

大規模テストコレクション NTCIR-2 の構築 - 言語横断的プーリングの評価への影響 -

栗山和子[†] 吉岡真治[‡] 神門典子[†]

[†] 国立情報学研究所 [‡] 北海道大学大学院工学研究科

[†] {kuriyama,kando}@nii.ac.jp [‡] yoshioka@db-ei.eng.hokudai.ac.jp

概要. 大規模テストコレクション NTCIR-2 の構築において、正解文書リストは、プーリング法に基づいて作成された。本稿では、NTCIR-2 の正解文書リストの作成過程で行なった言語横断的プーリングおよび対話型検索システムを用いた追加検索が、NTCIR-2 を使用した検索システムの評価に与える影響について考察する。

本研究では、NTCIR-2 の正解文書リストと、NTCIR ワークショップ 2 の参加チームの提出結果を用いて評価実験を行なった。まず、NTCIR-2 の最終的な正解文書リスト F と、 F から追加の対話型検索 I だけで見つかった文書を除いたリスト $F - I$ を用いて、提出結果の評価を行なった。次に、各サブタスクごとの提出結果からのプーリングを行ない、このサブタスクごとのプールを正解文書リストとして評価を行なった。

結果として、いずれの文書リストを正解文書リストとして提出結果の評価を行なっても、提出結果の相対的な順位はほとんど変化せず、プーリング法に基づいて作成したテストコレクションの信頼性を確めることができた。

Construction of a Large Scale Test Collection NTCIR-2 - The Effect of Cross-Lingual Pooling on Evaluation -

Kazuko Kuriyama[†] Masaharu Yoshioka[‡] Noriko Kando[†]

[†] National Institute of Informatics (NII)

[‡] Graduate School of Engineering, Hokkaido University

[†] {kuriyama,kando}@nii.ac.jp [‡] yoshioka@db-ei.eng.hokudai.ac.jp

Abstract. The purpose of this study is to examine whether there is an effect on the relative evaluation of the IR systems using the relevance judgments made by the pooling method and additional interactive searches.

We carried out experiments using different lists of relevance judgments and search results submitted for the test of the 2nd NTCIR Workshop. First, we evaluated the search results using the list of the final relevance judgments F of NTCIR-2 and $F - I$, that is, the F without the unique relevant documents found by the additional interactive searches I . Second, we made pools from the search results for each of the sub-tasks and evaluated the search results using the pools as lists of relevance judgments.

Almost the same rankings of the search results were produced by using the pools as lists of relevance judgments for system evaluation. Therefore our results verified the reliability of test collection as an evaluation tool, which was based on pooling method.

1 はじめに

1.1 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) 正解文書リストの網羅性
- (2) 正解文書リストの公平性
- (3) 正解判定の無矛盾性

筆者らは、テストコレクション NTCIR-1 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た[2],[3],[4]。(1) について、NTCIR-1 では、上位一定数のプーリングと対話型検索システムを用いた追加検索によって正解文書リストの網羅性を高めることができた。具体的には、正解文書が100件以上の検索課題については、各提出結果からの上位100件ずつのプーリングでは、NTCIR ワークショップ 1[6]の予備テストでは全正解文書数の51.9%、評価テストでは76.4%しか網羅できなかったが、対話型検索システムで手作業で再現率を重視した追加検索を行なった結果を追加すると、予備テストでは89.7%、評価テストでは98.0%をカバーすることができた。また、(2) に関して、予備テストと評価テストの提出結果のそれぞれについて、プーリングと追加検索によって作成した数種類の正解文書リストを用いて評価を行ない、プーリングと追加検索による正解文書リストの作成が相対的なシステム評価に影響を与えないことを確めた。

本稿では、NTCIR-1 構築の経験を踏まえて、NTCIR-2 について、プーリング法と追加検索による正解文書リストの作成が、相対的なシステム評価にどのような影響を与えるか考察する。

以下、2 節では、日本語・英語検索タスクのサブタスクと提出結果、および、プーリングによる正解文書リストの作成法について述べる。3 節では、プーリング実験と作成した数種類のプールを用いた評価実験について述べ、4 節では、本研究でわかったことをまとめる。

1.2 テストコレクションとプーリング法

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を構築するとき、各検索課題について文書データベース中の全ての文書の正解判定を行ない、完全に網羅的な正解文書リストを作成することが理想ではあるが、数万から数十万件の文書を含む大規模なデータベースに対してそのようなことを行なうのは、実際には不可能であり、別の方法で正解文書を網羅的に収集する必要がある。

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、プーリング法 (Gilbert and Sparck Jones 1979[1]) が、効率的で効果的な方法として有名であり、世界的な評価型ワークショップ TREC(Text REtrieval Conference)[8],[9] では、1992 年から、プーリング法によって大規模テストコレクションを作成している。

プーリング法では、異なる検索手法を用いた様々な検索システムによって検索された結果のそれぞれから、各検索課題ごとに上位一定数 (X 件) ずつをプールし、プール中の全ての文書について人間の正解判定者が正解判定を行なう。プールに含まれない文書は判定されず、不正解と仮定されるため、どのように、正解文書候補を網羅的に、また、どのような検索システムに対しても公平になるように集めるかが問題となる。

2 NTCIR-2 の正解文書リスト

NTCIR-2 の正解文書リストは以下のような手順で作成された。(1) プーリング法を用いて正解文書候補を収集する、(2) 人間の正解判定者によって正解文書候補の正解判定を行なう、(3) ある一定数以上の正解文書を持つ検索課題について、正解文書リストの網羅性を高めるため、対話型検索システムを用いて、再現率を重視した検索を行ない、追加の正解文書候補を収集する、(4) 追加の正解文書候補について正解判定を行なう。次項以下で、サブタスクと正解文書作成手順の各ステップについて詳しく説明する。

2.1 日本語・英語検索タスクのサブタスクと検索対象文書

2.1.1 サブタスク

本項以下では、NTCIR ワークショップ 2[7]の「日本語・英語検索タスク (Japanese & English IR Task)」

を「JEIR タスク」と略記する。JEIR タスクには 2 つのサブカテゴリがあり、そのサブカテゴリ中のサブタスクは以下の通りである。

単言語検索タスク：

- J-J タスク：日本語検索課題を用いて日本語文書を検索する、
- E-E タスク：英語検索課題を用いて英語文書を検索する。

言語横断検索タスク：

- E-J タスク：英語検索課題を用いて日本語文書を検索する、
- J-E タスク：日本語検索課題を用いて英語文書を検索する、
- J-J,E タスク：日本語検索課題を用いて日本語文書と英語文書を検索する、
- E-J,E タスク：英語検索課題を用いて日本語文書と英語文書を検索する。

2.1.2 サブタスクにおける検索対象文書

JEIR タスクでは、J コレクションと E コレクションという 2 つの文書コレクションが使用された。J コレクションと E コレクションは、国立情報学研究所 (NII) が提供している「学会発表データベース」と「科学研究費補助金研究成果概要データベース」の一部を抽出したものである。元のデータベースの一部は、日本語文書と英語文書が対訳として組になっている。

J コレクションは、*ntc1-j1.mod*、*ntc2-j1g*、*ntc2-j1k* という 3 つの文書セットから成り、E コレクションは、*ntc1-e1.mod*、*ntc2-e1g*、*ntc2-e1k* という 3 つの文書セットから成る。*ntc1-j1.mod*、*ntc1-e1.mod*、*ntc2-j1g*、*ntc2-e1g* は「学会発表データベース」から抽出された文書セット、*ntc2-j1k* と *ntc2-e1k* は「科学研究費補助金研究成果概要データベース」のから抽出された文書セットである。

日本語文書と英語文書が元のデータベース中で 1 組の対訳になっているとき、それらは同じ文書番号 (ACCN) が付与されている。NTCIR-2 では、J コレクションと E コレクションを独立な文書コレクションとして扱うため、組になっている日本語文書と英語文書を分け、英語文書には新たな ACCN を付与した。例

えば、元の文書の ACCN が「gakkai-000040700」であるとき、英語文書には、新たな ACCN 「gakkai-e-000104007」を付与した。

各文書コレクションの文書数と、文書コレクション中に含まれる、元のデータベースでは組になっている文書数を表 1 に示す。また、JEIR タスクのサブタスクと、NTCIR-2 の検索課題、検索対象文書、および正解文書リストのファイル名との対応を表 2 に示す。

表 1: 日本語・英語検索タスクで使用された文書コレクションと文書数

Document Collections	Number of docs
<i>ntc1-j1.mod</i>	332,918
<i>ntc1-e1.mod</i>	187,080
<i>pairs in ntc1-j1&e1</i>	181,485
<i>ntc2-j1g</i>	116,177
<i>ntc2-e1g</i>	77,433
<i>pairs in ntc2-j1g&e1g</i>	74,180
<i>ntc2-j1k</i>	287,063
<i>ntc2-e1k</i>	57,545
<i>pairs in ntc2-j1k&e1k</i>	57,510

2.1.3 提出結果からのプーリング

JEIR 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、J コレクションあるいは (および) E コレクションを検索し、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。

図 1 に JEIR タスクのプーリングの過程を示す。JEIR タスクのプーリングでは、(1) 全てのサブタスクの run について、上位 *X* 件の文書をプーリングし、(2) プーリング中の日本語文書と英語文書の ACCN を、元の ACCN に戻す。例えば、日本語文書の ACCN 「gakkai-j-000040700」と英語文書の ACCN 「gakkai-e-000104007」をそれぞれ元の ACCN 「gakkai-000040700」(数字部分は日本語文書の文書番号と同一)に戻し、英語文書を日本語文書に対応付ける。この過程は、タスク横断的・言語横断的であるので、本稿では、言語横断的プーリングという言葉で表わす。

プーリングは、全ての run について行なったわけではなく、次のような理由から、いくつかの run は

表 2: サブタスク、検索課題、文書、正解文書リストの関係

Task	Topics	Document Collections	Relevance Judgments	
			Level1 (S or A)	Level2 (S, A or B)
J-J	topic-j101-149	ntcl-j1.mod, ntc2-j1g, ntc2-j1k	rel1_ntc2-j2_0101-0149txt	rel2_ntc2-j2_0101-0149
E-J	topic-e101-149			
J-E	topic-j101-149	ntcl-e1.mod, ntc2-e1g, ntc2-e1k	rel1_ntc2-e2_0101-0149	rel2_ntc2-e2_0101-0149
E-E	topic-e101-149			
J-J,E	topic-j101-149	ntcl-j1.mod, ntc2-j1g, ntc2-j1k,	rel1_ntc2-je2_0101-0149	rel2_ntc2-je2_0101-0149t
E-J,E	topic-e101-149	ntcl-e1.mod, ntc2-e1g, ntc2-e1k		

topic-j101-149 は日本語検索課題のリスト、 topic-e101-149 は英語検索課題のリスト。

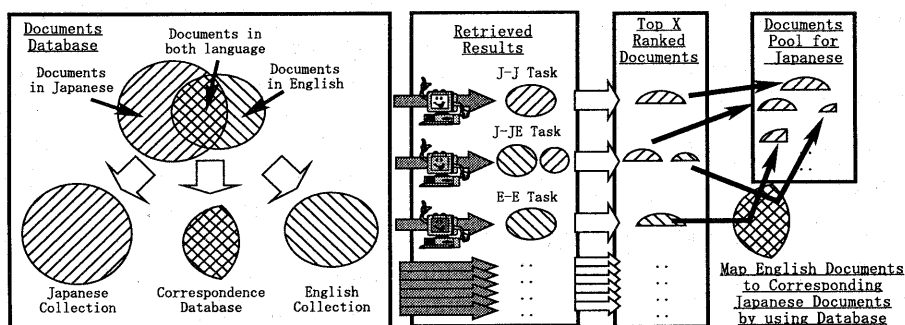


図 1: 言語横断的プーリングの過程

プーリングには使用しなかった。

(a) 表 2 からわかるように、J-J,E タスクと E-J,E タスクでは、J コレクションと E コレクションを使用している。E コレクションの文書は、J コレクションの文書の一部と対訳になっているため、ACCN を変換した後、重複した ACCN の文書を除くと、J-J,E タスクと E-J,E タスクの 1 つずつの run からプールされる文書数は実際は上位 X 件よりも少なくなり、プーリングの効率が落ちる。効率的なプーリングを行なうため、J-J,E タスクと E-J,E タスクの run については、同じ検索システムを用いた同じ参加チームの run が、他の 4 つのタスクのうちのいずれにも含まれていない場合を除いて、プーリングに使用しないことにした。

(b) 一つの参加チームが同じ検索システムを用いた run を複数提出している場合には、同じシステムは同じような正解文書と異なる不正解文書を見つけるので

はないかという仮定から、効率的なプーリングと正解判定のために、参加チームが提出時に run に付けた優先順位に基づいて、一つの参加チームにつき、タスク別に、優先順位の高い順に 2 つまでの run をプーリングに使用した。表 3 に実際のプーリングに使用した run の個数を示す。

各 run からプールされる文書数 X は一つの検索課題に対してプールされた総文書数が 2000~2500 件以下になるように検索課題ごとに 70,80,90,100 のいずれかに調整した。同一の検索課題については、どの run から同じ上位 X 件をプールしている。

プーリングの過程で、ステップ (1) で作成されたプールのうち、日本語文書部分を $J1$ 、英語文書部分を $E1$ 、ステップ (2) で $J1$ から変換された英語文書を $E1$ に追加したプールを $E2$ 、 $E1$ から変換された日本語文書を $J1$ に追加したプールを $J2$ とすると、それぞれに含まれる正解文書数の、最終正解文書リ

表 3: 提出された run の数とプールに使用された run の数

Task	Submitted runs	Pooled runs
J-J	93	29
J-E	41(1)	23(1)
J-J,E	15(1)	1
E-E	18	12
E-J	30	17
E-J,E	11	0

J-E タスクと J-J,E タスクの括弧内の数字 n は、JEIR タスクのオーガナイザの提出した run の数である。

スト F のうちの単言語の正解文書リスト $J(F)$ 、 $E(F)$ に対する割合は、 $J1:89.6\%$ 、 $E1:91.8\%$ 、 $J2:96.6\%$ 、 $E2:98.1\%$ となっている。このことから、言語横断的プーリングが、対訳になっている文書の中で新たな正解文書を見つけるのにある程度効果的であることがわかった。

2.2 正解判定

正解判定作業では、検索課題ごとに、主判定者 1 人と副判定者 1 人の計 2 人の正解判定者が正解判定とクロスチェックを行なった。主判定者は、基本的には、その検索課題の作成者である。最終判定結果は、2 人の判定者の協議に基づき、主判定者が決定した。正解判定は、高正解（高適合）highly-relevant (S)、正解 relevant (A)、部分的正解 partially-relevant (B)、不正解 non-relevant (C) の 4 つのレベルで行なわれた。

検索課題を作成する時点で、検索課題作成者には、対話型検索システムを用いて予備的な検索を行なってもらい、5~10 件の正解文書候補のリストを提出してもらった。この予備検索結果文書のうち、参加者の run から作成したプールを P としたとき、 P に含まれていない、予備検索だけで見つかった文書の集合をプール PP とする。正解判定は、 P と PP を合わせた正解文書候補リスト $P+PP$ に対して行ない、全 49 件の検索課題のうちの 29 件の検索課題についてはクロスチェックを行なった。最後に全ての検索課題について、各主判定者が最終的なチェックを行い、最終判定結果を決定した。

JEIR タスクでは、TREC で使用されている評価

スクリプトを用いて、各 run の評価を行なった。その評価スクリプト trec_eval による評価計算は、2 値の正解（正解：1、不正解：0）に対して行なわれる。そのため、JEIR タスクでは、正解判定結果の 4 段階の正解を 2 段階に直し、2 つの異なる正解レベルの正解文書リストを作成した。すなわち、レベル 1 として、「S」と「A」を正解 (1) とし、「B」と「C」を不正解 (0) とした正解文書リスト (Level1) と、レベル 2 として、「S」、「A」、「B」を正解 (1) とし、「C」を不正解 (0) とした正解文書リスト (Level2) を作成した。本稿では、レベル 2 の正解「S」、「A」、「B」を「正解文書」として定義し、以下では、その意味で使用する。

2.3 追加検索

参加者からの run のプールについての正解判定後、正解文書を 110 件以上持つ検索課題あるいは各 run について上位 70 件の文書をプールした検索課題 16 件については、正解文書リストの網羅性を高めるため、対話型検索システムを用いて、図書館情報学専攻の大学院生によって、再現率を重視した追加検索を行なった。そして、その追加検索で新たに見つかった文書の集合 I について主判定者によって追加判定を行ない、 I を $P+PP$ に加えて最終的な正解文書リスト F を作成した。

run からのプール P 、予備検索のプール PP 、追加検索のプール I 、の正解文書数の、最終正解文書リスト F のうちの日本語部分 $J(F)$ と英語部分 $E(F)$ に対する検索課題ごとの割合の平均を表 4 に示す。ave%16 追加検索を行なった 16 件の検索課題についての平均、ave%all は、全検索課題についての平均である。

表 4 からわかるように、正解文書を 110 件以上持つ、あるいは、プール数が 70 件であった検索課題 16 件についての $J(P)$ と $E(P)$ の、それぞれ $J(F)$ と $E(F)$ に対する割合の平均は、91.4% と 95.3% であり、NTCIR ワークショップ 1 の予備テストと評価テストでの run からだけのプールに含まれる正解文書の割合よりもかなり大きい。これは、NTCIR ワークショップ 2 では、NTCIR ワークショップ 1 よりも多くの run が提出されたため、run からのプールの網羅性が高まったからだと考えられる。しかしながら、再現率重視の追加の対話型検索は、日本語の正解文書 $J(F)$ の 8.4% を見つけており、ある程度、網羅性を高めるのに効果的であったと言える。

表 4: プール中の正解文書数の割合

topic	J(P)	J(PP)	J(I)	J(F)	E(P)	E(PP)	E(I)	E(F)
ave % all	96.6	0.1	3.3	100	98.1	0.2	1.7	100
ave % 16	91.4	0.2	8.4	100	95.3	0.7	4.0	100

ave%16 追加検索を行なった 16 件の検索課題についての F に対する割合の平均であり、ave%allは、全検索課題についての平均である。

3 プーリングおよび評価の実験

3.1 追加検索について

対話型検索システムを用いた追加検索がシステム評価に影響を与えるかどうか調べるために、J-J タスクの run について、最終的な正解文書リスト F と、 F から対話型検索でのみ見つかった文書集合 I を除いた文書リスト $F-I$ を正解文書リストとして用いて評価実験を行なった。また、追加型検索結果 I は、正解文書収集において、一つの検索システムからの複数の run と同様の働きをしているのではないかと考え、一つのシステムからの複数の run がシステム評価に影響を与えるかどうか調べた。2 つの異なる検索システム（参加チーム）CRL と DOVE の run を用いて、CRL の run のみが見つけた正解文書を CRL 、DOVE が見つけた正解文書 $DOVE$ として、最終正解文書リスト F から除いた正解文書リスト $F-CRL$ と $F-DOVE$ をそれぞれ作り、評価実験を行なった。

3.2 サブタスクごとのプールについて

サブタスクごとの run からのプールが、正解文書リストの網羅性にどれくらい貢献しているか、また、相対的なシステム評価に影響を与えるのかどうか調べるため、サブタスク J-J、E-J、E-E、E-J について、サブタスクごとのプール $P(task)$ と、最終正解文書リスト F から、各サブタスクのプールだけに含まれる文書の集合を除いたプール $F-P(task)$ を作成し、評価実験を行なった。

その各プール中の正解文書数の正解文書全体に対する割合の平均を表 5 に示す。

3.3 評価実験

最終的な正解文書リスト F 、追加検索についてのプール $F-I$ 、 $F-CRL$ 、 $F-DOVE$ 、サブタスクごとのプール $P(J-J)$ 、 $P(J-E)$ 、 $P(E-E)$ 、

$P(J-E)$ 、 F からサブタスクごとのプール中のユニークな文書を除いたプール $F-P(J-J)$ 、 $F-P(J-E)$ 、 $F-P(E-E)$ 、 $F-P(E-J)$ を正解文書リストとして用いて評価を行なったときの、J-J タスクの run の平均精度の平均とそれによる順位を表 6 に示す。

評価には、提出された全 run のうち、参加チームにつき 1 つずつ、検索に使用した検索課題のフィールドが「DESCRIPTION(D)」である run を用いた。DESCRIPTION を使用した run を提出していないチームについては、使用フィールドに DESCRIPTION を含む run を代わりとして用いた。

表 6 から、 F 、 $F-I$ 、 $F-CRL$ 、 $F-DOVE$ を用いて評価した結果、各 run の順位は全く同じになることがわかる。特に、最終正解文書リスト F で評価しても、 F から追加の対話型検索だけが見つけた文書集合 I を除いたリスト $F-I$ で評価しても、対話型手法を用いている 4 つの run のいずれの順位も変わらないことに注目したい。したがって、追加検索は、正解文書リストの網羅性の向上には貢献しているが、システム評価には影響を与えないということがわかった。

また、表 6 から、 F とその他のサブタスクごとのプールを用いた評価による順位も、ほとんど変わらないことがわかる。ゆえに、言語横断的プーリングはシステム評価にほとんど影響を与えないことがわかった。

4 おわりに

言語横断的プーリングと、対話型検索システムによる再現率重視の追加検索を用いて作成された正解文書リストの公平性を調べるために、NTCIR ワークショップ 2 の日本語・英語検索タスクの評価テストの提出結果 (run) を用いて、実験的なプーリングと評価を行なった。実験によって、次のようなことがわかった。

表 5: サブタスクごとのプール中の正解文書数の割合

topic	$J(P(J-J))$	$J(P(J-E))$	$J(P(E-E))$	$J(P(E-J))$	$E(P(J-J))$	$E(P(J-E))$	$E(P(E-E))$	$E(P(E-J))$
ave % F	81.6	35.3	31.9	73.5	72.9	86.5	78.8	67.3
topic	$J(F-P(J-J))$	$J(F-P(J-E))$	$J(F-P(E-E))$	$J(F-P(E-J))$	$E(F-P(J-J))$	$E(F-P(J-E))$	$E(F-P(E-E))$	$E(F-P(E-J))$
ave % F	87.8	97.3	98.6	94.4	97.4	93.3	96.9	98.5

(1) 追加の対話型検索のシステム評価への影響：追加検索は、正解文書を 110 件以上持つ検索課題および上位 70 件だけをプールした検索課題との合計 16 件について、行なわれた。最終的な正解文書リスト F と、 F から追加検索で見つかった正解文書 I を除いた $F-I$ を用いて、各 run を評価し、検索課題全体に対する平均精度の平均で順位付けを行なった結果、各 run の相対的な順位は全く変わらなかった。この実験的な結果は、対話型検索システムによる追加検索を行なっても、システム評価には影響がない [2],[3],[4] という仮定を補強するものである。

(2) 追加検索の必要性：追加検索は、正解文書リストの網羅性を高めるのにある程度有効であるが、プーリングに使用できる run の個数が十分多く、多様であれば、追加検索は必要ではないかもしれない。しかし、何件の run をプーリングに使用すれば十分であるかは不明であるので、その件数と多様性について検討する必要がある。

(3) 言語横断的プーリングのシステム評価への影響：サブタスクごとのプールを作成し、そのプールを用いて各 run を評価したとき、各プールがどれくらい正解文書リストの網羅性に貢献しているかは異なるので（表 5 参照）、各 run の平均精度の平均値の絶対的な大きさは異なる（表 6 参照）。しかし、どのプールを用いても、検索課題全体に対する平均精度の平均での相対的な順位は、最終正解文書リスト F による順位とほとんど変わらなかった。

結果として、言語横断的プーリング、すなわち、単言語文書の検索結果からプールした正解文書候補を、異なる言語の文書における正解文書候補として加えたことは、システム評価にほとんど影響を与えず、また、正解文書候補を効率的に集めるためにある程度有用であることがわかった。

謝辞

システム評価について貴重なご助言をしてくださった、駿河台大学岸田和助教授に深く感謝いたします。

本研究は、日本学術振興会未来開拓学研究推進事業「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] Gilbert, G., Sparck Jones, K., "Statistical Bases of Relevance Assessment for the 'Ideal' Information Retrieval Test Collection". BL R&D Report 5481, 1979.
- [2] 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.
- [3] 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.
- [4] 栗山和子ほか, "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". 情報処理学会第 59 回全国大会, pp.3-105-3-106, 1999.
- [5] NII-NACSIS Test Collection for Information Retrieval systems.
<http://research.nii.ac.jp/ntcir/>
- [6] NTCIR Workshop 1: Proceedings of the First NTCIR Workshop on Retrieval in Japanese Text Retrieval and Term Recognition, Tokyo, Japan, Aug.30-Sep.1, 1999, ISBN 4-924600-77-6.
- [7] NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, Tokyo, Japan, March.7-9, 2001, ISBN 4-924600-89-X.
- [8] Text REtrieval Conference (TREC).
<http://trec.nist.gov/> (visited January 12th, 2001).
- [9] Voorhees, E., Harman, D. eds. The Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-242, Maryland, U.S.A., 2000.

表 6: J-J タスクの run の平均精度の平均と順位

Run-ID	DOVE9	CRL16	LAPIN6	JSCB1	R2D22	sstut1	FXSD2	sstut6	apljj2	DOVE3
Query Field	D N	D	D	D	D	D	T D N C F	D	D	D
Method	interact	auto	auto	auto	auto	auto	interact	interact	auto	auto
F	1	2	3	4	5	6	7	8	9	10
	0.4138	0.3686	0.3620	0.3377	0.3051	0.3024	0.2847	0.2810	0.2687	0.2683
F-I	1	2	3	4	5	6	7	8	9	10
	0.4173	0.3720	0.3659	0.3396	0.3085	0.3059	0.2863	0.2814	0.2713	0.2713
F-CRL	1	2	3	4	5	6	7	8	9	10
	0.4120	0.3686	0.3634	0.3382	0.3061	0.3039	0.2842	0.2811	0.2692	0.2685
F-DOVE	1	2	3	4	5	6	7	8	9	10
	0.4097	0.3682	0.3630	0.3382	0.3055	0.3035	0.2838	0.2798	0.2685	0.2677
P	1	2	3	4	5	6	7	8	10	9
	0.4175	0.3721	0.3660	0.3398	0.3086	0.3061	0.2863	0.2815	0.2714	0.2715
P(J-J)	1	2	3	4	5	6	7	8	9	10
	0.4565	0.4088	0.4056	0.3747	0.3425	0.3345	0.3130	0.3019	0.2970	0.2964
P(J-E)	1	2	3	4	5	7	6	11	8	9
	0.1938	0.1777	0.1661	0.1595	0.1439	0.1351	0.1425	0.1166	0.1258	0.1251
P(E-E)	1	2	3	4	5	7	6	12	10	9
	0.1951	0.1801	0.1682	0.1652	0.1512	0.1354	0.1464	0.1177	0.1261	0.1271
P(E-J)	1	3	2	4	5	6	8	10	9	7
	0.4575	0.4056	0.4082	0.3713	0.3428	0.3228	0.3011	0.2922	0.2945	0.3031
F-P(J-J)	1	2	3	4	5	6	7	8	10	9
	0.4213	0.3729	0.3709	0.3441	0.3104	0.3013	0.2807	0.2825	0.2715	0.2720
F-P(J-E)	1	2	3	4	5	6	7	8	10	9
	0.4175	0.3735	0.3686	0.3428	0.3100	0.3066	0.2871	0.2770	0.2711	0.2770
F-P(E-E)	1	2	3	4	5	6	7	8	9	10
	0.4145	0.3710	0.3654	0.3398	0.3071	0.3051	0.2856	0.2816	0.2704	0.2696
F-P(E-J)	1	2	3	4	5	6	7	8	9	10
	0.4218	0.3771	0.3722	0.3468	0.3139	0.3114	0.2915	0.2855	0.2750	0.2749

Run-ID	FXSD1	Brkly2	SRGDU1m	STIX6	MPIN55	smlab	sato2	WUSKL	OASIS9	trans4
Query Field	D	D	D	D	D	D N C	D	D	D	D
Method	auto	auto	auto	auto	auto	interact	auto	auto	auto	auto
F	11	12	13	14	15	16	17	18	19	20
	0.2567	0.2432	0.2309	0.2101	0.2067	0.2059	0.2016	0.1591	0.1210	0.0138
F-I	11	12	13	14	15	16	17	18	19	20
	0.2579	0.2454	0.2329	0.2121	0.2093	0.2076	0.2044	0.1599	0.1205	0.0141
F-CRL	11	12	13	14	15	16	17	18	19	20
	0.2572	0.2437	0.2318	0.2108	0.2079	0.2052	0.2030	0.1591	0.1198	0.0140
F-DOVE	11	12	13	14	15	16	17	18	19	20
	0.2568	0.2427	0.2314	0.2102	0.2079	0.2051	0.2019	0.1593	0.1194	0.0139
P	11	12	13	14	15	16	17	18	19	20
	0.2579	0.2455	0.2330	0.2122	0.2093	0.2077	0.2044	0.1600	0.1206	0.0141
P(J-J)	11	12	13	16	14	15	17	18	19	20
	0.2831	0.2750	0.2584	0.2339	0.2367	0.2342	0.2251	0.1753	0.1337	0.0156
P(J-E)	12	10	15	13	16	17	14	18	19	20
	0.1152	0.1241	0.1070	0.1099	0.1017	0.1005	0.1088	0.0906	0.0390	0.0078
P(E-E)	11	8	14	15	16	17	13	18	19	20
	0.1190	0.1278	0.1094	0.1088	0.1055	0.1012	0.1144	0.0908	0.0396	0.0082
P(E-J)	11	12	13	15	14	17	16	18	19	20
	0.2780	0.2707	0.2559	0.2392	0.2401	0.2190	0.2245	0.1759	0.1324	0.0158
F-P(J-J)	11	12	13	14	15	17	16	18	19	20
	0.2560	0.2472	0.2361	0.2180	0.2166	0.2003	0.2074	0.1622	0.1211	0.0146
F-P(J-E)	11	12	13	14	15	16	17	18	19	20
	0.2597	0.2457	0.2339	0.2126	0.2107	0.2082	0.2040	0.1608	0.1206	0.0140
F-P(E-E)	11	12	13	14	15	16	17	18	19	20
	0.2580	0.2444	0.2329	0.2111	0.2085	0.2064	0.2034	0.1599	0.1203	0.0139
F-P(E-J)	11	12	13	14	15	16	17	18	19	20
	0.2642	0.2513	0.2389	0.2173	0.2156	0.2120	0.2086	0.1637	0.1235	0.0145

Query Field は、その run の検索に使用された検索課題中のフィールドである。