ネットワークコンピューティング(1)
情報検索

関西学院大学理工学部情報科学科
北村泰彦

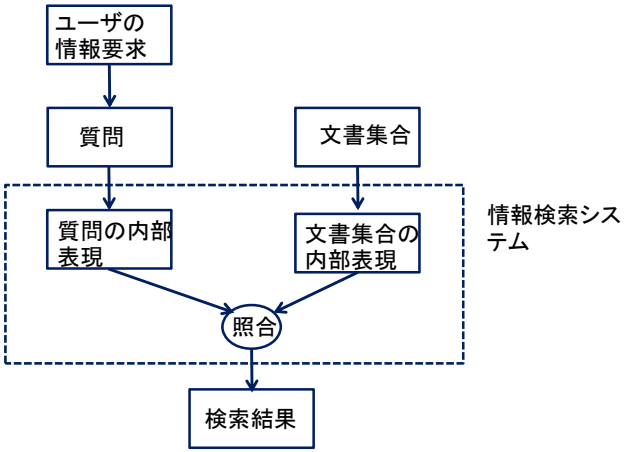
情報検索とは

- 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出すこと．
- ここでは，検索対象となる情報は，[　　](テキスト)を想定する．
- 検索の対象となる複数の文書を[　　　　](document collection)，ユーザの情報要求を[　　](query)と呼ぶ．

情報検索とは

- 代表的な情報検索システムに[　　　　　　　　]がある．
  - 文書集合：インターネット上に存在するWebページ
  - 質問：ユーザの与えるキーワード
  - 機能：キーワードを含むWebページを探し出す．

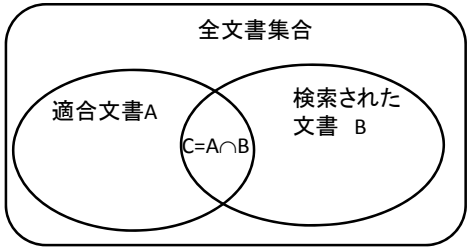
情報検索のモデル



データベースとの違い

- データベースシステムは情報検索システムである.
  - 文書集合:データベース
  - 質問:SQL文などの検索式
  - 機能:検索式を満たすデータをデータベースより抽出する
- ただし，データベースで扱うデータは一般的に（複数の属性値から構成されている）されている．情報検索システムで扱うデータは文書（テキスト）であり，

情報検索の評価尺度



- (recall):検索の完全性を評価するための尺度で，適合文書のうち，検索された文書の割合を示す．検索漏れの少なさを示す尺度.
- あるいは（precision）:検索の正確性を評価するための尺度で，検索された文書のうち，適合文書の割合を示す．検索ノイズの少なさを示す尺度.

情報検索の評価尺度

- 10000件の文書集合の中に，関西学院大学に関する文書は100件あるとする．検索質問文「関西学院大学」で検索したところ，200件の文書が得られたが，その中で関西学院大学に関する文書は50件であった.
- その時の再現率は である.
- その時の適合率は である.

Web情報検索の歴史

- 第一世代:人手による情報収集
- 第二世代: による情報収集
- 第三世代: による選別強化

第一世代

- 人海戦術：ウェブページを選別収集し，それを内容に応じてカテゴリに分類し，[　　]（データベース）に登録する．
- 「Yahoo!カテゴリ」(http://dir.yahoo.co.jp/)やOpenDirectory(http://www.dmoz.org)が代表例．他の多くの検索エンジンはサービスを終了している．
- 人間の判断に基づいて選りすぐりのページを集めることができる反面，労力の限界から，規模拡大や情報鮮度の維持が難しい．

ディレクトリ

- 人手で作った，ウェブページの索引集

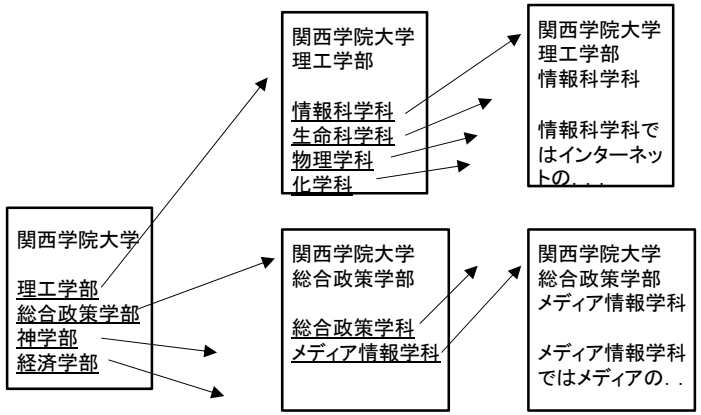
第二世代

- [　　](Crawler)：インターネットを巡回して，ウェブページを自動収集するシステム．[　　]，[　　]とも呼ばれる．
- 当初のLycos, AltaVista, gooなどが代表例
- 検索エンジンの規模，カバー率は拡大されたが，検索キーワードに多数のページがヒットし，必要な情報が埋もれてしまう．情報検索の[　　]は向上したが，[　　]は低下した．

Web情報検索の手順

- Webページの収集：クローラ
- 索引語の抽出
  - 形態素解析
  - 不要語の除去
- 索引語の重み付け
  - 局所的重み付け：索引語頻度(term frequency: TF)
  - 大域的重み付け：文書頻度(document frequency: DF)
- 文書検索
  - ベクトル空間モデル

ウェブページ中のそれぞれのリンクに対して，その先のページ収集を再帰的に行う．



- 文書を構成する文字列を単語に分割し，各単語に品詞や語形変化などの情報を与える処理．

検索	ケンサク	名詞-サ変接続
エンジン	エンジン	名詞-一般
の	ノ	助詞-連体化
仕組み	シクミ	名詞-一般
と	ト	助詞-並立助詞
技術	ギジュツ	名詞-一般
の	ノ	助詞-連体化
発展	ハッテン	名詞-サ変接続

不要語の除去

- (stop word)：文書を特定する能力が低く，索引語として適当でない単語．
  - 日本語の助詞：「は」，「が」など
  - 英語の冠詞：“a”, “the”など
  - 英語の前置詞：“at”, “of”, “in”など

索引語の重み付け

- 索引語の中には文書の内容と密接に関係したものもあるし，そうでないものも存在する．文書の内容を示す上での重要度を表す重み付けを検索語に行うと精度の高い検索が可能になる．
- 局所的重み：l_{ij}
  - 索引語w_iの文書d_jにおける出現頻度に基づき計算される重み．文書中に頻繁に出現する索引語に大きな値が与えられる．
  - 例： TF: f_{ij}(索引語w_iの文書d_jにおける出現頻度)

索引語の重み付け

- 大域的重み: g_i
  - 文書集合全体における索引語w_iの分布を考慮して決定される重み．特定の文書に集中して出現する索引語に対して大きな値が与えられる．
  - 例: [　　]の逆数IDF: $\log \frac{n}{n_i}$ ただしnは文書の総数，n_iは索引語w_iを含む文書数．対数化するのはIDFの値の変化を小さくするため．

[　　]

- 検索対象となる[　]を$d_1, d_2, \ldots, d_n$とする．これらの文書集合に対する[　　]を$w_1, w_2, \ldots, w_m$とする．このとき文書d_jを以下の[　　　]で表現される．

$$\boldsymbol{d}_j = \begin{bmatrix} d_{1j} \\ d_{2j} \\ \vdots \\ d_{mj} \end{bmatrix}$$

ここでd_{ij}は索引語w_iの文書d_jにおける重みである．

ベクトル空間モデル

- 文書集合全体は$m \times n$の[　　　]Dによって表現できる．

$$D = [\boldsymbol{d}_1 \quad \boldsymbol{d}_2 \quad \cdots \quad \boldsymbol{d}_n] = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix}$$

索引語・文書行列の各列は[　　　]であり，り，各行は[　　　]である．

ベクトル空間モデル

- 検索質問文に含まれる索引語w_iの重みをq_iとすると，[　　　]$\boldsymbol{q}$は以下のように表される．

$$\boldsymbol{q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{bmatrix}$$

ベクトル空間モデル

- 文書検索は検索質問ベクトルqと各文書ベクトルdの間の[redacted]を計算することで行う．類似度の計算にはコサイン尺度や内積が用いられる．
- コサイン尺度

$$\cos(\boldsymbol{d}_j, \boldsymbol{q}) = \frac{\boldsymbol{d}_j \cdot \boldsymbol{q}}{\|\boldsymbol{d}_j\|\|\boldsymbol{q}\|} = \frac{\sum_{i=1}^{m} d_{ij} q_i}{\sqrt{\sum_{i=1}^{m} {d_{ij}}^2} \sqrt{\sum_{i=1}^{m} {q_i}^2}}$$

- 内積

$$\boldsymbol{d}_j \cdot \boldsymbol{q} = \sum_{i=1}^{m} d_{ij} q_i$$

ベクトル空間モデル

- 索引語
  - w_1: Bioinformatics
  - w_2: Biology
  - w_3: Chemistry
  - w_4: Enzymes
  - w_5: Evolution
  - w_6: Genes
  - w_7: Genome(s)
  - w_8: Proteins

ベクトル空間モデル

- 文書
  - d_1: Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins
  - d_2: Proteins, Enzymes, Genes: The Interplay of Chemistry and Biology
  - d_3: Adaptive Evolution of Genes and Genomes
  - d_4: Advances in Genome Biology: Genes and Genomes
  - d_5: Bioinformatics and Genome Research
  - d_6: Data Analysis in Molecular Biology and Evolution

ベクトル空間モデル

- 索引語・文書行列（索引語の重みは頻度）

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

ベクトル空間モデル

- 検索質問文：Genes and Genomes

$$\boldsymbol{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

ベクトル空間モデル

- コサイン尺度を用いた場合の類似度計算

$$\cos(\boldsymbol{d}_1, \boldsymbol{q}) = \frac{1}{\sqrt{3}\sqrt{2}} = 0.408$$
$$\cos(\boldsymbol{d}_2, \boldsymbol{q}) = \frac{1}{\sqrt{5}\sqrt{2}} = 0.316$$
$$\cos(\boldsymbol{d}_3, \boldsymbol{q}) = \frac{2}{\sqrt{3}\sqrt{2}} = 0.816$$
$$\cos(\boldsymbol{d}_4, \boldsymbol{q}) = \frac{3}{\sqrt{6}\sqrt{2}} = 0.866$$
$$\cos(\boldsymbol{d}_5, \boldsymbol{q}) = \frac{1}{\sqrt{2}\sqrt{2}} = 0.5$$
$$\cos(\boldsymbol{d}_6, \boldsymbol{q}) = \frac{0}{\sqrt{2}\sqrt{2}} = 0$$

第三世代

- 第二世代の検索エンジンは検索精度が十分でなくなってきた．
  - ウェブページの数が膨大になった
  - ほとんどの検索語は1，2語であり，適合度を計算するための情報に乏しい．
  - ランキングの上位に位置することを意図した行為が存在する
- [　　]に基づく検索結果ランキング方式の導入
- Googleが代表例

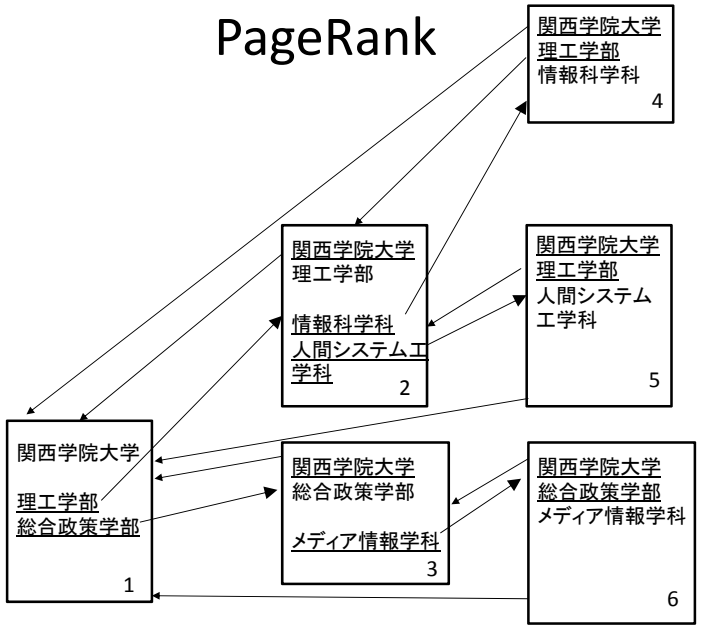
[　　]

- 「多数引用されているページは信頼できる．また，信頼できるページに引用されるページも信頼できる．」という考えに基づくスコア計算方式．
- PageRankは下式を繰り返すことにより求められる．

$$R(p) = \frac{\varepsilon}{n} + (1-\varepsilon) \cdot \sum_{(q,p)\in G} \frac{R(q)}{outdegree(q)}$$

- $R(p)$はページpのPageRank，nは対象とする（Webページをノードとし，それらのリンクをエッジとした）グラフGのノード総数（Webページ数），$outdegree(q)$はページqからの外向きリンク数．また$\sum_{p\in G} R(p) = 1$であり，εはdampening factorとよばれ，0.1~0.2の値を取る．
- ユーザは(1-ε)の確率で現在のWebページからのリンクをたどり，εの確率でまったく無関係なWebページにジャンプする．

PageRank



PageRank

- 初期値：R(1)=R(2)=R(3)=R(4)=R(5)=R(6)=1/6
- 1回目ε=0.1とする
  - $R(1) = \frac{0.1}{6} + 0.9 \cdot \left(\frac{\frac{1}{6}}{3} + \frac{\frac{1}{6}}{2} + \frac{\frac{1}{6}}{2} + \frac{\frac{1}{6}}{2} + \frac{\frac{1}{6}}{2}\right) = 0.367$
  - $R(2) = \frac{0.1}{6} + 0.9 \cdot \left(\frac{\frac{1}{6}}{2} + \frac{\frac{1}{6}}{2} + \frac{\frac{1}{6}}{2}\right) = 0.242$
  - $R(3) = \frac{0.1}{6} + 0.9 \cdot \left(\frac{\frac{1}{6}}{2} + \frac{\frac{1}{6}}{2}\right) = 0.167$
  - $R(4) = R(5) = \frac{0.1}{6} + 0.9 \cdot \left(\frac{\frac{1}{6}}{3}\right) = 0.067$
  - $R(6) = \frac{0.1}{6} + 0.9 \cdot \left(\frac{\frac{1}{6}}{2}\right) = 0.092$
- 得られた値を初期値として，この計算を繰り返し行う．

参考文献

- 北研二，津田和彦，獅々堀正幹：情報検索アルゴリズム，共立出版，2002
- 福島俊一：検索エンジンの仕組みと技術の発展，情報の科学と技術，54(2):66-71, 2004.
- 原田昌紀：WWWサーチエンジンの作り方，情報処理，40(11):1280-1283, 2000.