**に変換する．**Qs**も正規化．**
  - ○Qs=(q(1,s),q(2,s),...,q(3,s))
  - ○**ただし，**q(i,s)**は，質問**s**における語**Ti**の重み**

転置ファイルによる実装 (cont.)

- □2) 0**でない**q(i,s)**すべてについて以下を計算する．**
  - ○Inverted file**を調べ，**Ti**に対応するすべての** w(i,m)**について以下の計算をする．**
  - ○Sm ← Sm + w(i,m)*q(i,s)
- □3) Sm**に類似度が得られる．**0**でないものを整列する．**

検索質問文に登場する語を含まない文書は全く計算しないので，文書数が多くなっても速度が低下しない．(辞書ファイルの作り方にも依存)

情報検索システムの性能評価

検索質問文に対して定義される，再現率と適合率が良く用いられる．

以下の定義において「適合文書」は「その検索質問文に適合する文書」

□再現率(recall) R

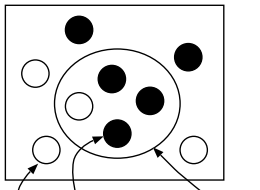
$$R = \frac{\text{検索された適合文書数}}{\text{全文書中の適合文書数}}$$

□適合率(precision) P

$$P = \frac{\text{検索された適合文書数}}{\text{検索された文書数}}$$

情報検索システムの性能評価(cont.)

再現率=3/5　適合率=3/4



適合文書　検索結果

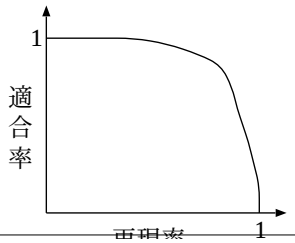
適合しなかった文書

情報検索システムの性能評価(cont.)

□理想は

- ○再現率=1.0
- ○適合率=1.0

□しかし，実際には再現率と適合率の間にトレードオフの関係



検索効率の改善

□再現率が低いとすると，利用者はすべての適合文書のうち一部だけしか得ることができない

- ▹検索要求の条件を利用者に緩めてもらう?
- ▹検索要求を変更してもらう? (語の選択のしなおし)
- ○検索条件を適応的に変化させる
  - ▹「適合性フィードバック」，「質問拡張」，「概念検索」

□適合率が低いとすると，利用者はかなりの数の無関係な文書を読まなければならない

- ▹検索要求の条件を利用者にきつくしてもらう?
- ○文書の，より細かい類似性を見つける
  - ▹「係受け情報の利用」，「共起情報の利用」

適合性フィードバック

□Rocchio他，多数の研究あり

□一度ではなく，数回の検索を繰り返す間に，徐々に結果を利用者の求めるものに近付けていく．

□検索が行なわれるたびに，利用者に検索された文書の評価をしてもらう(明示的/非明示的に)．

- ○(検索質問に対する)「適合」/「不適合」

□検索質問ベクトルを「適合」した文書のほうに移動すれば，検索結果の中の「適合」文書が多くなるはず．

適合性フィードバック (cont.)

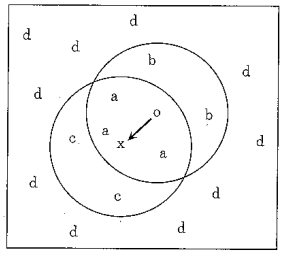


図 2.8　関連フィードバックの概念図
先の検索結果に対するユーザーのフィードバック情報を元に，検索質問ベクトルを修正し，該当文書の中心により近付けることを行なう．

適合性フィードバック (cont.)

質問拡張 + 再重みづけ による方法(Standard Rocchio)

$$Q = Q_0 + \beta \sum_{i}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{j}^{n_2} \frac{S_j}{n_2}$$

- ○Q0: 初めの質問に対応する検索質問ベクトル
- ○Ri: 適合文書Diの文書ベクトル
- ○n1: 適合文書の数
- ○Sj: 適合しない文書Djの文書ベクトル
- ○n2: 適合しない文書の数
- ○β，γ: パラメタ

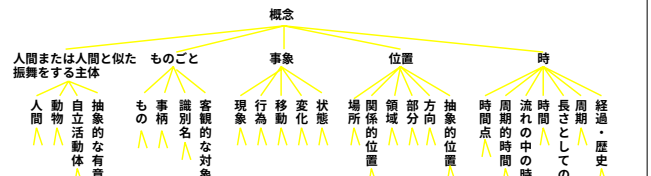
質問拡張(Query Expansion)

索引(辞書)にない言葉が検索質問に現れたら?

□無視してしまう方法では再現率が落ちる

□索引に現れる言葉のうち，質問中の語と「関連する」語を検索質問に加える

□シソーラス(類義語辞書)の活用



概念検索(Concept based IR)

□検索質問文における表層表現だけではなく，その「意味」(概念)をあつかう．

□(外部から与える)シソーラスによる質問拡張は概念検索の第一歩

○ポイントは語と語の間の類似性を導くこと

□シソーラス自動構築には，文書集合全体の情報を使い，語と語の間の類似度を求める

○「出現する文書(複数)が同じであるほど，類似度が高い」

- LSI
- CLARIT(Concept Base)
- Concept based Query Expansion

○「周りに現れる語の傾向が同じであるほど，類似度が高い」

- InfoMap

概念検索(Concept based IR) (cont.)

例:「出現する文書が同じであるほど，類似度が高い」

文書1	文書2	文書3	文書4	文書5
..宇宙..	..自動車..	..宇宙..	..鉄道..	..宇宙..
..ロケット..	..鉄道..	..ロケット..	..自動車..	..ロケット..
	..ロケット..			

	文書1	文書2	文書3	文書4	文書5
語ベクトル	「宇宙」	(1,0,1,0,1)			
	「ロケット」	(1,1,1,0,1)			
	「自動車」	(0,1,0,1,0)			
	「鉄道」	(0,1,0,1,0)			

類似度が高い組は 「宇宙」と「ロケット」
「自動車」と「鉄道」

検索効率の改善(再掲)

□再現率が低いとすると，利用者はすべての適合文書のうち一部だけしか得ることができない

- 検索要求の条件を利用者に緩めてもらう?
- 検索要求を変更してもらう? (語の選択のしなおし)

○検索条件を適応的に変化させる

- 「適合性フィードバック」，「質問拡張」，「概念検索」

□適合率が低いとすると，利用者はかなりの数の無関係な文書を読まなければならない

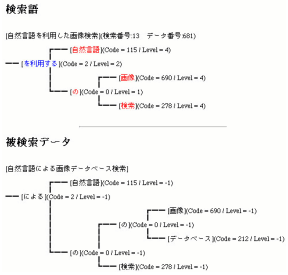
- 検索要求の条件を利用者にきつくしてもらう?

○文書の，より細かい類似性を見つける

- 「意味構造(係受け情報)の利用」，「共起情報の利用」

意味構造の利用(例:国立情報学研究所(旧学術情報センター))

意味ネットワークで文を表現．係受け構造の一致をみる．

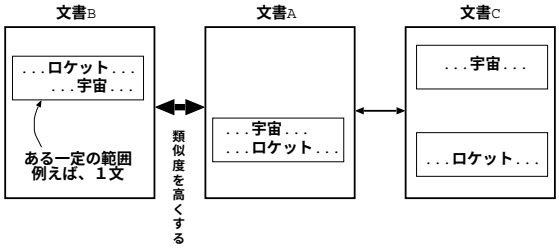


語の関係を利用する(1)

□語の共起

□語彙連鎖

語の共起(近距離に現れる語の組)



語の関係を利用する(2)

□語彙連鎖(同じや関連のある語の繰り返し)