講義
「情報システム」,「情報科教育法II」
内容と担当者

- [配当学年]
  - 3年後期「情報システム」(情報学科)
  - 「情報科教育法II」(教育学部)
- [担当者]
  - 田中克己
    - tanaka@dl.kuis.kyoto-u.ac.jp 内線5385, 文学部東館465号室
    - ksato@dl.kuis.kyoto-u.ac.jp 内線5385, 文学部東館465号室,(教務補佐)
  - 田島敬史
    - tajima@i.kyoto-u.ac.jp 内線5385,文学部東館465号室
- [内容]
  - 情報システムを構築するための基礎となる理論および構築技術について講述. 特に, 情報検索・情報フィルタリングのモデルや基本的な手法, 物理的ファイル編成技術, Web 情報などに代表される半構造データ処理, Web情報検索とWebマイニングなどについて講述.

授業計画

- 第1回 (10/03) 情報検索(I) 適合率, 再現率, ベクトル空間モデル,類似検索 (田中)
- 第2回 (10/10) 情報検索(II) tf/idf法 , 適合フィードバック , クラスタリング (田中)
- 第3回 (10/17) 情報検索(III) 情報検索の評価尺度 (田中)
- 第4回 (10/24) 情報検索(IV) 協調フィルタリング, 推薦システム (田中)
- 第5回 (11/07) 情報システムの歴史 : ハイパーテキストから Webサービスまで (田島)
  - Dexterモデル, Smalltalk, HyperCard, SGML, HTML, スタイルシート,XML, Xlink, SMIL, SOAP, REST, Ajax
- 第6回 (11/14) XMLの基本 , XMLのための問合せ言語 (田島)
  - XPath , XQuery , XSLT , UnQL , 各言語のパラダイムの違い
- 第7回 (日程未定) XMLのためのスキーマ言語 (田島)
  - DTD, XML Schema, RELAX NG, 各言語の表現能力の違い
- 第8回 (11/28) XMLの問合せ処理 (田島)
  - 索引(DataGuide), Region Algebra, ノードラベリング方式, Join アルゴリズム, パス索引
- 第9回 (12/05) 副次索引 (田中)
  - 転置ファイル , B木 , グリッドファイル , k-D木 , シグニチャファイル
- 第10回 (12/12) 空間アクセス法 (田中)
  - Z-ordering, R木
- 第11回 (12/19) マルチメディア情報検索 (田中)
  - 画像検索, ビデオ動画像検索 , Gemini
- 第12回 (12/26) Web 情報検索(I) : ランキング (田島)
  - PageRank, VisualRankなど
- 第13回 (01/16) Web 情報検索(II) : コミュニティ発見と知識抽出 (田島)
  - HITS , Webマイニング
- 第14回 (01/23) Web 情報検索(III) : (田島)
- 第15回 (01/30) 試験

履修にあたっての注意

- [教材 (配付資料)]
  - 教材は講義ノート(Powerpoint)およびプリント配布
- [参考書]
  - C.Zaniolo, S.Ceri,C.Faloutsos,R.T.Snodgrass, V.S.Subrahmanian, R.Zicari, "Advanced Database Systems" Morgan Kaufmann Pub.Part IV -Spatial,Text and Multimedia Databases-
  - 鈴木, 中川, 福岡, 森, 細谷著, 「情報データベース技術」, 電気通信協会
  - David A. Grossman and Ophir Frieder, "Information Retrieval –Algorithms and Heuristics-", Kluwer Academic Publishers (1998)
- [予備知識]
  - データ構造, データベース, コンピュータネットワークに関する予備知識を有するのが望ましい。
- [評価]
  - 試験

情報学科CSコース情報システム(3年後期)

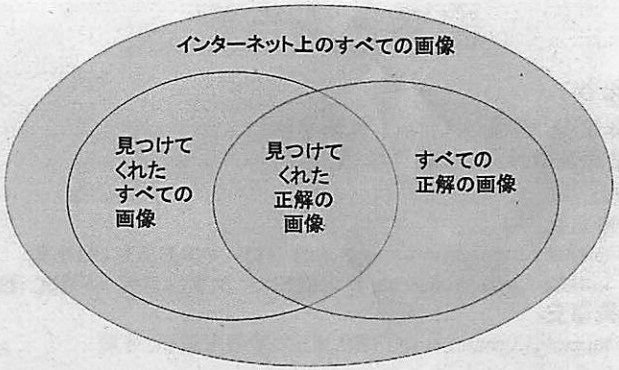
講義ノート
- 第1回 -

情報検索(I)
適合率, 再現率, ベクトル空間モデル, 類似検索

グーグルの画像サーチ

- ホームページの中の画像の周辺の文章や、画像ファイル名に、質問のキーワードが含まれているような画像を検索

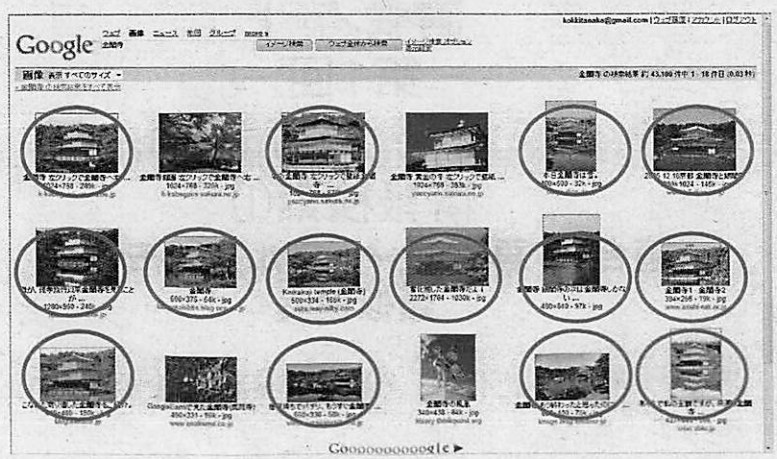
画像サーチの性能：どのぐらい正解の画像を見つけてくれるの？



- **再現率** $= \dfrac{\text{見つけてくれた正解の画像数}}{\text{すべての正解の画像数}}$
- **適合率** $= \dfrac{\text{見つけてくれた正解の画像数}}{\text{見つけてくれた画像数}}$

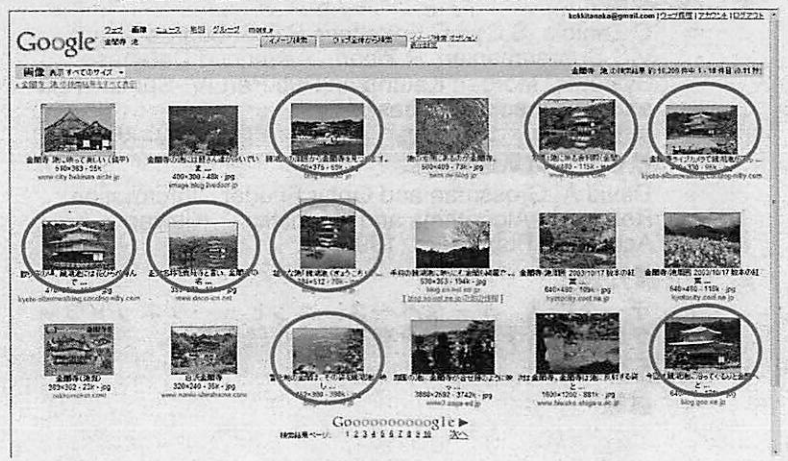
画像サーチの性能

- 金閣寺（らしい）画像を検索したい
- 正解は金閣寺と池が写っている写真(仮に正解数を50としよう)
- 質問キーワードは、「金閣寺」
- 1ページ目だけが見つけた画像とすると、再現率は14/50、適合率は14/18

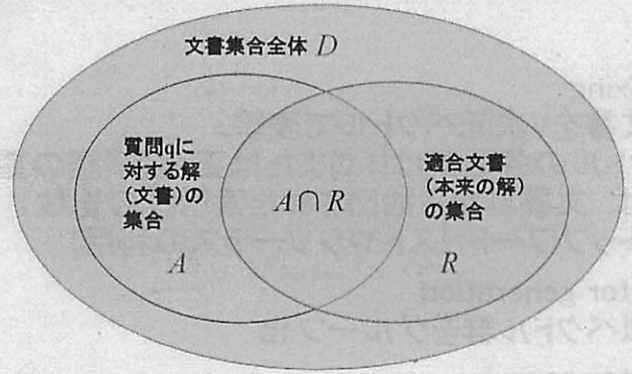


画像サーチの性能

- 金閣寺（らしい）画像を検索したい
- 正解は金閣寺と池が写っている写真(仮に正解数を50としよう)
- 質問キーワードは、「金閣寺　池」
- 1ページ目だけが見つけた画像とすると、再現率は8/50、適合率は8/18



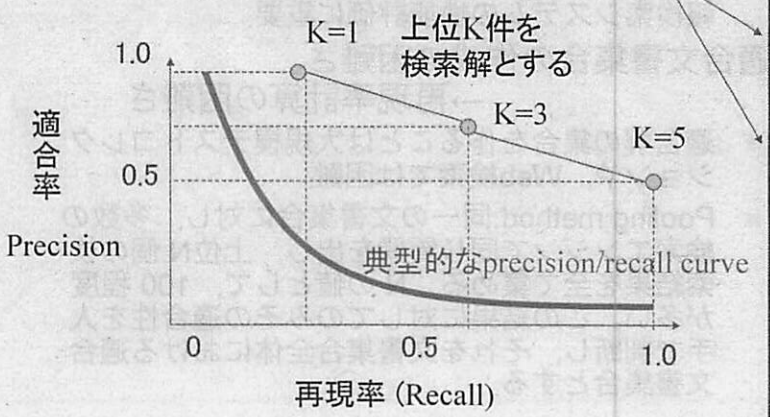
情報検索システムの評価尺度



- 再現率 $recall = \frac{|A \cap R|}{|R|}$
- 適合率(精度) $precision = \frac{|A \cap R|}{|A|}$

再現率 vs 適合率

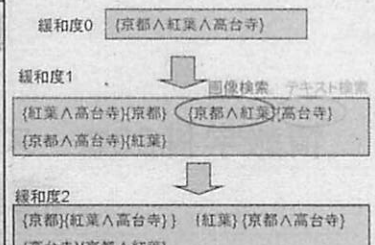
- 一般に，互いにtrade-offの関係
- Precision/recall curve



検索で得られたデータ(ranking順)

d1
d2
d3
d4
d5
d6
d7
d8
d9
d10

ゆるめて画像サーチ
(質問緩和法による再現率向上)

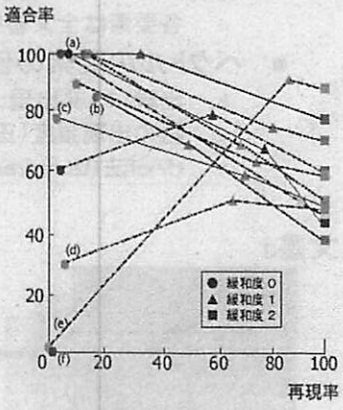


- 画像やビデオ検索に用いるキーワードを減らし，質問を緩和することで再現率を向上
- 画像やビデオ検索エンジンの検索結果のページで，テキスト部分に緩和されたキーワードを含むページを適合と判定
- ページ内から，検索キーワードを含む文章を抽出して統合
- GoogleImage，YahooVideoなど各種サーチエンジンに対応
- http://www.dl.kuis.kyoto-u.ac.jp/cc-society/index.html

質問緩和法

検索キーワード
「富士山」∧「夕日」∧「雪」

Google画像検索	Google テキスト検索	ヒット数	有効ページ
富士山、夕日、雪		0件	0件
富士山、夕日	雪	8件	6件
富士山、雪	夕日	12件	12件
夕日、雪	富士山	3件	3件
富士山	夕日、雪	4件	3件
雪	富士山、夕日	1件	1件
夕日	富士山、雪	0件	0件



テストコレクション

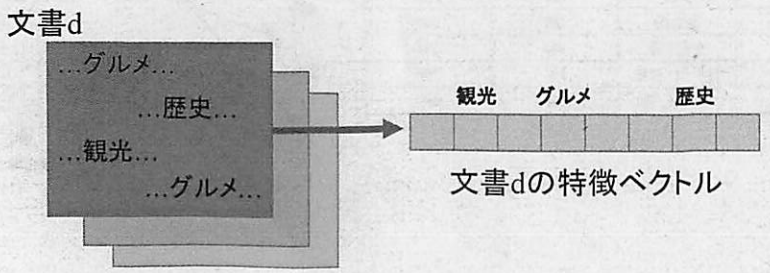
- テストコレクション
  - (a) 文書集合, (b) 多数の質問, (c)各質問に対する適合文書の集合を組にしたデータベース. 情報検索システムの性能評価に重要.
- 適合文書集合の作成の困難さ
  →再現率計算の困難さ
  - 適合解の集合を作ることは大規模テストコレクションや, Web検索では困難.
  - Pooling method:同一の文書集合に対し, 多数の検索エンジンで同じ質問を出し, 上位N 個の検索結果を全て集める. N の値として, 100 程度が多い。この結果に対してのみその適合性を人手で判断し, それを文書集合全体における適合文書集合とする.

ベクトル空間モデル(Vector Space Model)

- indexing
  各文書をV次元ベクトルで表現.
  ベクトルの各要素は{1,0}または正実数(語の重み)
  (Vは, 文書群から抽出された索引語の総数)
  (ストップワードリストやシソーラスの利用)
- cluster generation
  類似ベクトル群をグループ化
- cluster search
  質問(ベクトル)にもっとも類似のクラスタを検索

ベクトル空間モデルの特徴ベクトル

- 特徴ベクトルの各要素
  - 各要素は全文書集合から抽出した語(ターム)に対応
- ベクトルの要素の値の決定方法
  - 出現:1, 非出現:0
  - 語の出現頻度(正規化)
  - tf/idf法(term frequency/inverse document frequency)



情報学科CSコース情報システム(3年後期)

講義ノート

- 第2回 -

情報検索(II)

tf/idf法, 適合フィードバック, クラスタリング

tf/idf法(1)

- 語出現頻度(term frequency: tf)

$$tf_{ij} = freq(i,j)$$ 文書D_iにおけるタームt_jの出現頻度

$$tf_{ij} = K + (1-K)\frac{freq(i,j)}{\max_{i,j}(freq(i,j))}$$

語の最大出現頻度数で正規化

$$tf_{ij} = \frac{\log(freq(i,j)+1)}{\log(\text{文書 j 中の総ターム種類数})}$$

小さい語彙数の文書で特定の語の出現回数が大きな場合、tf値を高める

tf/idf法(2)

- 文書頻度 document frequency
  df_j = タームt_jが出現する文書数
- 実際はその逆のinverse document frequencyを使う.
  文書総数Nによる正規化

$$idf_j = \log\frac{N}{df_j}$$

- 文書D_iのタームt_jの重み $w_{ij} = tf_{ij} \times idf_j$

類似度(1)

- 文書D_jの特徴ベクトル

$$D_i = (w_{i1}, w_{i2}, \ldots, w_{in})$$

- 質問Qの特徴ベクトル
  - タームt_jを含めば1，含まなければ0という値からなるベクトル

$$Q = (w_{q1}, w_{q2}, \ldots, w_{qn})$$

- nは文書集合における全ての異なるターム数

類似度(2)

- 内積

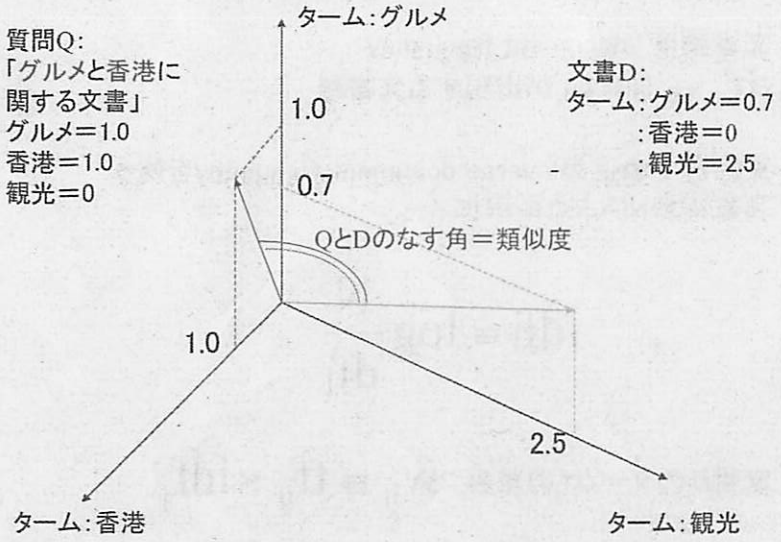
$$sim(Q, D_i) = w_{q1}w_{i1} + \ldots + w_{qn}w_{in}$$

- コサイン相関値

$$sim(Q, D_i) = \frac{w_{q1}w_{i1} + \ldots + w_{qn}w_{in}}{\sqrt{w_{q1}^2 + \ldots + w_{qn}^2} \times \sqrt{w_{i1}^2 + \ldots + w_{in}^2}} = \cos\theta$$

- 質問と文書の類似度
  文書と文書の類似度

質問と文書の類似度(コサイン相関値)



練習問題

- 質問Q: "gold silver truck"
- 文書
  - D1: "Shipment of gold damaged in a fire"
    D2: "Delivery of silver arrived in a silver track"
    D3: "Shipment of gold arrived in a truck"
- tf/idf法で各文書の特徴ベクトルを求めよ
  Qと各文書の類似度(内積, コサイン相関値)で求めよ.
- David A. Grossman and Ophir Frieder, "Information Retrieval – Algorithms and Heuristics–", Kluwer Academic Publishers (1998)から引用

	a	arrived	damaged	delivery	fire	gold	in	of	silver	shipment	truck
D1	0	0	.477	0	.477	.176	0	0	0	.176	0
D2	0	.176	0	.477	0	0	0	0	.954	0	.176
D3	0	.176	0	0	0	.176	0	0	0	.176	.176
Q	0	0	0	0	0	.176	0	0	.477	0	.176

内積の場合, ランキングはD2>D3>D1

パッセージ検索

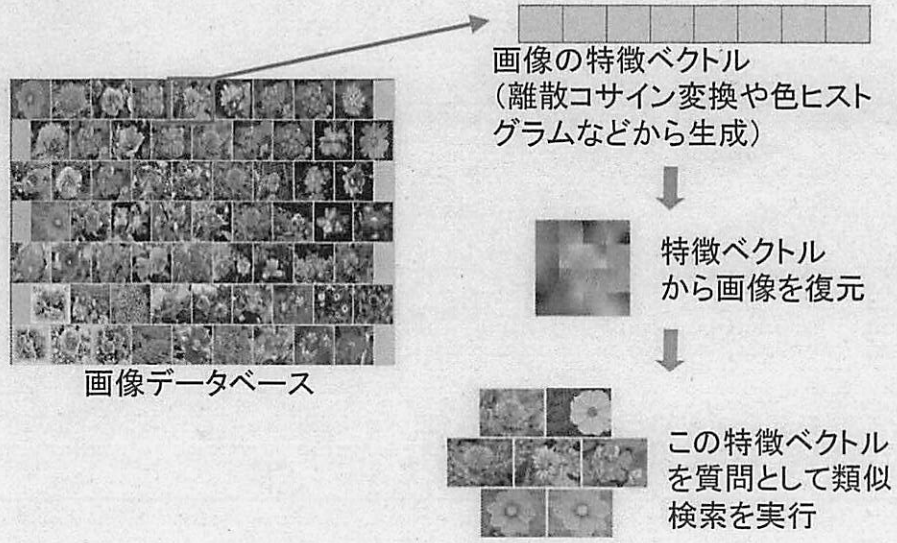
- パッセージ (passage)
  - 文書の内容を特徴付けるのは文書全体よりはむしろ特定の部分(段落など)
  - 文書Dの代わりにパッセージP1, ...,Pkの各特徴ベクトルと質問ベクトルとの類似度を計算しこれをマージする.
- パッセージの候補
  1 固定長に分割したテキストの部分
  2 形式段落
  3 形式的な節、章

適合フィードバック

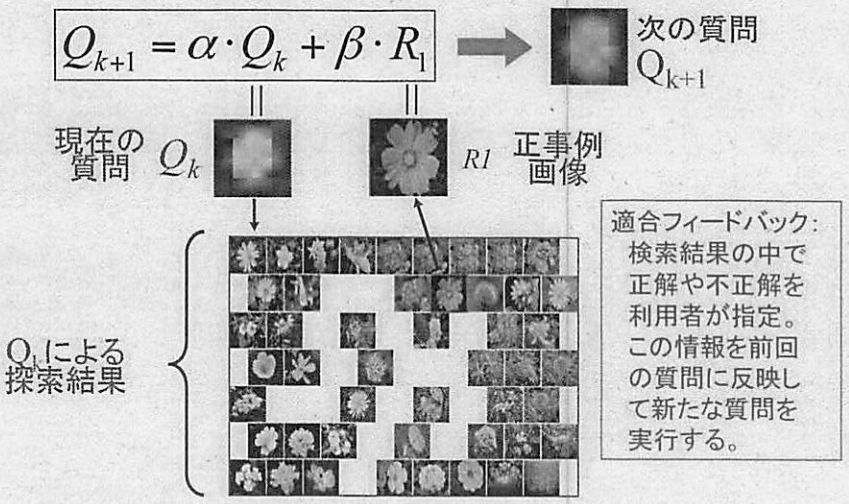
- Rocchio (1971)
  - ベクトル空間モデルでの適合フィードバック
  - Rocchio,J.J. "The SMART Retrieval System Experiments in Automatic Document Processing," chapter Relevance Feedback in Information Retrieval, pp.313-323, Prentice Hall.
  - Qは元の質問. Q'はQとユーザの反応から修正された質問. Qの検索結果集合の内, $R_1,..,R_{n1}$はユーザが適合と判断したもの, $S_1,...,S_{n2}$は不適合と判断したもの. このプロセスを繰り返す. (各項に適当な重み α, β, γを付加)

$$Q' = Q + \frac{1}{n_1}\sum_{i=1}^{n1} R_i - \frac{1}{n_2}\sum_{i=1}^{n2} S_i$$

画像検索における適合フィードバック



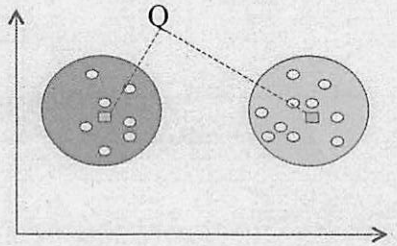
画像検索における適合フィードバック



次段階の 質問 Q_{k+1} は正事例画像を利用して現在の質問 Q_k を修正することによって生成される.

クラスタリング（clustering）

- 文書–クラスタ間の類似度（例えばコサイン相関値）
  - 文書とクラスタの中央値（centroid）
  - クラスタ内の文書との距離のうち最小のもの
  - クラスタ内の文書との距離のうち最大のもの



クラスタ生成

- 健全なクラスタ生成の方法
  - グラフ理論的アプローチ
  - 文書間の類似度がある閾値を超えたものを枝（edge）で結ぶ → 無向グラフ
  - 無向グラフ中の連結成分（connected component）または極大クリーク（clique，部分完全グラフ）を1つのクラスタとする．文書数Nに対してO(N×N)以上の計算量必要
- 反復法
  - サンプルから適当なクラスタ（seeds）作成．ある文書をそれに最も近いクラスタに追加．クラスタのセントロイドを修正，これを繰り返す．高速．