データベースシステム 118－4
情 報 学 基 礎 54－4
（1999. 5. 17）

大規模テストコレクション構築のためのプーリングについて：NTCIR-1の予備テストの分析

栗山和子 神門典子 野末俊比古 大山敬三
学術情報センター
{kuriyama,kando,nozue,oyama}@rd.nacsis.ac.jp

概要. 本研究の目的は、(1) 大規模テストコレクションを構築する手法としてのプーリングの有効性を検証し、(2) プーリング件数が検索システムの評価に関連があるかどうか調べ、(3) 正解判定の際の判定のゆれがシステムの評価に関係してくるかどうかを明らかにすることである。

(1),(2) のために、NTCIR-1 の訓練用正解セットを使用した予備テストで提出された結果を用いてプーリング実験を行なった。また、プーリング法の改良である Move-to-Front 法の簡略なヴァリエーションを提案し、平均精度が上位である提出結果からの文書をプーリング法によるプールに追加する実験も同時に行なった。その結果、プーリング法の有効性、すなわち、作成された正解リストの網羅性、および、プーリングによって作成された正解文書リストの公平性が確かめられた。

(3) のために、NTCIR-1 の訓練用セットを構築する際に行なった、異なる判定者による 3 種類の正解判定結果（判定者 A,B それぞれによる判定, 両者の協議による最終判定）を用いて評価実験を行なった。結果として、30 件の検索課題を用いて検索結果を評価したとき、検索精度の平均は異なる正解判定リスト間においてほとんど差がなくなり、他数の検索課題を用いて評価を行なえば、判定者間の判定のゆれは評価においては問題ではないということがわかった。

Pooling for a Large Scale Test Collection : Analysis of the Search Results for the Pre-test of the NTCIR-1 Workshop

Kazuko Kuriyama Noriko Kando Toshihiko Nozue Keizo Oyama
National Center for Science Information Systems

Abstract. The purposes of this study are; (1) to verify the effectiveness of the pooling method to construct a large scale test collection, (2) to examine whether the number of documents in a pool affects the evaluation of retrieval results, and (3) to verify the reliability of a test collection through investigating the effect of variations in relevance assessments have on the evaluation of search effectiveness since individual relevance assessments are known to be vary widely.

For (1) and (2), we carried out an experiment using the training qrel's (queries and their relevance assessments) and the submitted search results for the pre-test of the NTCIR-1 workshop. The result of it is that we verified the efficiency and effectiveness of the pooling, the exhaustiveness of the relevance assessments, the reliability of the evaluation using the test collection based on the pooling method, and the effectiveness of the modified Move-to-Front pooling method.

For (3) we compared search effectiveness of retrieval strategies using the three different sets of relevance assessments judged by the assessor A, B and the final judgment based on the negotiation between A and B. As a result, we found very high similarity among the rankings of retrieval systems produced using different set of relevance assessments when a sufficient number of search topics were used in the evaluation.

1 はじめに

1.1 NTCIRプロジェクト

著者らは、学術情報センター研究開発部の「情報検索システム評価用テストコレクション構築」プロジェクトにおいて、情報検索システム評価用テストコレクション NTCIR（エンティサイル：NACSIS Test Collection for Information Retrieval systems）の構築を行なっている。その過程において、昨年 11 月から今年 9 月まで、テストコレクション 1 (NTCIR-1)（予備版）を用いた、コンペティション形式のワークショップを開催している [4],[5]。

本稿では、NTCIR-1 を使用して、昨年 12 月に行なった予備テストの結果を用いたプーリング実験の結果とプーリングの有効性について報告する。また、NTCIR-1 を構築する際に行なった複数の正解判定者による正解判定の違いによって、検索結果の評価がどのような影響を受けるかということを合わせて報告する。

1.2 テストコレクションについて

テストコレクションとは、情報検索システムの検索性能評価に用いられる実験用セットのことであり、(1) 文書データベース、(2) 検索課題群、(3) 各検索課題に対する正解文書の網羅的リスト、からなる。テストコレクション構築においては、各検索課題についてデータベース中の全ての文書のそれぞれが適合するかどうか正解判定を行ない、網羅的な正解文書リストを作成する。しかしながら、数万件以上の文書を含む大規模データベースに対してこれを行なうことは非現実的である。

大規模テストコレクション構築における正解文書候補の収集の手法としては、TREC(Text REtrieval Conference)[7] などで用いられているプーリング (Pooling) 法がある。プーリング法では、同一課題に対して複数の検索手法を用いた複数の検索システムによる検索結果の上位一定数を集めてプールし正解候補として、それに対して正解判定を行なう。TREC では、1992 年から毎年コンペティションを行ない、多くの研究者から検索結果を収集してプーリングをすることによって、大規模テストコレクションの効率的な構築を実現している。

また、プーリング法を改良したものとして、Move-to-Front プーリング法がある [2]。Move-to-Front 法では、正解である確率の高い文書をプールするため、検索精度の高い検索結果の文書に優先順位を付けて、上位一定数をプーリングする。そのため効率的に網羅的な正解リストを構築できることが確かめられているが、一方で、システム間の相対的評価に公平さがなくなるのではないかという問題がある。

1.3 本研究の目的

大規模テストコレクションにおける、正解文書リストのプーリングによる構築、および、正解文書リストを用いたシステム評価についての問題点は以下の通りである。

(1) 正解文書リストの網羅性：プーリングによる正解文書収集では、プールに入れられなかった文書を不正解文書と仮定しているため、正解文書候補をいかに網羅的に集めてプールすることができるかということが問題となる。

(2) 正解文書リストの公平性：テストコレクションの構築の目的である、検索システムの評価という点から見ると、複数のシステムの検索結果が同数でなく複数あるとき、検索システムごとに上位一定数の文書をプーリングするか、検索結果ごとにするか、あるいは、検索結果の評価（平均精度）を用いて検索結果ごとに上位何件までプーリングするかを変化させるのか、などによって、システム間の相対的な評価が変化する可能性がある。したがって、なるべく、どのシステムにも公平になるような方法で正解文書リストを作成する必要がある。

(3) 正解判定の一致度：正解文書候補は、一般的には、検索課題ごとに、複数の判定者によって判

定される。このとき、判定者間の判定のゆれが正解判定リストを用いてシステムを評価するときにどの程度影響を与えるか検証する必要がある。

本稿では、以上の点について、特に検索結果の評価への影響という面から考察を行なった。

(1) について検証するため、NTCIR-1 の訓練用正解セットを使用した予備テスト（昨年 12 月）の提出結果を用いてプーリング実験を行ない、日本語の大規模テストコレクション構築におけるプーリングの有効性について考察した。また、プーリング法の改良である Move-to-Front 法のヴァリエーションとして、プーリング法でのプールにおいて正解文書を多く含んでいるとされた提出結果からより多く（一定数）の文書をプールに入れる実験も同時に行なった。

(2) については、プーリングに入れる各提出結果からの上位 X件を変化させて、プーリングする件数と評価との関係を考察した。

(3) について、NTCIR-1 では、一つの検索課題の正解判定を複数の判定者によって行なっている。一つの検索課題に対する複数の判定者の判定は完全に一致することはなく、最終的には決められた一人の最終判定者が正解を決定することになる。システムの評価の観点からは、一般的には、複数の判定者の判定結果を用いて、それぞれ別に検索結果を評価したときに、複数の検索課題に対する評価の平均はほとんど差がなくなるので、判定者間の判定のゆれは評価においては問題ではない、という報告がある [1],[6],[8]。本稿では、このことを検証するため、予備テスト用の訓練用セットを構築する際に行なった複数の判定者による正解判定結果を用いて評価実験を行ない、判定結果のゆれによる評価の違いについて考察した。

また、本稿では考察しないが、評価に関することで、検索課題の性質とシステムの特徴との関連の問題がある。あるシステムが得意としている検索方法がある検索課題の性質に適していた場合、その検索課題についての評価が特によくなり、他の検索課題いついての評価が平均的であったとしても、全体の平均精度があがってしまうということが考えられる。このような場合を区別できるように、検索課題の性質とシステムの特徴との関連を調べておく必要がある。これについては、次回の課題とする。

2 予備テストの提出結果を用いたプーリング実験

2.1 予備テストの概要

NTCIR-1 では、訓練用検索課題については、事務局で予め作成した正解文書リストがある。このリストの正解文書の網羅性、システム評価に対する公平性を検証するため、NTCIR ワークショップでは、昨年 12 月 2 日に予備テストを行なった [5]。

予備テストでは、訓練用検索課題 30 件に対する検索結果を、ワークショップ参加者から自由参加で提出してもらい、内部で用意した正解文書リストの網羅性を評価し、新たな正解文書を追加した。

この予備テストでは、10 チームで合計 23 セットの検索結果が提出された。23 の内訳は、随時検索タスク 8 チーム 16 セット、言語横断タスク 4 チーム 5 セット、単言語（言語横断検索のための baseline として）1 チーム 2 セットである。本稿では、この 23 セットのうち、随時検索タスクの提出結果 16 セットを対象として実験を行なった。

本稿では、予備テストの「提出結果」を「検索結果」という言葉と区別するため、以下では、submission と呼ぶ。一つの submission は、ある検索システムによる検索結果の、30 件の検索課題に対するそれぞれ上位 1000 件ずつを一つのファイルに順にリストとして並べたものである。システムの検索性能は、この 30 件の検索課題に対する検索結果の評価尺度の平均によって順位付けた。

2.2 プーリング実験

2.2.1 プーリング法によるプーリング

著者らの以前の論文 [5] では、事務局で準備した正解文書リスト (ver.1) と予備テストで提出された検索結果との関係、正解文書リスト (ver.2) の網羅性、および評価の公平性を明らかにするため、正解文書リスト (ver.1) と 23 submission の全ての文書をプーリングした文書リストを用いて評価実験を行なった。結果として、以下のことがわかった。(1) 正解文書リスト (ver.1) では、97.1%の正解を探すことができた。(2)ver.1 の中の Auto (プロジェクト内部での検索結果によるプーリング)、Interactive(I) (図書館情報学専攻の大学院生が対話型システムを用いて検索した結果)、ver.1、Pretest (23 submission の上位 1000 件によるプーリング)、ver.2 のいずれを用いても、システムの評価に影響はなかった。(3) 対話型検索 (I) は他の方法では探せなかったユニークな正解を見つけた。(4) 対話型検索 (I) によって見つかった正解のみを用いてシステム評価を行なっても、対話型システムに有利になることはなかった。

以上のようなことを踏まえて、本稿では、提出された検索結果だけを用いたプーリング法の有効性とプーリングによって作成された正解文書リストの評価に対する公平性を検証するため、予備テストの随時検索タスクの 16 submission を用いてプーリング実験を行なった。

$X = 10, \cdots, 90, 100, \cdots, 1000$ について、各 submission から上位 X 件のプーリングを行なった。そのプールのそれぞれを順に P10, … , P1000 とする。全ての正解文書リストを R (Relevance Assessment) とする。正解文書リストの網羅性を高めるため、対話型システムを用いて検索した結果 (前述) を I (Interactive) とする。P100 に I を加えたものを P100I とする。

表 2-1 に R, I, P10, P30, P100, P1000, P100I の検索課題ごとの正解文書数を示す。「正解」は「正解 (relevant)」と「部分的正解 (partical relevant)」を合わせて「正解」とする。今回のワークショップでは、実際には、予備テスト以前にプロジェクト内部のシステムで検索した結果に I を加えたものを正解文書リスト (ver.1) として予め用意しておき、新たにみつかったものを正解文書リスト (ver.1) に加えて正解文書リスト (ver.2) とした [5] が、ここでは、プーリングの効率性、有効性を考えるため、正解文書リスト (ver.2) を網羅的な正解文書リスト (R) であると仮定する。

表 2-1. プール中の正解文書数

Tpcs	R	I	P10	P30	P100	P1000	P100I
0001	293	289	55	109	169	283	289
0002	19	19	13	17	18	19	19
0003	14	11	3	9	12	14	14
0004	38	37	23	30	34	37	37
0005	13	2	4	5	10	13	10
0006	72	61	22	36	59	71	66
0007	16	0	4	10	12	15	12
0008	25	0	8	16	25	25	25
0009	8	8	6	6	6	8	8
0010	55	51	25	35	45	53	53
0011	7	3	6	7	7	7	7
0012	70	68	25	49	57	67	68
0013	38	31	7	10	26	37	37
0014	317	283	55	96	210	310	287
0015	20	20	11	15	17	20	20
0016	5	5	4	5	5	5	5
0017	16	13	6	8	13	15	14
0018	167	100	21	39	75	155	120
0019	92	90	33	54	71	92	90
0020	16	16	12	15	16	16	16
0021	11	11	6	9	10	11	11
0022	82	68	29	43	65	82	75
0023	98	98	31	49	76	98	98
0024	158	156	38	70	122	157	156
0025	23	23	21	23	23	23	23
0026	23	23	21	21	23	23	23
0027	23	23	11	16	22	23	23
0028	1590	1586	21	75	205	1018	1587
0029	180	113	27	61	94	162	141
0030	23	23	11	17	20	22	23
%av1	100	82.0	44.8	61.8	79.9	97.0	94.8
%av2	100	81.5	44.8	63.7	82.2	98.1	94.7
%av3	100	84.8	14.8	23.3	51.9	90.1	89.7
%av4	100	92.5	33.8	47.7	79.5	98.9	95.7
%av5	100	77.9	50.2	63.7	88.2	98.4	95.7
%av6	100	81.0	80.2	91.7	91.7	100	100

Tpcs:検索課題 (Search Topics) 番号 (0001~0030 の 30 件)
R:正解文書リスト (ver2)
I:事務局内部で行なった対話型システムによる検索結果
PX:各 submission から上位 X 件ずつをとったプール
%av1:平均正解文書数
%av2:0028 を除いた場合の平均正解文書数
%av3:正解文書数≥100 の検索課題についての平均正解文書数
%av4:50≤正解文書数<100 の検索課題についての平均正解文書数
%av5:10≤正解文書数<50 の検索課題についての平均正解文書数
%av6:正解文書数<10 の検索課題についての平均正解文書数

%av1 は、各プール中の文書の正解文書リストに

対する正解文書数の割合の平均であり、%av2 は、0028 を除いた場合の平均である。0028 は、正解文書数が 1000 件以上あるため、最大でも上位 1000 件までしかとらないプーリングではカバーしきれないと考え、除外した場合も考えた。%av3 は、正解文書リスト中の正解文書数が 100 件以上の検索課題についての平均、%av4 は正解文書数が 50 件以上 100 件未満の検索課題についての平均、%av5 は正解文書数が 10 件以上 50 件未満の検索課題についての平均、%av6 は正解文書数が 10 件未満の検索課題についての平均である。

図 2-1 に、P10, … , P1000 に含まれる正解文書の割合のグラフを示す。

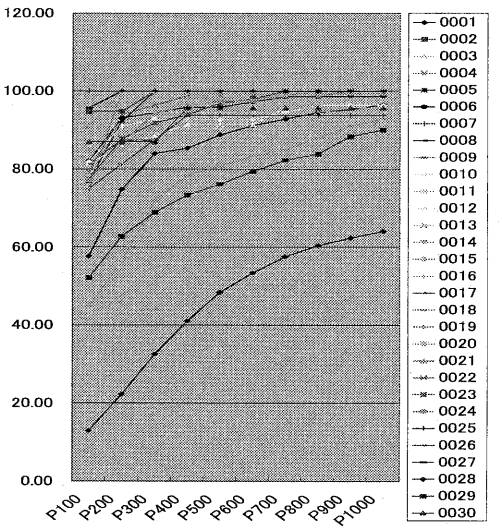

図 2-1. プール中の正解文書

PX:各 submission から上位 X 件ずつをとったプール
横軸の数値:PXに含まれる正解文書数/正解文書リストに含まれる正解文書数 (%)

表 2-1 の %av1, %av2 からわかるように、上位 100 件をプーリングした P100 の場合、対話型で検索した場合 (I) とほぼ同程度の網羅性で正解文書を含んでいると考えられる。%av3（正解文書数 100 以上）～%av6（正解文書数 10 未満）からわかるように、I では、正解文書数が少ないほど正解文書の網羅性が低くなる傾向があるのに対し、P100, P1000 では、正解文書数が少ないほど、網羅性が高くなっている。これは、図 2-1 からもわかる。すなわち、正解文書数が多いほど、より多くの文書をプールする必要がある。このことから、正解文書数が多い検索課題については、上位 X 件のプーリングだけでは不十分である可能性があるので、対話型検索によって補完することが考えられる。P100 に I を加えたものは、6 種類の平均のいずれでも、ほぼ、P100 と I の間くらいの値になり、全検索課題についての平均 %av1 では、1000 件全部をプーリングした P1000 の 97.0 % には及ばないものの、94.8 % をカバーしている。よって、プーリングする文書数が少ないときには、対話型検索による補完が有効であることがわかった。

2.2.2 簡易 Move-to-Front 法によるプーリング

次に、より効率的で簡便なプーリング法を試みるため、Move-to-Front プーリング法を簡略化した方法で実験を行なった。

本来の Move-to-Front 法では、プーリング法で一定数プーリングして作成した正解文書リストを用いたシステム評価によって、検索性能の高いシステムから正解である確率の高い文書を優先順位列 (priority queue) に並べ、正解である確率が高い順に一定数プーリングする [2]。本稿では、より少ない X で正解の確率の計算をせずに効果的にプーリングを行なうために、$X = 10, 30, 100$ について、P10, P30, P100 でそれぞれ評価したとき、評価の高い 4 つの submission の上位 X 件に続く 100 件の文書をそれぞれ追加することを考えた。

P10, P30, P100 のそれぞれに追加したプールを MP10, MP30, MP100 とする。ただし、複数チームが複数 submission を提出しているので、単純に評価が上位の submission からとるのではなく、1 チーム 1 submission までとして 4 つの submission を追加した。MP10, MP30, MP100 と対話型システムで検索した結果 (I) を合わせた結果を MP10I, MP30I, MP100I とする。

表2-2に、MP10, MP30, MP100, MP10I, MP30Iして検索結果の評価に使用するかによって、評価MP100I に含まれる正解文書数を示す。

表 2-2. プール中の正解文書数

Tpcs	MP10	MP30	MP100	MP10I	MP30I	MP100I
0001	135	145	194	289	289	289
0002	17	18	18	19	19	19
0003	8	12	12	14	14	14
0004	31	34	35	37	37	37
0005	5	7	11	6	7	11
0006	43	54	67	62	62	69
0007	11	11	12	11	11	12
0008	24	25	25	24	25	25
0009	6	6	7	8	8	8
0010	38	42	49	51	51	53
0011	7	7	7	7	7	7
0012	57	53	59	68	68	68
0013	15	13	30	35	36	37
0014	146	175	231	283	284	287
0015	14	17	17	20	20	20
0016	5	5	5	5	5	5
0017	12	10	14	13	13	14
0018	60	64	87	116	117	124
0019	62	67	74	90	90	90
0020	15	16	16	16	16	16
0021	10	9	10	11	11	11
0022	59	62	76	74	76	80
0023	62	73	80	98	98	98
0024	93	111	132	156	156	156
0025	23	23	23	23	23	23
0026	22	23	23	23	23	23
0027	19	22	23	23	23	23
0028	124	136	265	1586	1587	1587
0029	78	88	100	133	136	144
0030	15	18	20	23	23	23
%av1	69,0	74.2	93.1	92.5	93.2	95.6
%av2	71.1	76.5	93.1	92.3	93.0	95.5

Tpcs:検索課題 (Search Topics) 番号（0001〜0030 の 30 件）
MPX:PXに上位 4 チームの各 submission から上位 100 件ずつをとった文書を追加したプール
MPXI:MPXに対話型システムでの検索結果を加えたプール
%av1:平均正解文書数
%av2:0028 を除いた場合の平均正解文書数

表2-2からわかるように、追加のプーリングでは、どのプールでも元のプールよりもほぼ10%程度の正解文書数の向上が見られる。また、Iと組合せたときには、どのプールでも正解文書数の割合が90%以上になり、対話型検索と組合わせることによって、どの submission からも上位の一定数をプーリングする方法に比べて、より少ない数でのプーリングが可能であると考えられる。

2.3 プーリングする件数による評価の違い

プーリング件数の異なるプールがあるとき、どのプールに含まれる正解文書を正解文書リストとは異なってくる。プーリングに用いた各 submission からの文書数のシステム評価への影響を調べるために、前節で示したプールのうち、P30, P100, MP30, MP100, P100I, MP30I のそれぞれを用いて正解文書リストを作成し、各 submission の評価を行なった。評価した submission は随時検索タスクの検索結果 8 チーム 16 セットの内、それぞれ異なる検索システムによって提出された特徴的な 8 セットを選んだ。「正解」、「部分的正解」と判定されたものをそれぞれの「正解」として各 submission の精度を計算し、平均精度（補完なし）の値で順位を付けた結果を表 2-3 に示す。A, B, C, D, E, F, G, H はそれぞれの submission の run-id を表わす。

表 2-3. プーリング数システム評価への影響

run-id	A	B	C	D	E	F	G	H
R	1	2	3	4	5	6	7	8
P30	1	2	3	5	4	6	7	8
P100	1	2	3	5	4	6	8	7
MP30	1	2	3	5	4	6	8	7
MP100	1	2	3	5	4	6	8	7
P100I	1	2	3	5	4	6	8	7
MP30I	1	2	3	5	4	6	8	7

run-id:8 セットの submission の run-id
R:正解文書リスト (ver2)
PX:各 submission から上位 X 件ずつをとったプール
MPX:PXに上位 4 チームの各 submission から上位 100 件ずつをとった文書を追加したプール
MPXI:MPXに対話型システムでの検索結果を加えたプール

表 2-3 から、一般的なプーリング法において、プール数を変えたプーリングでは、ほとんど相対的な評価の順位には影響がないということがわかる。また、P30 と MP30、P100 と MP100 のそれぞれで順位が同じであることから、上位 100 件の追加のプーリングを行なっても、相対的な評価には影響がないと考えられる。

3 異なる判定結果による評価

正解判定のゆれがシステム評価に影響を及ぼすかどうか調べるため、予備テスト用の訓練用セットを構築する際に行なった、複数の判定者による正解判定結果と正解文書リスト (ver.2) を用いて、評価を行なった。

正解判定は、各検索課題について、異なる判定者A、Bにより判定A、Bとして行ない、最終判定は両者の協議によって行なった。表3-1にそれぞれの判定結果含まれる正解文書数、括弧内にその判定で正解とされた文書数、正解文書リストに対する割合を示す。正解文書リストには、判定A、Bの時点での判定用文書リストには含まれていない文書があるため、それについては除外した。含まれていない文書を除外した正解文書リストを正解文書リストR′と呼ぶことにする。各検索課題ごとの、正解文書リストR′に含まれる正解文書リスト中の正解文書の割合の平均は、77.1 % である。検索課題0018については、判定Bが他の検索課題とは異なる状態で行なわれたため、除外した。また、判定Aにおいて、正解判定が行なわれなかった、検索課題 0007,0009,0010,0011,0013,0014,0028 についても除外した。

表3-1. 異なる判定結果の正解文書数

Tpcs	jdgA		% jdgA	jdgB		% jdgB
0001	107	(120)	52.5	172	(354)	84.3
0002	5	(9)	27.8	11	(17)	61.1
0003	2	(2)	16.7	4	(10)	33.3
0004	29	(43)	90.6	10	(20)	31.3
0005	9	(13)	75.0	9	(197)	75.0
0006	13	(13)	24.1	49	(201)	90.7
0008	3	(2)	12.5	13	(49)	54.2
0012	22	(39)	44.0	0	(1)	0
0015	9	(10)	50.0	2	(28)	11.1
0016	5	(5)	100	4	(12)	80.0
0017	5	(5)	45.5	6	(14)	54.6
0019	58	(58)	68.2	81	(349)	95.3
0020	1	(1)	6.7	5	(8)	33.3
0021	3	(7)	27.3	8	(65)	72.7
0022	47	(68)	75.8	46	(280)	74.2
0023	72	(74)	91.1	28	(96)	35.4
0024	93	(100)	66.9	123	(279)	88.5
0025	17	(18)	89.5	17	(23)	89.5
0026	17	(17)	81.0	21	(24)	100
0027	8	(11)	40.0	14	(91)	70.0
0029	81	(95)	70.4	88	(211)	76.5
0030	8	(13)	57.1	12	(27)	85.7
ave	21.2	(25.0)	41.8	38.1	(100.4)	60.6

Tpcs:検索課題 (Search Topics) 番号
jdgA:判定Aに含まれる正解文書数
jdgA 括弧内:判定Aで正解と判定された文書数
%jdgA:判定Aに含まれる正解文書数/正解文書リストに含まれる正解文書数 (%)
jdgB:判定Bに含まれる正解文書数
jdgB 括弧内:判定Bで正解と判定された文書数
%jdgB:判定Bに含まれる正解文書数/正解文書リストに含まれる正解文書数 (%)
ave:検索課題全体についての平均

評価するsubmissionは、前節と同様に、随時検索タスクの検索結果8チーム16セットの内の特徴的な8セットを選んだ。判定A、判定B、正解文書リストR′のそれぞれで「正解」、「部分的正解」と判定されたものをそれぞれの「正解」として各submissionの精度を計算し、平均精度（補完なし）の値で順位を付けた結果を表3-2に示す。A, B, C, D, E, F, G, H はそれぞれのsubmissionのrun-idを表わす。

表3-2. 判定結果のシステム評価への影響

run-id	A	B	C	D	E	F	G	H
R′	1	2	3	4	5	6	7	8
jdgA	1	3	2	5	4	6	7	8
jdgB	1	2	3	4	5	6	7	8

run-id:8セットのsubmissionのrun-id
R′:正解判定文書リスト (ver.2) から判定A・判定Bに含まれていない文書を除外したもの
jdgA:判定Aに含まれる正解文書数
jdgA 括弧内:判定Aで正解と判定された文書数
%jdgA:判定Aに含まれる正解文書数/正解文書リストに含まれる正解文書数 (%)
jdgB:判定Bに含まれる正解文書数
jdgB 括弧内:判定Bで正解と判定された文書数
%jdgB:判定Bに含まれる正解文書数/正解文書リストに含まれる正解文書数 (%)

表3-1に示したように、正解文書リストに含まれる正解文書は、判定Aでは平均41.8%、判定Bでは平均60.6% しか正解と判定されていない。しかし、表3-2からわかるように、判定のゆれによる順位の変動は少ない。また、それぞれの正解文書を用いて計算した各submissionの平均精度の平均は、判定Aで0.3263、判定Bで0.2856、正解文書リストR′で0.3168とこれも大幅な変化は見られなかった。判定A、Bで正解とされた文書数、および、判定A、Bの文書リストの中で最終的に正解とされた文書数の割合からわかるように、判定A、Bと正解文書リストの間ではかなり判定が異なっている。しかし、表3-2や平均精度から見ても、相対的な評価にはほとんど影響が出ていない。したがって、日本語に対する情報検索についても、複数の異なる正解文書リストのそれぞれでシステムを評価した場合でも、システムの相対的な評価にはほとんど影響がないと言える。

4 まとめ

本稿では、大規模テストコレクションを構築する際に必要となる、正解文書候補をどうすれば効果的で公平に収集することができるかという観点から、NTCIR-1 の予備テストの提出結果を用いて、プーリング実験を行なった。プーリング実験の結果から、NTCIR-1 については、以下のことがわかった。

(1) 各 submission の上位 100 件をとるプーリングは、正解文書数が 1〜50 件程度である場合には、ほぼ正解文書を網羅的に収集できるが、正解文書数が 100 件以上である場合には、網羅的であるとはいえない。すなわち、正解文書数が多い検索課題ほど、多くの文書をプーリングする必要がある。その場合でも、上位 100 件のプールに、対話型で検索した結果を加えるとほぼ正解を網羅的に収集することができる。

(2) 上位 X件をプーリングする場合、平均精度で順位付けしたときに上位の submission から追加で上位 100 件をプーリングすれば、対話型検索を併用しなくても、ほぼ網羅的に正解文書が収集できる。また、対話型検索を併用すれば、100 件より少ない件数のプーリングでもほぼ網羅的になる。

(3) 一般的なプーリング、追加のプーリングでも、上位何件をプーリングするかによる、システム (submission) の相対的評価に大きな変化は見られない。

判定者の異なる複数の正解文書リストを用いた提出結果の評価によって、以下のことがわかった。(4) 日本語のテストコレクションの構築過程においても、正解判定のゆれによるシステム間の大きな評価の違いは見られなかった。

以上のようなことから、今年 3 月に行なった本テストの提出結果に対しては、各 submission から上位 100 件のプーリングを行なった。現在、各検索課題につき二人の判定者による正解判定作業を行なっている。正解がある程度以上多い場合には、その検索課題の分野の研究者による対話型検索での検索結果の追加を検討する予定である。

謝辞

本研究は、日本学術振興会未来開拓学術研究推進事業「高度分散情報資源活用のためのユービキタス情報システム」(課題番号 JSPS-RFTF96P00602) による。

参考文献

[1] Burgin, R. "Variations in Relevance Judgments and the Evaluation of Retrieval Performance". Information Processing and Management, Vol.28, No.5, pp.619–627, 1992.

[2] Cormack, G.V. et.al. "Efficient Construction of Large Test Collections". In Proc. of ACM–SIGIR'98, pp.282–289, Melbourne, 1998.

[3] Harman, D. "Overview of the Third Text Retrieval Conference(TREC-3)", NIST Special Publication 500-225.

[4] Kando, N. et.al. NTCIR: "NACSIS Test Collection Project". [Poster] the 20th Annual Collection of BCS–IRSG, France, 1998.

[5] 神門典子ほか. "NTCIR-1：情報検索システム評価用テストコレクション構築の方針と実際". 99–FI–53–5, pp.33–40, 1999.

[6] Lesk, M.E.; Salton, G. "Relevance Assessments and Retrieval System Evaluation". Information Storage and Retrieval, Vol.4, pp.343–359, 1969.

[7] Text REtrieval Conference (TREC). http://trec.nist.gov/

[8] Voorhees, E.M. "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness". In Proc. of ACM–SIGIR'98, pp.315–232, Melbourne, 1998.