社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－FI－73 (2)
2003／11／13

NTCIR-3 言語横断検索タスクの分析：プーリングを中心として

栗山和子* 江口浩二† 岸田和明‡ 神門典子§

概要. 大規模テストコレクションNTCIR-3の言語横断検索システム評価用データの適合文書リストは、NTCIRワークショップ3の言語横断検索タスクにおいて、各参加者から提出された検索結果を用いて、プーリング法に基づいて作成された。本研究では、その過程で用いられたサブタスク混合プーリングが、検索結果の性能評価の公平性にどのような影響を与えるかについて考察する。
日本語文書についてのプーリングおよび評価実験を行なった結果、サブタスク混合プーリングで作成した適合文書リストと、サブタスクごとのプーリングで作成した適合文書リストとでは、相対的な評価結果に違いがある場合があり、1つのサブタスクのみからのプーリングは評価に影響を与える可能性があることがわかった。ただし、単言語検索の検索結果全てから文書をプールした場合には、そのプールは、サブタスク混合プーリングで集めた適合文書の9割以上をカバーするため、相対的評価にはほとんど影響がなかった。

Analysis of CLIR Task Results for the 3rd NTCIR Workshop : Focusing on the Pooling

Kazuko Kuriyama* Koji Eguchi† Kazuaki Kishida‡ Noriko Kando§

Abstract. The purposes of this study is to examine whether there is any effect on the relative evaluation of the IR systems using the relevance assessments of the NTCIR-3 made by the subtask-mixed pooling. We carried out an experiment using the relevance assessments and the search results submitted for the CLIR task at the third NTCIR workshop. The result is that the relative evaluation by using the relevance assessment which was made by the subtask-mixed pooling is different from the one using the pool for each of the subtasks. However, the pool for the SLIR task covers more than 90% of the relevant documents in the relevance assessment of NTCIR-3 and it has little effect on the relative evaluation.

1 はじめに

1.1 NTCIRプロジェクト

著者らは、国立情報学研究所中心とした「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の構築を行なっている[7]。その過程において、2001年9月から2002年10月まで、評価型ワークショップNTCIRワークショップ3を開催し[8]、テストコレクションNTCIR-3の構築および検索システムの評価を行なってきた。

1.2 テストコレクション

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適合文書の網羅的リスト、からなる。

適合文書の網羅的リストを作成するためには、

* 白百合女子大学 Shirayuri College
† 国立情報学研究所 National Institute of Informatics
‡ 駿河台大学, 国立情報学研究所 Surugadai University
§ 国立情報学研究所 National Institute of Informatics

各検索課題についてデータベース中の全文書の適合判定を行うことが必要である。しかし、数万件以上の文書を含む大規模データベースの全文書についてこれを行うことは、時間と人的資源の面から考えて、非現実的である。

そのため、大規模テストコレクションの適合文書リストの構築法としては、各検索課題ごとに、複数の異なる検索結果の上位一定数の文書をプールし、それを人間の判定者が検索課題に適合か不適合かを判定して、適合文書のリストを作成する、プーリング法が一般的に採用されている [1],[9],[10]。これは、異なる検索手法を用いた検索システムは異なる適合文書を探すということが知られているからである [3]。

1.3 目的

プーリング法による大規模テストコレクションの構築については、情報検索システムの評価という側面から以下のような点について考慮する必要がある。

(1) 適合文書リストの網羅性:
プーリングによる適合文書収集では、プールに入れられなかった文書は不適合文書であるものと仮定される。そのため、適合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) 適合文書リストの公平性:
検索システムの評価という観点から、適合文書リストはどのような検索システムに対しても公平になるような方法で作成する必要がある。

(3) 適合判定の無矛盾性:
適合判定が複数の判定者によって行われるとき、判定者間の判定にはゆれがある。そのゆれによって、システムの相対的評価がどのような影響を受けるかを検証する必要がある。

筆者らは、テストコレクション NTCIR-1 および NTCIR-2 構築の過程において、上記の点について、実験と考察を行なってきた [4] [5]。

NTCIR ワークショップ 1 および 2 の日本語・英語検索タスクで使用した文書は、日英の対訳データを多く含み、部分的には、ほぼ対訳コーパスとなっていた。NTCIR ワークショップ 3 の言語横断検索タスクの使用文書は、同年の同主題の文書を含むが、対訳ではなく、コンパラブルコーパスになっている。

NTCIR ワークショップにおいては、3 言語以上の多言語コーパスを使用したプーリングによる適合文書リストの作成は初めての試みであったため、NTCIR-3 の構築過程では、適合文書リストの網羅性と各 run への公平性を考慮して、サブタスク、および、検索課題と検索対象言語の組合せを区別せずに、全サブタスクの全提出結果から上位一定数をプールした。

本稿では、コンパラブルコーパスに対する、網羅的で効率的なプーリング手法による信頼性の高いテストコレクション作成を目的として、サブタスク混合プーリングによる NTCIR-3 の適合文書リストの作成が各検索結果の相対的評価に対して影響を与えているかどうかを、プーリングと評価実験を行なって検証する。

2 NTCIR-3 言語横断検索タスク

2.1 サブタスクと検索対象文書

本項以下では、NTCIR ワークショップ 3[8] の「言語横断検索タスク (Cross-Lingual IR Task)」を「CLIR タスク」と略記する。CLIR タスクには 3 つのサブタスクがある。

- 単言語検索 (SLIR): ある言語で書かれた検索課題を用いて、検索課題を同じ単言語の文書セットを検索する。
- 2 言語検索 (BLIR): ある言語で書かれた検索課題を用いて、検索課題とは異なる単言語の文書セットを検索する。
- 多言語検索 (MLIR): ある言語で書かれた検索課題を用いて、複数言語の文書セットを検索する。

各サブタスクの検索課題の言語と検索対象文書の言語の組合せ、および、各文書セットの文書数については、以前の論文 [6] を参照されたい。検索課題は中国語 (C)、日本語 (J)、英語 (E)、韓国語 (K) の4ヶ国語であり、検索対象文書は、その4ヶ国語それぞれの単言語の文書セット、および、単言語の文書セットを組合せた多言語の文書セットである。

2.2 検索課題の形式

検索課題は、利用者の検索要求を一定の書式の自然言語で明文化したものである。NTCIR-3 の CLIR タスクの検索課題は、タイトル (title)、検索要求文 (description)、検索要求説明 (narrative)、概念語リスト (concept) から成る。タイトルは、検索課題を1つの単語あるいは句として表現したものである。検索要求文は、利用者の検索要求を1文で表わしたものである。検索要求説明は、背景知識、検索の目的、適合判定の基準、用語の定義などを含み、その検索要求を作成したのではない第三者が読んでも理解できるように説明している。概念語リストは、検索課題における重要語、同義語、関連語のリストである。図1に検索課題の例を示す。

```
〈TOPIC〉
〈NUM〉014〈/NUM〉
〈SLANG〉JA〈/SLANG〉
〈TLANG〉JA〈/TLANG〉
〈TITLE〉 コンピュータウィルスによる被害
〈/TITLE〉
〈DESC〉
コンピューターウィルスによって被害を受けた事件について報じた記事が読みたい。
〈/DESC〉
〈NARR〉
コンピュータウィルスに感染すると、記憶装置（ハードディスク）上のデータが消去されるというような被害を受ける。また、メールなどを介して、他のシステムに伝染させてしまった場合、信用に関わる問題にもなる。そのような被害を受けた事件について書かれた記事であれば要求を満たす。新種のウィルスに関する記事も要求を満たす。ただし、特定の事件やウィルスではなく、一般的な被害統計だけを報じる記事は要求を満たさない。
〈/NARR〉
〈CONC〉
コンピュータウィルス, ウィルス, メリーサ, メリッサ, チェルノブイリ, 感染, 伝染, 被害, ネットワーク, インターネット, クラッカー, ハッカー, クラッキング, ハッキング
〈/CONC〉
〈/TOPIC〉
```

図 1: A Sample of Search Topics

2.3 提出結果からのプーリング

CLIR タスクのサブタスクの参加チームは、各自の検索システムを用いて、各検索課題について、単言語の文書セットあるいは多言語の文書セットを検索し、検索結果を提出する。以下では、提出された検索結果を「run」と呼ぶ。

NTCIR ワークショップ 3 の CLIR タスクの fomal run には、23 チームが参加し、199run を提出した。そのうち、正式な run として、189run が評価されたが、プーリングでは、適合文書リストの網羅性を高めるため、提出された 199run 全てを使用し、検索課題と検索対象言語の組合せに関係なく、どの run からも、同一の検索課題については同一の一定数の上位 X 件（中国語、日本語、英語の文書セットについては、X は 80,90,100 のいずれか）の文書をプールし、プールした文書を各言語ごとに分けて、適合判定を行なった。

このプーリングの過程では、各サブタスクの run を区別しないので、本稿では、サブタスク混合プーリングと呼ぶ。実際のプーリングに使用した run 数とチーム数を検索課題 (Topic) と検索対象文書 (Doc) の組合せごとに表1に示す.

3 プーリングと評価

3.1 プーリング

以前の論文 [6] では、199run のうち、日本語文書を検索対象文書セットに含む 70run を用いて、SLIR、BLIR、MLIR というサブタスクごとに、日本語文書セットについてのプーリング実験を行い、適合文書数の比較を行なった。その結果、SLIR の検索結果からのプール (PS) だ

表 1: Number of Pooled Runs and Groups

Sub task	Topic -Doc	Pooled		Total	
		Run	Group	Run	Group
SLIR	C-C	35	14	113	21
	E-E	28	14		
	J-J	33	14		
	K-K	17	8		
BLIR	E-C	16	6	57	14
	J-C	5	3		
	K-C	2	1		
	C-E	3	1		
	J-E	1	1		
	K-E	7	1		
	C-J	4	2		
	E-J	13	6		
	E-K	6	2		
MLIR	C-CE	3	1	29	7
	E-CE	6	2		
	C-JE	3	1		
	E-JE	1	1		
	J-JE	2	2		
	C-CJ	3	1		
	C-CJ E	4	2		
	E-CJ E	4	2		
	J-CJ E	3	1		
Total		199	23	199	23

けでも、ある程度網羅的に適合文書を集めることはできるが、BLIR と MLIR の検索結果からのプール (PB, PM) もそれぞれユニークな適合文書の収集に貢献し、網羅性を高めていることがわかった。

本稿では、各サブタスクの検索結果からのプール (PS, PB, PM) と NTCIR-3 の正式な適合文書リスト R を用いて評価を行い、各検索結果の評価にどのような影響があるかを考える。

SLIR における日本語の検索課題を用いた日本語の文書セットに対する検索を J-J、BLIR における日本語以外の言語 x（x は中国語 (C)、英語 (E) のいずれか）の検索課題を用いた日本語の文書セットに対する検索を x-J、MLIR における、ある言語 y（y は中国語 (C)、日本語 (J)、英語 (E) のいずれか）の検索課題を用いた日本語の文書セットを含む多言語の文書セットに対する検索を y-Jz（z は C, E, CE のいずれか）とする。NTCIR-3 の適合文書リストを R とする。R のうち J-J の 33run のみからプールした部分を PS、R のうち x-J の 17run のみからプールした部分を PB、R のうち y-Jz の 20run のみからプールした部分を PM とする。R から PS だけに含まれるユニークな文書を除いたプールを R-PSonly、R から PB だけに含まれるユニークな文書を除いたプールを R-PBonly、R から PM だけに含まれるユニークな文書を除いたプールを R-PMonly とする。

NTCIR ワークショップの CLIR タスクでは、1 つの参加チームが複数の run を提出することを認めている。各 run が検索課題のどのフィールド（図 1 参照）を使用したものかということと run の優先順位がわかるように、各 run の名前は、「Test-J-J-D-01」のように、チーム ID、検索課題の言語、検索対象文書の言語、検索課題フィールドの頭文字 (T, D, N, C)、優先順位をハイフン (-) で組み合わせた形になっている。R、PS、PB、PM から、各チーム 1run ずつ、優先順位が 1 である run のみをプールした結果を取り出したものを、それぞれ R-1run、PS-1run、PB-1run、PM-1run とする。

表 4 に、各プールに含まれる文書の合計と、R の全文書数に対する割合の、検索課題全体についての平均 (ave)、適合文書が 50 件未満の検索課題についての平均 (r<50)、適合文書が 50 件以上 100 未満の検索課題についての平均 (50≤r<100)、適合文書が 100 以上の検索課題についての平均 (100≤r) を示す。表 5 に、各プールに含まれる適合文書の合計と、R の全適合文書数に対する割合の、検索課題全体についての平均 (ave)、適合文書が 50 件未満の検索課題についての平均 (r<50)、適合文書が 50 件以上 100 未満の検索課題についての平均 (50≤r<100)、適合文書が 100 以上の検索課題についての平均 (100≤r) を示す。

3.2 評価実験

プールの違いが各サブタスクの run の評価にどのような影響を与えるか調べるため、SLIR の J-J の 33run と BLIR の x-J の 17run のうち、検索課題中の検索要求文 (D) の部分を検索に用いた、J-J の 14run と x-J の 8run を使用し、各プールを適合文書リストとして評価を行なった。

J-J の 14run をそれぞれ J-J-1～J-J-14、x-J の 8run をぞれぞれ E-J-1～E-J-6、C-J-1、C-J-2 とする。E-J-i は英語の検索課題を使用して日本語文書を検索した run、C-J-j は中国語の検索課題を使用して日本語文書を検索した run である。

プールごとの各 run の平均の平均精度 (mean average precision) とそれに基づく順位を表 2、表 3 に示す。

表 2 から、各サブタスクごとのプールを適合文書リストとして用いた場合、上位 6 位までの run では相対的評価（順位）がプールによって異なることがわかる。特に、PB、R-PSonly、PB-1run を用いた結果では、順位が入れ替わっている。R-PSonly で入れ替わりが起こっているのは、R 中の適合文書に対する、PS に含まれるユニークな適合文書の割合が 18.3%[6] と大きいため、PS 中のユニークな文書を除いていることが影響を与えていると考えられる。それに対して、PB、PM 中のユニークな文書の割合は、それぞれ、2.3%、0.5%と小さいため、R-PBonly と R-PMonly でユニークな文書を除いても、それほど評価に影響を与えない。表 3 から、E-J、C-J の run に対する、各プールによる相対的評価は、R による評価と同じであることがわかる。

適合文書リストの網羅性を高めるためには、なるべく多くの文書をプールして適合判定をすることが望ましい。しかし、効率的なプーリングと適合判定のためには、各プールに含まれる判定用文書の件数はなるべく少ないことが期待される。したがって、網羅性と効率性の両方を満たすようなプーリングを行うことが理想的である。

表 4 と見ると、R に対する PS、PB、PM のサイズ（プールされ、判定された文書数の合計）の割合の平均は、それぞれ、49.5、59.1、30.2%になっている。SLIR、BLIR、MLIR でプールされた run の数は、33、17、20 であることを考えると、PS では上位 X の重複が多いのに対して、PB ではその重複が少なく、PB ではプールされた run が PS の約半分であるのにもかかわらず、より多くの異なる文書がプールされていることがわかる。ところが、表 5 を見ると、R 中の適合文書数に対する PS、PB、PM の中の適合文書数の割合は、それぞれ、97.0、77.7、66.5%であり、PB よりもプールされた文書数が少ない PS の方がずっと多くの適合文書を含んでいることがわかる。また、PM の文書数は、R の 3 割程度であるが、適合文書数の割合は、R の 6 割以上になっている。MLIR からのプーリングでは、上位 X をプールしたとしても、検索対象である複数言語文書の言語でプール件数を分けるため、各言語ごとのプール数は、X 件よりも少ない。PM のプールサイズが小さくなるのは、このプールの言語ごとの分割のためである。このことから、PM には、より少ない文書でより効率的に適合文書を集めていると言える。

したがって、網羅性を高めるためには、サブタスク混合プーリングを行うことが望ましいが、SLIR のみからのプーリングでも、ほとんどの適合文書を集めることができ、MLIR からのプーリングでは、より小さいプールで効率的に適合文書を集められることがわかる。ゆえに、サブタスクごとのプーリングを行うのであれば、SLIR からプールをし、BLIR と MLIR から追加のプーリングを行なって、適合文書リストを補完するのが適当であるのではないかと考えられる。

プーリングの効率性を考えた場合、サブタスクごとのプーリングでは、全ての run をプールに入れるかどうかが問題になる。表 4 と表 5 から、各サブタスクの優先順位 1 の run からのみプールした場合 (PS-1run、PB-1run、PM-1run) には、プールサイズは、全 run をプールした場合の 6～8 割に減っているが、適合文書の割合は、平均で 5%程度しか変わらないことがわかる。このことから、各チームの優先順位の高い run をプールすることによって、より効率的なプーリングが行えることがわかる。また、表 2 と表 3 から、PS-1run、PB-1run、PM-1run を用いた相対的評価（順位）は、R による評価とあまり変わらず、プールに入れる run を減らしても、相対的評価にはほとんど影響がないことがわかる。

以上のことから、各チームの優先順位の高い run をプールすることによって、効率的なプーリングが行え、評価の公平性にもあまり影響はないと考えられる。

表 2: Mean Average Precisions and Rankings of J-J Runs for SLIR

run	J-J-1	J-J-2	J-J-3	J-J-4	J-J-5	J-J-6	J-J-7
method	Pr	Pr	Pr	Pr	Pr	Pr	VS
R	1	2	3	4	5	6	7
	0.3965	0.3939	0.3891	0.3516	0.3492	0.3350	0.3264
PS	1	2	3	4	5	6	7
	0.4001	0.3962	0.3961	0.3547	0.3533	0.3392	0.3308
PB	2	3	1	5	4	6	7
	0.3984	0.3956	0.4066	0.3500	0.3526	0.3363	0.3188
PM	1	2	3	5	6	4	7
	0.4098	0.4009	0.3888	0.3600	0.3445	0.3832	0.3215
R-PSonly	1	3	2	5	4	6	7
	0.4041	0.3980	0.4013	0.3548	0.3556	0.3487	0.3263
R-PBonly	1	2	3	4	5	6	7
	0.3989	0.3950	0.3928	0.3533	0.3516	0.3376	0.3290
R-PMonly	1	2	3	4	5	6	7
	0.3967	0.3942	0.3898	0.3521	0.3494	0.3354	0.3268
R-1run	1	2	3	4	5	6	7
	0.3985	0.3979	0.3941	0.3534	0.3527	0.3390	0.3295
PS-1run	1	3	2	5	4	6	7
	0.4033	0.4029	0.4032	0.3585	0.3595	0.3454	0.3362
PB-1run	3	2	1	5	4	6	7
	0.3907	0.4005	0.4158	0.3510	0.3568	0.3360	0.3228
PM-1run	1	2	3	5	6	4	7
	0.4109	0.4035	0.3920	0.3615	0.3476	0.3885	0.3248

run	J-J-8	J-J-9	J-J-10	J-J-11	J-J-12	J-J-13	J-J-14
method	Pr	Pr	Pr	VS	VS	Pr	VS
R	8	9	10	11	12	13	14
	0.3102	0.2682	0.2338	0.2228	0.1582	0.0854	0.0517
PS	8	9	10	11	12	13	14
	0.3149	0.2728	0.2375	0.2263	0.1588	0.0861	0.0523
PB	8	9	10	11	12	13	14
	0.3097	0.2681	0.2343	0.2270	0.1575	0.0782	0.0500
PM	8	9	10	11	12	13	14
	0.3112	0.2889	0.2439	0.2388	0.1719	0.0826	0.0535
R-PSonly	8	9	10	11	12	13	14
	0.3144	0.2744	0.2376	0.2297	0.1611	0.0820	0.0513
R-PBonly	8	9	10	11	12	13	14
	0.3131	0.2709	0.2362	0.2248	0.1588	0.0859	0.0520
R-PMonly	8	9	10	11	12	13	14
	0.3106	0.2685	0.2340	0.2232	0.1584	0.0854	0.0517
R-1run	8	9	10	11	12	13	14
	0.3142	0.2710	0.2359	0.2256	0.1583	0.0857	0.0528
PS-1run	8	9	10	11	12	13	14
	0.3211	0.2778	0.2402	0.2302	0.1582	0.0861	0.0540
PB-1run	8	9	10	11	12	13	14
	0.3153	0.2685	0.2332	0.2314	0.1561	0.0780	0.0504
PM-1run	8	9	10	11	12	13	14
	0.3134	0.2930	0.2444	0.2420	0.1717	0.0832	0.0537

表 3: Mean Average Precisions and Rankings of x-J Runs for BLIR

run	E-J-1	E-J-2	C-J-1	E-J-3	E-J-4	C-J-2	E-J-5	E-J-6
method	Pr	Pr	Pr	Pr	Pr	VS	VS	Pr
R	1	2	3	4	5	6	7	8
	0.3382	0.2206	0.1519	0.1094	0.0607	0.0460	0.0402	0.0000
PS	1	2	3	4	5	6	7	8
	0.3417	0.2171	0.1483	0.1022	0.0639	0.0434	0.0414	0.0000
PB	1	2	3	4	5	6	7	8
	0.3642	0.2385	0.1608	0.1173	0.0681	0.0486	0.0410	0.0000
PM	1	2	3	4	5	6	7	8
	0.3150	0.2318	0.1632	0.1066	0.0889	0.0507	0.0420	0.0000
R-PSonly	1	2	3	4	5	6	7	8
	0.3525	0.2301	0.1581	0.1143	0.0662	0.0478	0.0408	0.0000
R-PBonly	1	2	3	4	5	6	7	8
	0.3392	0.2186	0.1502	0.1046	0.0629	0.0454	0.0410	0.0000
R-PMonly	1	2	3	4	5	6	7	8
	0.3386	0.2207	0.1518	0.1095	0.0609	0.0461	0.0402	0.0000
R-1run	1	2	3	4	5	6	7	8
	0.3418	0.2228	0.1526	0.1084	0.0623	0.0455	0.0406	0.0000
PS-1run	1	2	3	4	5	6	7	8
	0.3468	0.2204	0.1497	0.1022	0.0656	0.0424	0.0421	0.0000
PB-1run	1	2	3	4	5	6	7	8
	0.3736	0.2458	0.1642	0.1146	0.0786	0.0469	0.0415	0.0000
PM-1run	1	2	3	4	5	6	7	8
	0.3167	0.2340	0.1642	0.1039	0.0900	0.0477	0.0421	0.0000

表 4: Number of Documents in Pools

pool	R	PS	PB	PM	R-1run	PS-1run	PB-1run	PM-1run
total	72954	36592	42216	21318	47198	24346	25909	17779
ave	100	49.5	59.1	30.2	64.7	33.1	36.3	25.3
r<50	100	49.9	57.5	29.6	64.9	33.3	35.4	24.4
50≤r<100	100	46.1	63.9	29.1	60.2	28.9	37.6	24.4
100≤r	100	50.8	61.5	33.9	67.7	36.3	39.0	29.4

表 5: Number of Relevant Documents in Pools

pool	R	PS	PB	PM	R-1run	PS-1run	PB-1run	PM-1run
total	2538	2336	1834	1396	2332	2112	1657	1342
ave	100	97.0	77.7	66.5	96.6	92.6	72.5	65.1
r<50	100	98.9	79.3	70.8	98.3	96.4	74.5	69.6
50≤r<100	100	97.4	83.8	69.5	97.8	92.2	80.7	68.1
100≤r	100	89.0	66.1	45.9	88.4	76.9	57.4	43.7

4 まとめ

本稿では、パラレルコーパスを使用する言語横断検索のテストコレクションを構築する際に、多言語文書であるという特性を利用して、適合文書候補をどうすれば効果的かつ公平に収集することができるかという観点から、NTCIR-3の言語横断検索タスクの提出結果を用いて、プーリングと評価実験を行なった。

実験の結果から、以下のことがわかった。

(1) サブタスク混合プーリングで作成した適合文書リストと、各サブタスクごとにプーリングを行なって作成した適合文書リストのそれぞれ用いて、日本語文書を検索対象とするSLIRとBLIRのrunの評価を行なったところ、使用する適合文書リストによって相対的評価（順位）が異なる場合があることがわかった。これは、各プールの網羅性に差があるためであると考えられる。

(2) 各サブタスクごとにプールを行なった場合には、SLIRのrunからのプールではサブタスク混合プーリングとほぼ同程度の網羅性で適合文書を集めることができ、また、各runの相対的評価にも影響がないことがわかった。このことから、効率的なプーリングを行うためには、SLIRからプーリングを行い、BLIRとMLIRで網羅性を補完することが考えられる。

(3) より効率的なプーリングを行うためには、プールする文書を各チームの優先順位の高いrunのみに制限し、各サブタスクのプーリングで使用するrunの数を減らすことが考えられるが、run数の削減と選択方法によっては評価に影響を与える場合もある。

現在、NTCIRプロジェクトでは、2003年3月～2004年5月という期間で、NTCIRワークショップ4を開催しており、CLIRタスクの検索結果提出は2003年11月1日に予定されている。本稿の結果に基づいた、網羅的で効率的な適合文書リストの作成を行いたい。

今後の課題としては、適合文書リストの網羅性と公平性という観点から、適合文書が少ない検索課題と多い検索課題を同じ基準で計るのではなく、追加検索・追加プーリングなども利用した、適合文書の多い検索課題の網羅性を高めるようなプーリングを考えたい。また、それぞれのrunの、ユニークな適合文書を見つけることへの貢献度(unique contribution)とその評価への影響についても実験を行い、考察したい。

参考文献

[1] Buckley, C., Voorhees, E., "Tutorial: Theory and Practice in Text Retrieval System Evaluation". ACM–SIGIR'99, Berkeley, CA U.S.A, 1999.

[2] Chen, K. et al., "Overview of CLIR Task at the Third NTCIR Workshop". In Proc. NTCIR Workshop 3, Tokyo. (In printing)

[3] Gilbert, G., Sparck Jones, K., "Statistical Bases of Relevance Assessment for the 'Ideal' Information Retrieval Test Collection". BL R&D Report 5481, Cambridge, England., 1979.

[4] Kuriyama, K. et al., "Pooling for a Large-Scale Test Collection: An Analisis of the Search Results from the First NTCIR Workshop". Information Retrieval, Vol.5, No.1, pp.41–59, 2002.

[5] 栗山和子ほか., "大規模テストコレクションNTCIR-2の構築: 対話型追加検索と言語横断的プーリングの効果". 情報処理学会論文誌: データベース, Vol.43, No.SIG2 (TOD13), pp.48–59, 2002.

[6] 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:NTCIR-3言語横断検索タスクの分析". 情報処理学会研究報告, 2003-FI-72/NL-157, Vol.2003, No.98, pp.91–98, 2003.

[7] NTCIR (NII-NACSIS Test Collection for IR Systems) Project. http://research.nii.ac.jp/ntcir/

[8] NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop 2002. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/

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

[10] Voorhees, E. The Eleventh Text Retrieval Conference (TREC 2002), NIST Special Publication SP 500-251, Maryland, U.S.A., 2002.