4P－1

大規模テストコレクション NTCIR-1 の構築 (1)
－プーリングと正解判定の分析－

栗山和子　江口浩二　野末俊比古　神門典子
学術情報センター 研究開発部

1　はじめに

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、情報検索システム評価用テストコレクション NTCIR-1 を用いたワークショップ（1998 年 11 月～1999 年 9 月）[2] の評価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。

(3) のために、NTCIR-1 の評価用正解文書リス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。

2　プーリング実験

2.1　評価テストの概要

NTCIR ワークショップ [2] では、1999 年 3 月 4 日に評価テストを行ない、評価用検索課題 53 件について、24 チームで合計 121 セットの検索結果が提出された。本研究では、このうち、随時検索タスクの提出結果 18 チーム 47 セットを対象としてプーリング実験を行なった。一つの「提出結果」は、ある検索システムによる検索結果の、53 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。

NTCIR-1:NACSIS Test Collection for Information Retrieval systems-1 (1) – Analysis of the Pooling and the Relevance Assessments –
Kazuko Kuriyamal, Koji Eguchi, Toshihiko Nozue, Noriko Kando
R & D Dept., National Center for Sciecnce Information Systems (NACSIS)

2.2　プーリング法によるプーリング

$X = 10, \cdots, 1000$ について、各提出結果から上位 X 件のプーリングを行なった。そのプールをそれぞれを PX とする。全ての正解文書リストを R とする。正解文書数の多い検索課題 10 件について、事務局内部で対話型システムを用いて検索した結果を I(Interactive) とする。PX に I を加えたものを PXI とする。表 1-1 に R,I,P30,P100,P30I の検索課題ごとの正解文書数を示す。

表 1-1 の 10 件の平均正解文書数 (%)av2 からわかるように、P100 は、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。P30,P100 では、正解文書数が少ない方が、網羅性が高くなっている。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。P30I はは、av2 で P100 よりも大きい 93.7%をカバーしており、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。

2.3　簡易 Move-to-Front 法によるプーリング

以前の論文 [1] で提案した簡易 MTF 法を修正し、実験を行なった。$X = 10, \cdots$について、PX でそれぞれ評価したとき、評価の高い 5 つの提出結果の上位 X 件に続く X 件の文書をそれぞれ追加した。

PX に追加したプールを MPX とする。MPX と I を合わせた結果を MPXI とする。表 1-1 に、MP30,MP100,MP30I に含まれる正解文書数を示す。

表 1-1 より、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正解文書数の向上が見られる。また、MP30I から、対話型検索と組合わせることによって、通常のプーリング法に比べて、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。

表 1-1. プール中の正解文書数

Tp	R	I	P30	P100	M30	M100	P30I	M30I
31	21		20	20	20	20		
32	23		13	19	13	19		
33	162	161	89	136	94	144	161	161
34	15		10	15	11	15		
35	32		28	32	29	32		
36	14		13	14	13	14		
37	65	61	43	58	49	60	63	63
38	39		32	37	35	37		
39	16		16	16	16	16		
40	47		46	46	46	46		
41	16		12	16	14	16		
42	22		18	19	18	19		
43	35		33	35	35	35		
44	15		14	15	15	15		
45	18		15	18	17	18		
46	37		32	36	35	36		
47	30		23	30	27	30		
48	34		27	34	30	34		
49	20		13	18	16	19		
50	37		37	37	37	37		
51	20		15	17	15	18		
52	9		9	9	9	9		
53	84	67	65	78	68	79	79	79
54	584	504	123	253	146	295	521	524
55	40		35	40	35	40		
56	68	51	54	63	59	63	65	66
57	187	160	107	168	120	172	174	174
58	10		10	10	10	10		
59	61	56	44	57	46	60	58	58
60	10		9	10	10	10		
61	24		23	24	23	24		
62	22		17	22	19	22		
63	43		28	43	29	43		
64	59	59	45	56	50	58	59	59
65	10		10	10	10	10		
66	33		33	33	33	33		
67	23		23	23	23	23		
68	52	33	36	47	42	48	40	43
69	12		12	12	12	12		
70	111	104	71	98	77	102	107	107
71	13		12	12	12	12		
72	21		18	20	18	20		
73	11		10	11	10	11		
74	17		17	17	17	17		
75	14		14	14	14	14		
76	17		12	16	15	16		
77	6		6	6	6	6		
78	6		6	6	6	6		
79	10		10	10	10	10		
80	16		15	16	16	16		
81	9		9	9	9	9		
82	9		9	9	9	9		
83	36		27	34	32	35		
av1	100		84.0	85.6	88.1	96.4		
av2	100	86.9	63.8	85.9	70.1	89.1	93.7	94.4

表 1-2. プーリング数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
I	1	2	3	6	7	4	5	8
P30	1	3	2	4	6	5	7	8
P100	1	2	3	4	5	6	7	8
MP30	1	2	3	4	6	5	7	8
MP100	1	2	3	4	5	6	7	8
P30I	1	2	3	6	7	4	5	8
MP30I	1	2	3	6	7	4	5	8

2.4 プーリングする件数による評価の違い

プーリングに用いた各提出結果からの文書数のシステム評価への影響を調べるために、前節のプールを用いて正解文書リストを作成し、提出結果の評価を行なった。評価した提出結果はそれぞれ異なる検索システムによって提出された8セットである。各提出結果の精度を計算し、平均精度（補完なし）の値で順位を付けた結果を表 1-2 に示す。A,B,C,D,E,F,G,H,J,K はそれぞれの提出結果の run-id を表わす。

表 1-2 から、I と I を加えたプールについては他のプールと異なる傾向があるものの、上位の検索結果に対しては、プーリング数による相対的な評価の順位にほとんど影響がないことがわかった。

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、複数の判定者による正解判定結果（判定 A,B）と最終的な正解文書リスト R を正解文書リストとして用いて、評価を行なった。評価した提出結果は、前節と同じ随時検索タスクの8セットである。結果を表 2-1 に示す。

表 2-2. 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	2	3	4	7	5	6	8
jdgB	1	2	3	4	6	5	7	8

表 2-1 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。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

プーリング実験の結果、評価用セットについても、プーリング法の有効性、すなわち、作成された正解文書リストの網羅性、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。また、多数の検索課題を用いて検索結果の評価を行なえば、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、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。

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] 栗山和子ほか. “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”. 99-FI-54-4, pp.25-32, 1999.

[2] Proceedings of the 1st NTCIR Workshop on Research and Development in Japanese Text Retrieval, Tokyo, Aug.30-Sep.1, 1999. (to appear)