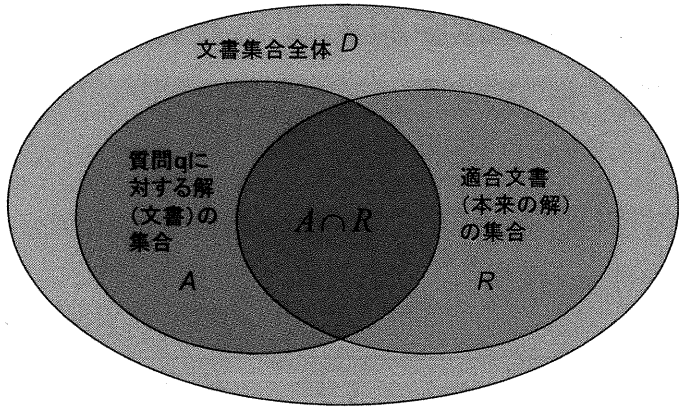
情報学科CSコース情報システム
（3年後期）
講義ノート
第3回

Web 情報検索(III)
情報検索の評価尺度

田中克己
ktanaka@i.kyoto-u.ac.jp
http://www.dl.kuis.kyoto-u.ac.jp

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) $precision = \frac{|A \cap R|}{|A|}$

適合率-再現率曲線
Precision-Recall Curve

i	$p(r_i)$	r_i
0	100%	0%
1	100%	10%
2	100%	20%
3	100%	30%
4	80%	40%
5	80%	50%
6	71%	60%
7	70%	70%
8	70%	80%
9	62%	90%
10	62%	100%

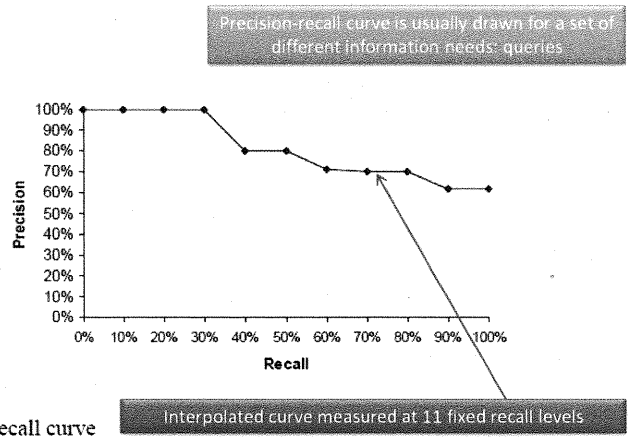


Fig. 6.4. The precision-recall curve

B. Liu: "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Springer Berlin Heidelberg, 2009

適合率-再現率曲線:
異なる情報検索システムの評価

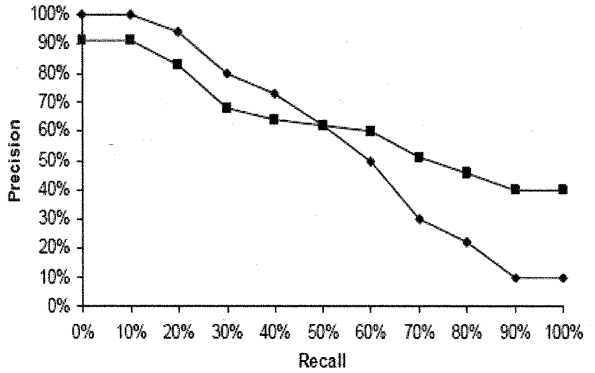


Fig. 6.5. Comparison of two retrieval algorithms based on their precision-recall curves

B. Liu: "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Springer Berlin Heidelberg, 2009

F値（F-score）

- 適合率Pと再現率Rを組み合わせた一つの尺度
- 調和平均 $F = \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2PR}{R+P}$
- F値
  - 重み付き「調和平均」（weighted harmonic mean）
  - 通常 $\beta=1$, $\alpha=0.5$

$$F = \frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P+R}$$

単なるPとRの「算術平均」は、PまたはRが大きい値の時に使えない。Rがほとんど1ならPはほぼ0となり、算術平均は0.5となるが、調和平均=2PR/(R+P)=2P/(1+P)=0となる。

Mean Average Precision: MAP

平均適合率(average precision)

- 再現率を動かした場合の適合率の平均
  「i個目の正解文書が初めて出てくるまでのランキングリストの適合率」の平均
- m個の正解文書
- $P(RankSet_i)$: 最上位からある正解文書$\mathbf{d}_i$までのリスト中に含まれる正解文書数の割合

$$MAP(q) = \frac{1}{m}\sum_{i=1}^{i=m} P(RankSet_i)$$

MAPは、平均適合率の平均

- クエリ$q_1, q_2, \ldots, q_n$に対する$MAP(q_i)$の平均

ランキング付き検索(Ranked Retrieval)の適合率と再現率

再現率を10,20,...,100%に固定して、その11点の適合率を計算

- 11-point interpolated average precision over fixed recall levels

Web サーチでは適合率のみが実際には使える

- 再現率はイントラネット（intranet)サーチでは有効

P@k (precision@k)

- Webサーチでは、検索結果の上位k件中の正解数が用いられる。P@5, P@10, P@15 など
- 通常、kは30〜50まで。（ユーザはそれ以上下位の検索結果は見ないため）

$$recall\ @\ k = \frac{|\text{正解集合} \cap \text{上位}\ k\text{件の解集合}|}{|\text{正解集合}|}$$

$$precision\ @\ k = \frac{|\text{正解集合} \cap \text{上位}\ k\text{件の解集合}|}{|\text{上位}\ k\text{件の解集合}|}$$

nDCG (Normalized Discounted Cumulative Gain)

- 文書の適合度を点数に置き換えて検索順位の重みをかけた指標
- ゲイン値(gain value): 適合度を点数化した値
  - （例）highly relevant 10, relevant 5, marginally relevant 1, nonrelevant 0
- ゲインベクトル(gain vector) G：ランク順の文書のゲイン値
  - （例）G= (5,10,0,5,1,10,0,0,.....)
- 累積ゲインベクトル(cumulative gain vector) CG：ランク順にゲイン値を累積
  - （例）CG=(5,15,15,20,21,31,31,31,...)

$$CG[k] = \sum_{i=1}^{k} G[i]$$

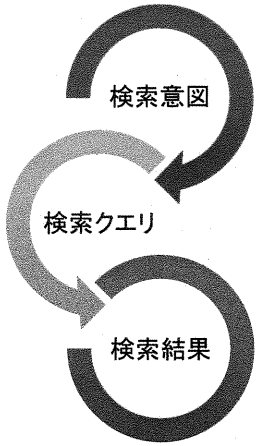
- ディスカウント累積ゲインベクトル (discounted cumulative gain vector)
  - 順位が下位の文書にペナルティを与える。（下位を見るのはコストがかかる）
  - （例）DCG=5.0,11,3,11.3,13.5,13.8,17.4,17.4,17.4,...)

$$DCG[k] = \sum_{i=1}^{k} \frac{G[i]}{\log_2(1+i)}$$

nDCG
(Normalized Discounted Cumulative Gain)

- 理想的なディスカウント累積ゲインベクトル DCG'
  (ideal discounted cumulative gain vector)
  - 理想的な順位が得られた場合のディスカウント累積ゲインベクトル
  - G' = (10,10,5,5,1,1,0,0,...)
  - CG' = (10,20,25,30,31,32,32,32,...)
  - DCG' = (10.0,16.3,18.8,21.0,21.3,21.7,21.7,21.7,...)
- 正規化ディスカウント累積ゲインベクトル：nDCG
  (normalized discounted cumulative vector)
  - DCGの各要素をDCG'の各要素で割り算することで正規化
  - （例）nDCG =(0.50,0.69,0.60,0.64,0.65,0.80,0.80,0.80,...)

$$nDCG\ [k] = \frac{DCG\ [k]}{DCG\ '[k]}$$



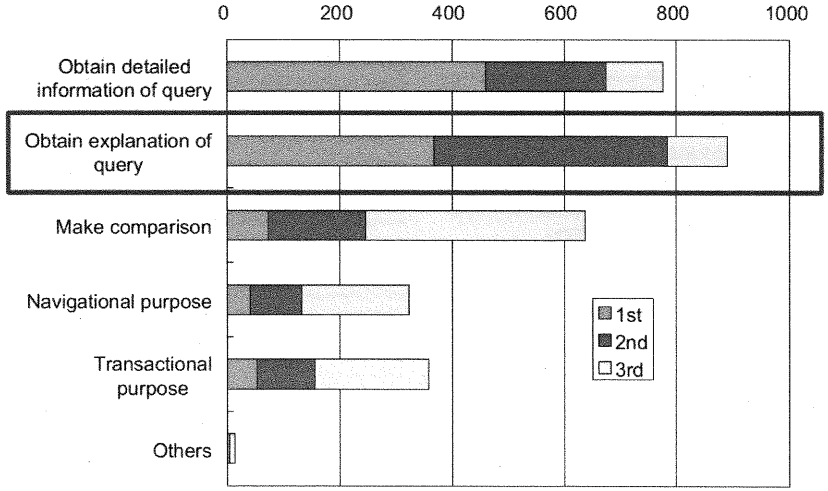
何のために検索するのか
― 検索の意図 ―

効果的(effective)な情報検索のための他の尺度

- ページの内容適合性(relevance)のみが、ユーザの満足を決定するファクターではない。
- 他のファクター:
  - Freshness (timeliness of results) 新鮮度
  - Credibility (=Trustworthiness + Expertise) 信頼性
  - Information coverage 網羅度, Diversity 多様性
  - Detailness (information depth) 詳細度
  - Typicality 典型度
  - Comprehensibility 理解容易性
  - Marginal relevance
  - etc.

Web検索の目的

- 知らないキーワードについて調べるためにWeb検索を用いる機会は多い。



何のために検索をするのか
―米ペンシルバニア州立大のB.J.Jansenらの研究―

3種の検索質問とその分布

Navigational型（10%）
- 少数特定のサイトに行き着くための質問

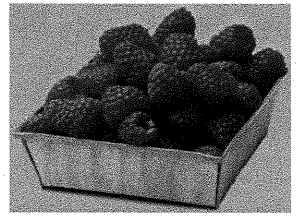
Informational型（80%）
- あるトピックに関して幅広く情報を集める質問

Transactional型（10%）
- 商品購入などのトランザクションを目的とするサイトへ行き着くための質問

与えられた質問の分類を高精度（精度74%）で行う手法を開発

阪急電車の時刻表を見る

京都大学の情報を収集



果物の購買発注

上位数件以外に どうせ見ないランキングリストの問題
― 検索結果の多様化（Result Diversification）の研究 ―

	Query: *sergey*		
	Google	Diver	MMR*
1	Sergey Brin - Wikipedia	Sergey Brin - Google Management	Sergey Brin - Wikipedia
2	Sergey Brin - Google Management	Sergey Korolyov - Wikipedia	Sergey Brin - Stanford
3	Sergey Brin - Stanford	Sergey Formin (at U. Mich.)	Sergey Brin (at forbe.com)
	Query: *hilton*		
	Google	Diver	MMR*
1	Hilton hotel	Hilton hotel (HHonors)	Hilton hotel
2	Hilton hotel online	Perez Hilton blog	Hilton hotel (at Germany)
3	Hilton hotel (HHonors)	Hilton Vacations Club	Hilton hotel online

- MMR (Maximal Marginal Relevance)
  - Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proc. of SIGIR Posters, 1998.
- DIVER
  - 米D.Rafiei（アルバータ大）, K.Bharat, A.（米Google社）の研究（WWW2010）
- 検索質問のみから検索の意図を検出するのは難しい。そこで、検索結果リストの上位に、多様な内容のページが並ぶようにランキング

Google 兵庫県立大学

兵庫県立大学 - トップページ
www.u-hyogo.ac.jp/ - キャッシュ
兵庫県立大学 公式ホームページ. ... 10月29日・30日に「**兵庫県立大学**・宮城大学合同シンポジウム」を開催します。 ■2011.10.7 神戸ポートアイランドキャンパスの施設を一般公開します。 ■2011.10.5 ・ 10月17日に「**兵庫県立大学**産学連携機構開設記念講演 ...
このページに 4 回アクセスしています。前回のアクセス: 09/05/20

兵庫県立大学 - Wikipedia
ja.wikipedia.org/wiki/兵庫県立大学 - キャッシュ
兵庫県立大学（ひょうごけんりつだいがく、英語: University of Hyogo）は、兵庫県神戸市西区学園西町8-2-1に本部を置く日本の公立大学である。2004年に設置された。大学の略称は定着したものはないが、「兵県大」と略記される場合が比較的多い。 ...

兵庫県立大学 合格目標偏差値の情報 | 進研ゼミ高校講座【定期テス...
shinken.zemi.ne.jp/hensachi/2199.html - キャッシュ
兵庫県立大学のお役立ち情報（**兵庫県立大学**の合格目標偏差値、先輩のクチコミなど）をお届けしています。**兵庫県立大学** 合格目標偏差値は学部、学科別にご紹介しています。**兵庫県立大学**に受かった先輩の合格体験記を是非ご覧ください。

【兵庫県立大学生活協同組合】
www.uhcoop.jp/ - キャッシュ
兵庫県立大学生活協同組合 ... 合同企業説明会. **兵庫県立大学**神戸学園都市キャンパス就職関連情報. 神戸・明石地区の住まい探しはこちら ... **兵庫県立大学**生協情報 **兵庫県立大学**生協学生委員会(工学部・理学部・環境人間学部)・大学生協阪神事業連合 ...

兵庫県立大学卓球部のHP
hyokentto.web.fc2.com/ - キャッシュ
兵庫県立大学卓球部のHPにようこそ(^^)/. 動画. 現在の閲覧者数: 動画. 新入生の皆さんご ... 卓球部は練習だけでなく他大学との交流や部員が企画した色々なイベントなどをしている活動的なクラブです。初心者からはじめた人もいます。 気軽に見学に来て ...

bing 兵庫県立大学

兵庫県立大学 - トップページ
www.u-hyogo.ac.jp
兵庫県立大学 公式ホームページ ... 平成23年4月、大学院シミュレーション学研究科（修士課程）を開設しました！ 平成23年3月22日、大学本部（神戸キャンパス）が移転しました。
受験生の方へ　交通アクセス
教育・入試情報　就職情報・学生生活
学部・大学院・研究所等　在学生・教職員向け
大学紹介案内　公募・案内等

兵庫県立大学工学部・大学院工学研究科
播磨工業大学は平成16年より**兵庫県立大学**工学部となって、新しい時代を迎えています。 ... 11月4日(金)に、はやぶさ小惑星探査に関する講演会を開催します <H23.10.21> 就業力育成支援学のホームページができました <H23.10.21>
www.eng.u-hyogo.ac.jp

兵庫県立大学 - Wikipedia
概観・沿革・基礎データ・教育および研究
兵庫県立大学（ひょうごけんりつだいがく、英語: University of Hyogo）は、兵庫県神戸市西区学園西町8-2-1に本部を置く日本の **公立大学** である。2004年 に設置された。 大学の略称 は定着したものはないが、「兵県大」と略記される場合が比較的多い。 全ての検索結果を表示した地図 - Google 全ての海域を示した地図 - Bing 全海域を出力 - KML Googleアース 全座標を出力 - GeoRSS 全 ...
ja.wikipedia.org/wiki/兵庫県立福祉大学

兵庫県立大学 - 教育・入試情報
兵庫県立大学 公式ホームページ　学部入試情報（入試日程等）　編成の連絡先および編集人員 大学入試センター試験において「地理歴史・公民」及び「理科」の成績利用方法について
www.[illegible].jp/ecu/index.html

兵庫県立大学　理学部　大学院　物質理学研究科　生命理学研究科
www.sci.u-hyogo.ac.jp/index-j.html　キャッシュ ページ

ランキングトップ5

検索結果の多様化:MMR

- 検索結果リストの上位に、多様な内容のページが並ぶようにランキング
  - ページと質問の類似度から、ページとリストに既出のページとの類似度を引き算した値の大きなものから並べる

$$MMR \stackrel{def}{=} \arg \max_{d_i \in R \setminus S} \left[\lambda (Sim_1(d_i, Q) - (1-\lambda) \max_{d_j \in S} Sim_2(d_i, d_j)) \right]$$

  - R: 検索結果のページ集合
  - S: すでにランキング表示で選ばれたページ集合
  - Sim1, Sim2:　文書やクエリの類似度

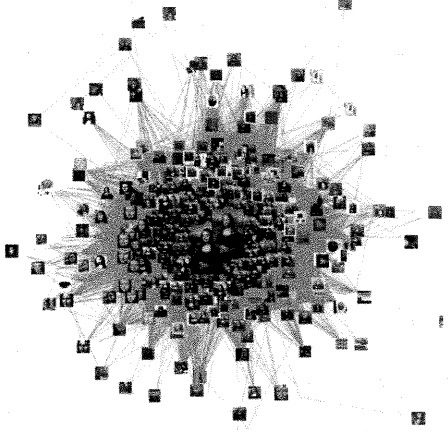
検索意図：
正解をどうやって決めているのか

		検索意図指標（正解指標）					
		内容類似性 (similarity)	人気度 (popularity)	典型度 (typicality)	多様性 (diversity)	理解容易性 (comprehensibility)	具体性 (concreteness)
検索性能評価尺度	適合率/再現率	・クエリと検索文書の内容が類似する文書 ・AltaVista等	・社会から高い評価を受けている文書 ・Google PageRank（参照重要度）	・全体の中で代表的・多数派的な文書	・重要な話題を網羅している文書	・理解しやすい文書	・内容が具体的な文書
	ランキングを考慮した尺度（MAP, nDCG等）						

典型性を重視した検索
[佃ら、WebDB2011フォーラム]

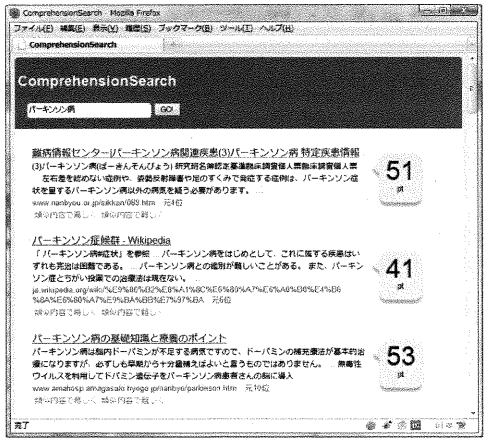
'典型的'な京都の観光地
（類似の観光地は少ないが、社会がそのように認知しているものが'典型的'）

'典型的'なモナリザの画像
（類似の画像が多い画像が'典型的'）



'典型的'な男（の中の男）
（類似の人物は少なく、社会の認知度も高いとは限らないが、その性格・行動等が男性らしさが傑出している人物が'典型的'）

理解容易性を重視した検索：
可読性と専門性



読みやすさ（可読性）と専門語・業界語の少ないページの検索とランキング
[Nakataniら：CIKM, DASFAA」

英語ページ		リンク先ページの可読性		
		容易	中	難解
リンク元ページの可読性	容易	**53.2%** **(3,594)**	42.1% (2,840)	**4.68%** **(316)**
	中	28.8% (3,233)	**56.8%** **(6,381)**	14.5% (1,630)
	難解	**18.8%** **(355)**	**47.8%** **(903)**	33.4% (631)

日本語ページ		リンク先ページの可読性		
		容易	中	難解
リンク元ページの可読性	容易	43.0% (15,601)	**49.9%** **(18,117)**	**7.13%** **(2,590)**
	中	3.37% (33,936)	**84.2%** **(846,576)**	12.4% (125,176)
	難解	**1.46%** **(6,605)**	34.6% (155,869)	**64.0%** **(288,468)**

[Akamatsuら：ACM WI2011]