LD-002

ひとつの高適合文書を高精度に検索するタスクのための評価指標

An Evaluation Metric for the Task of Retrieving One Highly Relevant Document with High Precision

酒井 哲也[†]
Tetsuya Sakai

1. はじめに

情報検索の分野では，ランクづけされた文書のリストを精度と再現率に基づき評価することが多い．例えば TREC や NTCIR[2] などで用いられている Average Precision (AveP) は，全ての適合文書を上位に (すなわち高再現率かつ高精度で) 検索するシステムに高いスコアを与える．しかし，例えば Web 検索のように再現率が重要でないかもしくは測定不能な大規模な検索環境では，高精度検索 (high precision search) が求められる．特に，適合文書を 1 件だけ高精度に検索することの実用上の意義は大きいであろう．なぜなら，例え複数の適合文書が存在する場合であっても，2 件目以降の適合文書に新規性 (novelty) がある保証はないので，現実には 1 件の適合文書の発見で満足する利用シーンが少なくないためである．質問応答の評価などに利用される Reciprocal Rank(RR) を高精度文書検索の評価に流用した研究もあるが [9]，厳密には，RR は適合文書を 1 件だけ高精度に検索するタスクのための指標である (2.3 節参照)．

しかし，高精度検索は近年の大規模検索環境における必要条件にすぎず，実際には部分適合文書よりも高適合文書をランク上位に出力する機能，すなわち高適合検索 (high relevance search) も同時に求められる．AveP や RR は二値適合性に基づくためこのようなタスクの評価には適さない．(NTCIR の Web タスクでは，RR を多値適合性向けに拡張した Weighted RR が提案されているが，実際には多値適合性は利用されておらず，部分適合文書も正解と見なした場合の RR と，見なさない場合の RR が別々に計算されている [3]．) そこで本稿では，高適合文書を 1 件だけ高精度に検索するタスクに適した評価指標 O-measure を提案し，以下の 2 点を明らかにする．

1. 全ての適合文書を高精度に検索するタスクと，1 件だけ高精度に検索するタスクの違い
2. 部分適合以上の文書を 1 件高精度に検索するタスクと，高適合文書を 1 件高精度に検索するタスクの違い

2 章で，本研究で扱う 4 つの情報検索評価尺度の関係について説明する．3 章では，各評価尺度により NTCIR-3 CLIR タスク [2] の参加システムを順位付けした場合の順位相関について考察する．4 章では，Buckley と Voorhees により提案された手法 [1, 11] を応用し，各評価尺度の安定性と判別能力について考察する．最後に，5 章において結論と今後の課題を述べる．

[†](株) 東芝 研究開発センター 知識メディアラボラトリー
tetsuya.sakai@toshiba.co.jp

2. 評価指標

2.1 Average Precision (AveP)

AveP は全ての適合文書を高精度に検索するタスクのための評価尺度であり，各検索課題に対して以下のように定義される．

$$AveP = \frac{1}{R} \sum_{1 \le r \le L} isrel(r) \frac{count(r)}{r} . \quad (1)$$

ここで，R は全適合文書数，$count(r)$ は検索結果上位 r 件中の適合文書数，$isrel(r)$ は第 r 位の文書が適合文書であるか否かを表すフラグ，L は検索結果のサイズである．$count(r)/r$ が上位 r 件における精度 (precision) を表すことは明らかであろう．本稿では，NTCIR CLIR タスクで用いられている Relaxed AveP(S,A,B 正解，すなわち高適合，適合，部分適合文書を全て正解と見なしたもの) を単に AveP と呼ぶ．

2.2 Q-measure

NTCIR-4 にて提案された Q-measure は，AveP を多値適合性向けに拡張したものと見なすことができ，AveP と非常に相関が高くかつ安定した指標である [5, 8]．例えば NTCIR CLIR タスクの多値適合性を用いる場合，理想的な検索システムは，全ての S 正解を最上位に検索し，その下に全ての A 正解を検索し，さらにその下に B 正解を検索する．このとき Q-measure は以下のように定義できる．

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \le r \le L} isrel(r) \frac{cg(r) + count(r)}{cig(r) + r} . \quad (2)$$

ここで，$cg(r)$ は cumulative gain と呼ばれる値であり [4]，第 i 位 $(1 \le i \le r)$ に X 正解 $(X \in \{S, A, B\})$ が検索される毎に $g(i) = gain(X)$ 点を加算することにより得られる．(本稿では gain values のデフォルト値として $gain(S) = 3, gain(A) = 2, gain(B) = 1$ を用いる．) また，$cig(r)$ は前述の理想的な検索結果に対する cumulative gain である．

Q-measure は ($L \ge R$ を満たす) 検索結果が理想的なものであるとき，またそのときに限って 1 となる．また，二値適合性の環境下においては，検索結果の上位 R 件より下位に適合文書を含まない場合に AveP と一致する．

2.3 Reciprocal Rank (RR)

RR は適合文書を 1 件だけ高精度に検索するための評価尺度と言える．検索結果が適合文書を含まない場合，$RR = 0$ とする．検索結果の第 r' 位に最初の適合文書がある場合，$RR = 1/r'$ とする．ここで，$1/r'$ は最初の適合文書に関する精度に他ならず，$R = 1$ なる検索課題に

表 1: O-measure values for systems that return exactly one relevant document for a topic s.t. $R(S)=1, R(A)=1, R(B)=1$.

	Relaxed AveP	Q-measure	RR	O-measure
(a) S at Rank 1	0.3333 (=(1/1)/3)	0.3333 (=(1/1)/3)	1 (=1/1)	1 (=4/4)
(b) A at Rank 1	0.3333 (=(1/1)/3)	0.2500 (=(3/4)/3)	1 (=1/1)	0.7500 (=3/4)
(c) B at Rank 1	0.3333 (=(1/1)/3)	0.1667 (=(2/4)/3)	1 (=1/1)	0.5000 (=2/4)
(d) S at Rank 2	0.1667 (=(1/2)/3)	0.1905 (=(4/7)/3)	0.5000 (=1/2)	0.5714 (=4/7)
(e) A at Rank 2	0.1667 (=(1/2)/3)	0.1429 (=(3/7)/3)	0.5000 (=1/2)	0.4286 (=3/7)
(f) B at Rank 2	0.1667 (=(1/2)/3)	0.0952 (=(2/7)/3)	0.5000 (=1/2)	0.2857 (=2/7)

ついては RR は AveP と一致することがわかる．すなわち，AveP が全適合文書の精度を見渡すのに対し，RR は最上位の適合文書の精度のみを調べる指標である．このため，RR は比較的不安定な指標であり，評価の際には検索課題を増やすなどの対策が必要となる [10].

2.4 O-measure

本稿では，高適合文書を 1 件だけ高精度に検索するための評価尺度として O-measure を提案する．検索結果が適合文書を含まない場合，$O\text{-}measure = 0$ とする．検索結果の第 r'位に最初の適合文書がある場合，

$$O\text{-}measure = \frac{cg(r') + count(r')}{cig(r') + r'} = \frac{g(r') + 1}{cig(r') + r'} \quad (3)$$

とする．すなわち，Q-measure が全適合文書の blended ratio $(cg(r) + count(r))/(cig(r) + r)$ を見渡すのに対し，O-measure は最上位の適合文書のみを調べる指標である．

Q-measure と同様の考察により [5]，O-measure が以下の性質を満たすことを容易に示すことができる．

- 最も適合レベルが高い文書のひとつがランク 1 位に検索された場合，かつその時に限り O-measure は 1 となる．
- 二値適合性の環境下では，上位 R 件以内に適合文書が含まれる場合，かつその時に限り O-measure は RR と一致する．
- 絶対値の小さい gain values を用いると，O-measure は RR と似た値をとる．

表 1 に，S,A,B 正解を各 1 件ずつ有する検索課題 $(R(S)=1, R(A)=1, R(B)=1)$ に対する 6 種類の検索結果に対する各評価尺度の計算例を示す．この課題に対する $cig(r)$ $(r = 1, 2, \ldots)$ は $(3, 5, 6, 6, \ldots)$ となる [4, 5]. 従って，例えば第 2 位に A 正解を 1 件だけ含む検索結果 (e) については $O\text{-}measure = (g(2)+1)/(cig(2)+2) = (2+1)/(5+2) = 3/7$ となる．この表から，O-measure が多値適合性の利用により検索結果 (a)-(f) を判別できることがわかる．

一方，O-measure は RR と同様にひとつの適合文書のみに着目した指標であるため，Q-measure などと比べて安定性が低いことが予想できる．

なお，$g(r')/cig(r')$ のような単純な指標では適切な評価が行えないことに注意する必要がある．例えば，$R = R(B) = 3$，従って $cig(r)$ が $(1, 2, 3, 3, \ldots)$ となる検索課題を考える．このとき，最初の適合文書を第 3 位にもつシステムと，最初の適合文書を第 1000 位にもつシステムの $g(r')/cig(r')$ は，ともに 1/3 になってしまう [5].

表 2: Kendall rank correlations for NTCIR-3 CLIR formal runs.

	Q-measure	RR	O-measure
(a) 45 C-runs			
Relaxed AveP	.9798	.6990	.6929
Q-measure	-	.6828	.6808
RR	-	-	.6828
(b) 33 J-runs			
Relaxed AveP	.9583	.7992	.7462
Q-measure	-	.8333	.7500
RR	-	-	.6742
(c) 24 E-runs			
Relaxed AveP	.9783	.7609	.7826
Q-measure	-	.7536	.7899
RR	-	-	.9058
(d) 14 K-runs			
Relaxed AveP	.9560	.8681	.8462
Q-measure	-	.8681	.8462
RR	-	-	.9780

表 3: Kendall rank correlations for NTCIR-3 CLIR formal runs with different gain values.

	O30:20:10	O0.3:0.2:0.1	O1:1:1	O10:5:1
(a) 45 C-runs				
RR	.5596	.7697	.9657	.5737
O-measure	.6869	.6909	.6768	.6687
(b) 33 J-runs				
RR	.6856	.8409	1.0000	.6591
O-measure	.7462	.7273	.6742	.7652
(c) 24 E-runs				
RR	.8768	.9275	.9928	.7826
O-measure	.8551	.8768	.9130	.7754
(d) 14 K-runs				
RR	.9121	1.0000	1.0000	.9121
O-measure	.8901	.9780	.9780	.8901

3. システム順位の相関

本章では NTCIR-3 CLIR タスクに提出された中国語，日本語，英語，韓国語文書の検索結果 (それぞれ C-runs, J-runs, E-runs, K-runs)[5, 8] を AveP, Q-measure, RR, O-measure により順位づけした場合の順位相関について考察する．(目下，これらは NTCIR から公開されている唯一のデータである．)

表 2 に NTCIR-3 CLIR タスクの公式提出結果を元に算出した各指標間のケンドール順位相関係数を示す．また，表 3 に O-measure の gain values を変化させた場合の RR およびデフォルトの O-measure とのケンドール順位相関係数を示す．例えば「O10:5:1」は $gain(S) = 10, gain(A) = 5, gain(B) = 1$ とした場合の O-measure を意味する．これらの結果より，以下のことがわかる．

- AveP/Q-measure(全ての適合文書に着目した指標) と RR/O-measure(ひとつの適合文書に着目した指標) との相関は比較的低い．すなわち，多くの適合文書を高精度に検索できるシステムと，1 件だけ高精度に検索できるシステムは必ずしも一致しない．
- RR と O-measure の相関も比較的低い．(K-runs については高いが，K-runs はデータサイズが最も小さい [5]．) すなわち，部分適合以上の文書を 1 件高精度に検索できるシステムと，高適合文書を 1 件高精度に検索できるシステムは必ずしも一致しない．
- O1:1:1 および O0.3:0.2:0.1 は RR との相関が高くなっており，これは理論的考察の結果と一致する (2.4 章)．
- Gain values の変更に対する O-measure の頑健性は，Q-measure の頑健性 [5] に比べると高くない．例えば表 3(a) における O10:5:1 と O-measure(すなわち「O3:2:1」) の相関は 0.6687 にとどまっている．

以上より，高適合文書を 1 件だけ高精度に検索するタスクには，O-measure を絶対値の小さい gain values とともに用いれば，従来の RR と比較的高い相関を保ちながら評価を行うことができると考えられる．ただし，gain values の変更に対する O-measure の頑健性は高くないので注意が必要である．

なお，紙面の都合上割愛したが，我々は上記 4 つの異なる評価指標の利用により，上位システム間でも平均的優劣の逆転が頻繁に起こること，および特定のシステム対を検索課題毎に比較しても逆転が頻繁に起こることを確認した．これらの分析結果を総合すると，以下のような知見が得られる．

- 全適合文書の高精度検索と適合文書 1 件の高精度検索では異なる検索戦略が要求される可能性がある．
- 部分適合以上の文書 1 件の高精度検索と高適合文書 1 件の高精度検索では異なる検索戦略が要求される可能性がある．

4. システム順位の信頼性

本章では，NTCIR-3 CLIR データのうち最も規模が大きい C-runs のうち AveP が最も高い 30 システムと [8]，Buckley/Voorhees[1] および Voorhees/Buckley[11] の手法を用い，評価尺度の安定性および判別能力について考察する．紙面の都合上ごく簡単に説明すると，Buckley/Voorhees の手法は，テストコレクションと参加システムのセットが与えられた場合に，各評価指標に対して，検索課題の入れ替わりに対するシステム対の優劣の安定性を表す少数派率と，優劣をつけられない割合を表す同順位率のトレードオフを明らかにする．一方，Voorhees/Buckley の手法は，検索課題セットをまるごと全く別のものに入れ替えた場合にシステム対の優劣が逆転する割合を，システム対の評価値の絶対差に対してプロットすることにより，与えられた信頼度を保証するために最低限必要なシステム間の評価値の差と，そのときの評価指標の判別能力を明らかにするものである．今回用いた具体的アルゴリズムは [6, 8] に示した通りである．

図 1 に 30 件の C-runs を元に算出した各評価尺度の少数派率-同順位率のトレードオフ曲線を示す．原点に近い評価尺度ほど理想的な尺度であるので，以下のことがわかる．

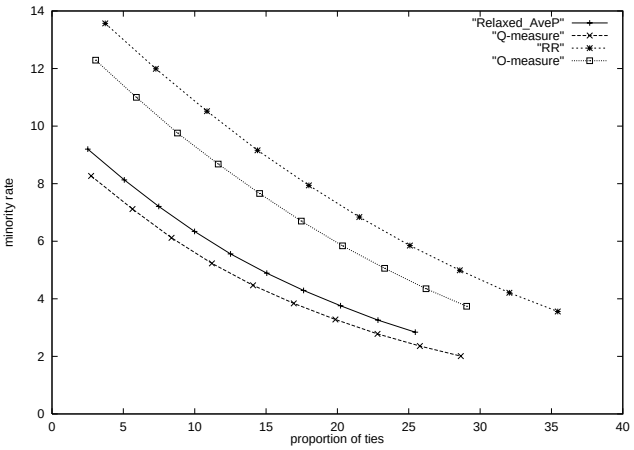

図 1: Minority rate / proportion of ties (Top 30 C-runs; 20 topics).

- AveP/Q-measure のほうが，RR/O-measure より安定している．
- O-measure のほうが RR よりも安定している．

一方，表 4 は Voorhees/Buckley の逆転率の計算に基づき，30 件の C-runs を 20 件の検索課題に基づき順位づけする場合の各指標の判別能力を示している．例えば表 4(a) の Q-measure の行を見ると，以下のことがわかる．

(i) 信頼度 95%以上の結論を得たい場合，システム間の Q-measure による絶対差は 0.10 以上必要である．
(ii) ランダムに選出した 20 件の検索課題からなる全ての課題セットについて，各システムの平均 Q-measure を算出した際の最大値は 0.5490 である．
(iii) 従って，(i) の絶対差を相対差に換算すると，0.10/0.5490=18% となる．
(iv) 全課題セットと全システム対の組み合わせのうち，上記の条件を実際に満たしたものの割合は 25.4%である (各行はこの値によりソートされている)．

表 4(a) より，20 の検索課題により信頼度 95%以上で C-runs を判別したい場合，AveP および Q-measure では全システム対の約 1/4 について判別可能であるが，RR および O-measure はこの信頼度を保証できないことがわかる．すなわち，全ての適合文書にもとづく評価指標よりも，ひとつの適合文書にもとづく評価指標のほうが判別能力が低い．そこで信頼度 90%以上とすると，表 4(b) に示したように O-measure は 16.5%のシステム対について判別可能となるが，RR は依然として判別ができないことがわかる．さらに，信頼度 80%以上とすると，やっと RR も 27.5%のシステム対について判別可能となる．しかしこの割合は，AveP および Q-measure の半分以下にとどまっており，また O-measure よりも僅かに低くなっている．これらをまとめると以下のようになる．

表 4: Discrimination power of metrics based on swap rate computation (Top 30 C-runs; 20 topics).

	(i) absolute diff required	(ii) max performance observed	(iii) relative diff required	(iv) %comparisons with required diff
(a) 95% confidence (5% swap rate)				
Q-measure	0.10	.5490	18%	25.4%
Relaxed AveP	0.11	.5392	20%	23.7%
O-measure	-	.8792	-	-
RR	-	.9750	-	-
(b) 90% confidence (10% swap rate)				
Q-measure	0.08	.5490	15%	36.7%
Relaxed AveP	0.09	.5392	17%	33.8%
O-measure	0.20	.8792	23%	16.5%
RR	-	.9750	-	-
(c) 80% confidence (20% swap rate)				
Relaxed AveP	0.05	.5392	9%	59.7%
Q-measure	0.05	.5490	9%	57.7%
O-measure	0.14	.8792	16%	33.2%
RR	0.16	.9750	16%	27.5%

- AveP/Q-measure のほうが，RR/O-measure より判別能力が高い．
- O-measure のほうが RR よりも判別能力が高い可能性がある．

ただし，Buckley/Voorhees および Voorhees/Buckley の手法はテストコレクションのみならず検索結果セットにも依存するものであり，本実験で得られた知見が例えば TREC/NTCIR Web などのタスクにどれくらい当てはまるかは今後検証していく必要がある．なお，E-runs でも C-runs と傾向の似た実験結果が得られているが，紙面の都合上割愛した．

5. まとめ

本稿では，高適合文書を 1 件だけ高精度に検索するタスクに適した評価指標 O-measure を提案し，NTCIR-3 CLIR に提出されたシステムの検索結果を用いて以下を明らかにした．

1. 全適合文書を高精度に検索するタスクと 1 件だけ高精度に検索するタスクでは異なる検索戦略が要求される可能性がある．また，ひとつの適合文書に基づく評価指標は全ての適合文書に基づく評価指標に比べ安定性および判別能力が低い．RR や O-measure により情報検索評価を行う場合，検索課題数を増やすなどの対策が必要である [10]．
2. 部分適合以上の文書を 1 件高精度に検索するタスクと高適合文書を 1 件高精度に検索するタスクでは異なる検索戦略が要求される可能性がある．高適合・高精度検索タスクには O-measure を絶対値の小さい gain values とともに用いると有用である可能性があるが，gain values の変更に対する頑健性は Q-measure と比べると高くないので注意が必要である．また，O-measure は RR よりも安定性および判別能力が若干高い可能性がある．

上記の頑健性に関する問題の解決策としては，O-measure の代りに，最初に見つかった適合文書についてのみ大きな gain value を与える Q-measure の変形版を利用することが考えられるが，これはより複雑な指標となる．今後，本研究で得られた知見が他のデータや実際の検索環境にどれくらい当てはまるか検証したい．なお，Voorhees/Buckley 法における標本抽出方法が検索指標の相対評価に与える影響については別途報告する [7]．

参考文献

[1] Buckley, C. and Voorhees, E. M.: Evaluating Evaluation Measure Stability. ACM SIGIR 2000 Proceedings (2000) 33–40

[2] Chen, K.-H. *et al.*: Overview of CLIR Task at the Third NTCIR Workshop. NTCIR-3 Proceedings (2003)

[3] Eguchi, K. *et al.*: Overview of the Web Retrieval Task at the Third NTCIR Workshop, NTCIR-3 Proceedings (2003)

[4] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems **20–4** (2002) 422–446

[5] Sakai, T.: Ranking the NTCIR Systems based on Multigrade Relevance. AIRS 2004 Proceedings (2004) 170–177 Also available in Myaeng, S. H. et al. (Eds.): AIRS 2004, Lecture Notes in Computer Science **3411**, Springer-Verlag (2005) 251–262

[6] Sakai, T.: A Note on the Reliability of Japanese Question Answering Evaluation. 情報処理学会研究報告 **FI-77-7** (2004) 57–64

[7] Sakai, T.: The Effect of Topic Sampling in Sensitivity Comparisons of Information Retrieval Metrics. 情報処理学会研究報告 **FI-80** (2005) to appear

[8] Sakai, T.: The Reliability of Metrics based on Graded Relevance. AIRS 2005 Proceedings, Lecture Notes in Computer Science, Springer-Verlag (2005) to appear

[9] Shah, C. and Croft, W. B.: Evaluating High Accuracy Retrieval Techniques. ACM SIGIR 2004 Proceedings (2004) 2–9

[10] Soboroff, I.: On Evaluating Web Search with Very Few Relevant Documents. ACM SIGIR 2004 Proceedings (2004) 530–531

[11] Voorhees, E. M. and Buckley, C.: The Effect of Topic Set Size on Retrieval Experiment Error. ACM SIGIR 2002 Proceedings (2002) 316–323