横浜国立大学 大学院 電子情報工学専攻 専門科目

森 辰則

mori@forest.dnj.ynu.ac.jp

ディジタルドキュメント論(9)

Digital Document Processing

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角手法

□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捜す

▹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「発表年月日」「製品名」「価格」をぬき出せ

▹例：複数の文書から中心的な情報を抽出し要約

□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

○「単語」＝語(形態素)，句，文，談話

□「タグ」を埋め込み構造を表現する

○HTML, XMLなど

以下では文書を「語の列」として扱う．

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

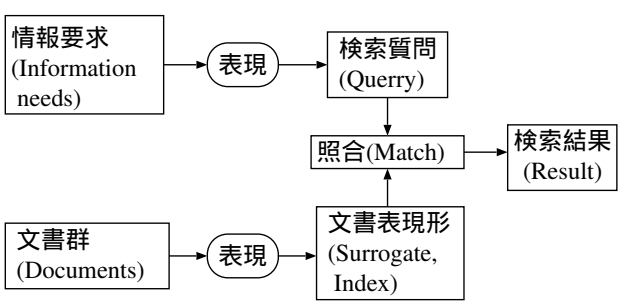
□具体化された要求

○要求の具体化された記述

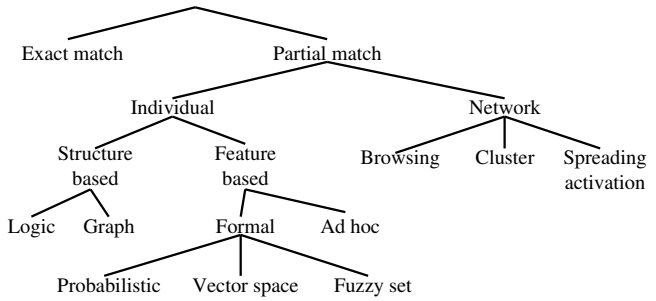
□調整された要求

○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.)

例
□「英語以外の言語に対する，情報検索や情報抽出に関する文献」
□¬ 英語 ∧ (情報検索 ∨ 情報抽出)
□問題点
○「英語以外の言語」⇔ ¬ 英語
○「情報検索に関する文献」⇔情報検索

情報要求(概念)と検索質問(語の組合せ)の間の隔たり

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
□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
○索引語の抽出ならびに重みづけ
□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．
□語の頻度情報を利用したTF・IDF法が広く用いられる

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

TF・IDF値は索引語の重みとしても利用される(後述)

TF・IDF法 (cont.)

TFのバリエーション

$$cfreq_{ij} = K + (1 - K)\frac{freq_{ij}}{maxfreq_j} \qquad (Croft\ 1983)$$

$$nfreq_{ij} = \frac{\log_2 (freq_{ij} + 1)}{\log_2 length_j} \qquad (Harman\ 1986)$$

where

$freq_{ij}$ = the frequency of term i in document j
$maxfreq_j$ = the maximum frequency of any term in document j
$length_j$ = the number of unique terms in document j

TF・IDF法 (cont.)

IDFのバリエーション(1)

$$IDF_i = \log_2 \frac{N}{n_i} + 1 \qquad (Sparck\ Jones\ 1972)$$

$$IDF_i = \log_2 \frac{maxn}{n_i} + 1 \qquad (Sparck\ Jones\ 1979)$$

$$IDF_i = \log_2 \frac{N - n_i}{n_i} \qquad (Croft\ and\ Harper\ 1979)$$

where

N = the number of documents in the collection
n_i = the total number of occurrences of term i in the collection
$maxn$ = the maximum frequency of any term in the collection

TF・IDF法 (cont.)

IDFのバリエーション(2)

$$normalized\ noise_i = maxnoise - noise_i,$$

$$noise_i = \sum_{k=1}^{N} \frac{Freq_{ik}}{TFreq_i} \log_2 \frac{TFreq_i}{Freq_{ik}} \qquad (Harman\ 1986)$$

$$entropy_i = 1 - \frac{\sum_{k=1}^{N} \frac{Freq_{ik}}{TFreq_i} \log_2 \frac{TFreq_i}{Freq_{ik}}}{\log_2 N} \qquad (Lochbaum\ \&\ Streeter\ 1989)$$

where

N = the number of documents in the collection
$maxnoise$ = the highest noise of any term in the collection
$Freq_{ik}$ = the frequency of term i in document k
$TFreq_i$ = the total frequency of term i in the collection

TF・IDF法 (cont.)

表 2.3　TF.IDF による索引語の抽出例

『岩波情報科学辞典』の全 4229 項目を全文書とした場合の，「情報学」項目と「情報検索」項目における，TF.IDF の値の上位 10 語．形態素解析には JUMAN 3.4 を用い，名詞(数詞を除く)と未定義語(カタカナ語等)のみを対象とした．

D=「情報学」				D=「情報検索」			
w	単語 (T)	tf	df	w	単語 (T)	tf	df
14.67	情　報	20	781	13.20	質　問	7	55
10.34	組織化	4	11	13.15	文　献	7	56
8.78	追　究	3	5	9.58	検　索	6	107
6.98	言　葉	4	76	8.85	キーワード	4	26
6.20	科　学	4	119	8.05	索　引	4	41
4.39	活　用	2	27	6.32	主　題	3	33
4.33	学　問	2	29	3.41	蓄　積	2	83

素性に基づく文書表現

□文書と質問を素性の組として表現

○(f1,f2,…,fn)

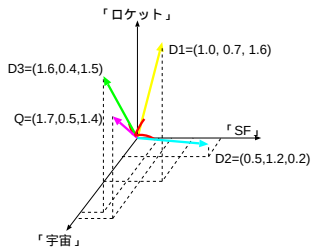
□素性の組の間の「類似度」を定義，その値により順位づけ
□代表的な方法

○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

$$(w(1,j), w(2,j), ..., w(t,j))$$

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

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

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
○cosの値が1に近いほど二つのベクトルが異なる方向を向いている

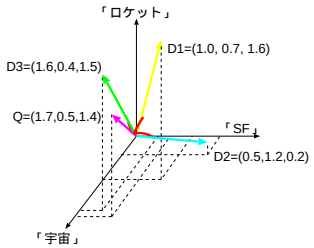
□類似度にベクトルの長さは反映されていない

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

(宇宙, ロケット, 自動車, 鉄道)

文書1 (1, 1, 0, 0)
文書2 (0, 1, 1, 1)
文書3 (1, 1, 0, 0)
文書4 (0, 0, 1, 1)
文書5 (1, 1, 0, 0)

- この例では簡単のために語の重みをすべて1にしている．
- 類似度の高い文書はどれ?

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

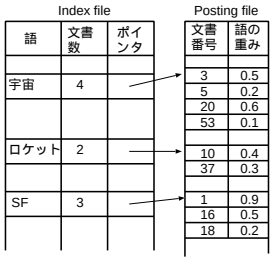
すべての文書と類似度計算を行なうのは効率が悪い...

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



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

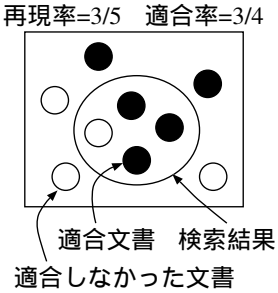
$$R = \frac{\text{検索された適合文書数}}{\text{全文書中の適合文書数}}$$

- 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.)

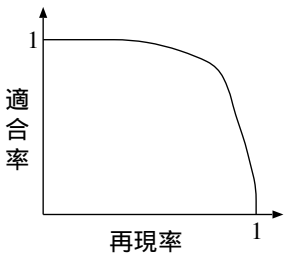
再現率=3/5　適合率=3/4



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，利用者はがなりの数の無関係な文書を読まなければならない
- ○文書要求の表現形式に類似性を見つける
  - ▹「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，(多数の研究者)
□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.)

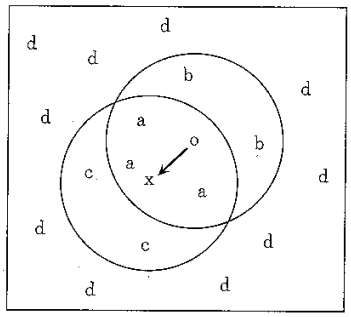


図 2.8　関連フィードバックの概念図
先の検索結果に対するユーザーのフィードバック情報を元に，検索質問ベクトルを修正し，該当文書の中心により近付けることを行なう

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)

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

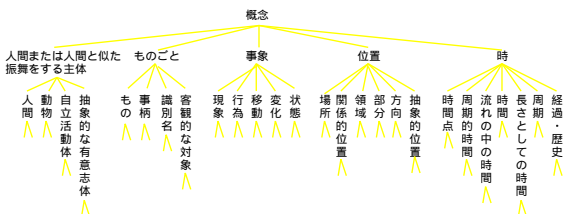
- ○Q0: 初めの質問に対応する検索質問ベクトル
- ○Ri: 適合文書Diの文書ベクトル
- ○n1: 適合文書の数
- ○Sj: 適合しない文書Djの文書ベクトル
- ○n2: 適合しない文書の数
- ○β, γ: パラメタ

効果的なアルゴリズム

□適合文書のベクトルの成分 → 値が大きくなり強化される
□非適合文書のベクトルの成分 → 値が小さくなり薄まる

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



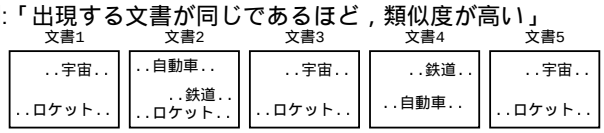
概念検索(Concept based IR)

□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」



語ベクトル	文書1	文書2	文書3	文書4	文書5
「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

○検索条件を適応的に変化させる

▹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(候補の選択のしなおし)

▹「適合性フィードバック」，「質問拡張」，「概念検索」

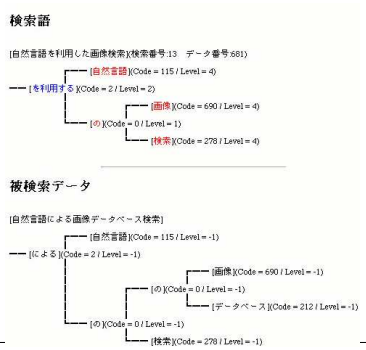
□適合率が低いとすると，利用者はがなりの数の無関係な文書を読まなければならない

○文書要求の表現力を細かくし，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．

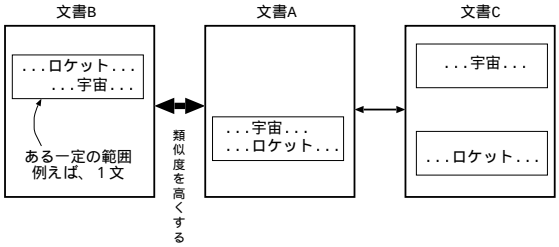


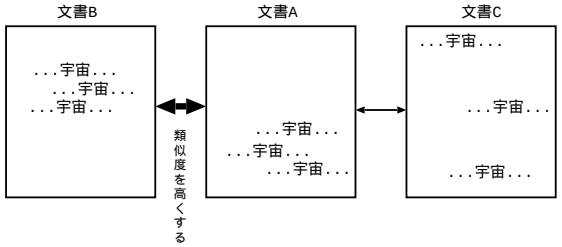
語の関係を利用する(1)

□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(同じ語の繰り返し)