情 報 学 基 礎 53－5
（1999．3．1）

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

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

抄録

日本語情報検索システム評価用テストコレクション(NTCIR)プロジェクトと現在構築中のテストコレクション1(NTCIR-1)、そのデータを用いたコンペティション型ワークショップの概要を報告する。テストコレクションに関する議論を踏まえて、NTCIR の基本方針を示した。NTICR-1 の検索対象文書は、約 33 万件で、半数以上は日英の対訳である。検索課題は、判定基準、検索の目的、背景知識などの詳細な検索要求説明を含む。ワークショップは、1998 年 11 月から 1999 年 9 月まで開催され、国内外の 31 チームが参加している。予備テストの結果で訓練用検索課題の正解判定の網羅性を評価したところ、対話型検索によるプーリングは特定の検索課題では特に有効で17.5%のユニークな正解文書を発見した。内部プーリングは全正解文書の 97%をカバーし、概ね良好であった。

NTCIR-1 (NACSIS Test Collection for Information Retrieval systems-1) : Its Policy and Practice

Noriko Kando　Kazuko Kuriyama　Toshihiko Nozue　Keizo Oyama
R & D Dept., National Center for Science Information Systems (NACSIS)

Abstract

This paper reports the outline of the NTCIR (NACSIS Test Collection for Information Retrieval systems) project, its Test Collection 1 (NTCIR-1), and the competition-typed workshop that uses the NTCIR-1. Based on the previous discussion about the data used for information retrieval laboratory testing, we discuss the fundamental policies of the project. NTCIR-1 contains ca.330,000 documents, more than half are Japanese-English paired. A search topic contains detailed narrative including term definition, relevance judgment criteria, the purpose of the search, and background knowledge, which are thought to be helpful for relevance assessment. The Workshop, which started from November, 1998, obtained thirty-one participating groups. The exhaustivity of the relevance assessment for the training topics is reported. The pooling by interactive searches was effective for particular topics and found 17.5% of unique relevant documents. The internal pooling worked well and covered 97% of the whole relevant documents.

1.はじめに

われわれは、日本語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)プロジェクトにおいてテストコレクションの構築と研究目的使用での公開を計画している。その過程で「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型ワークショップ(1998年11月～99年9月)を開催している[1]。

本稿では、NTCIR におけるテストコレクション構築の方針とテストコレクション1 (NTCIR-1)の概要を紹介する。また、1998 年 12 月に予備テストを行ない、正解文書リストの網羅性について評価したので、その結果も合せて報告する。

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る実験用セットで、(1)文書データベース、(2)利用者の検索要求を記述した検索課題群、(3)各検索課題に適合する正解文

書の網羅的リストからなる。

別稿[2]にまとめたように、1960 年代半ばの Cranfield 2 実験以降、各種の検索実験でテストコレクションが構築され、多くの研究者が共通に利用できる「標準テストコレクション」として情報検索の理論的実験的研究に活用されてきた。しかし、その規模は文書数が数千件程度であり、実験結果を大規模なデータベースを扱う実用システムに適応できるかは明らかではなかった[4-5]。

それに対し、1992 年から毎年、開催される情報検索会議 TREC (Text REtrieval Conference) [3]は、多くの研究者が同一課題を遂行するコンペティションによって大規模なテストコレクションを構築している。正解文書は、各参加チームの検索結果の上位一定数を集めて候補とし、それに対して正解判定を行なう。これはプーリング(Pooling)という大規模データベースでの正解文書候補の効果的・効率的な収集法[4]である。この TREC により現実の検索システムに匹敵する大規模なテストコレクションを用いた検索実験が可能になり、情報検索研究から実用システムへの技術移転が進むとともに、各国語での標準テストコレクション構築の動きも加速した[2,6]。

日本でも、近年、標準的テストコレクションの必要性が強く認識され、情報処理学会のワーキンググループにより、新聞記事を使った BMIR[7-8]が構築され、公開されている。これは、すでに 50 の研究グループが利用し、日本語情報検索の検索実験を容易にし、検索実験のスタイルを明確にし、相互比較を可能にすることによって、研究の促進、向上に貢献している。対象文書の種類、数量ともに一層の拡充が求められている。そこで、新たなテストコレクション構築をするため NTCIR プロジェクトを開始した。

以下、次節では NTCIR-1 とワークショップの概要を示す。3節では、予備テストの結果と考察を示し、4節はまとめである。

2. NTCIR プロジェクトと NTCIR-1

2.1 NTCIR プロジェクトの目的

プロジェクトの目的は、以下のとおりである。

- 大規模な日本語テストコレクションを構築することによって、日本語情報検索、日本語を含む言語横断検索研究を促進し、基礎的な研究成果を蓄積する
- テストコレクションの効率的構築法の検討
- 情報検索における日本語固有の問題の検討
- 対話型システムの評価方式の検討

情報検索研究では、提案手法は既存の方法より優れていることを示す必要がある。特に、TREC 開始以降は、提案手法の有用性を、現実のシステムにも外延できる環境で評価することが求めており、日本語検索に使用できる、実用システムに匹敵する大規模テストコレクションの構築は急務である。

また、ネットワーク環境下では、検索システムへの問合せとは異なる言語で書かれた文書も一括して検索できる「言語横断検索(Cross-lingual IR)」への要請が強い。Web や日本の学術文書では日本語と英語などの外国語の文書が混在し、しかも日本語文書中だけでも、同一概念が、日本語、外国語の原綴、外来語のカタカナ表記、略称など多様に表記される。このような表記の差違を超えて有効な検索を行なうためにも、言語横断検索技術が必要である[10]ことから、特に、言語横断検索研究へ利用可能性を考慮した。

空白など明示的な語の区切りがないなど、日本語固有の問題、対話型検索システムへ適応可能な評価方法の検討も目的の一つとした。

2.2 背景議論

テストコレクションを用いた実験室型検索実験の適切性、妥当性については、その成果を現実の検索システムへ外延できるかが重要である。これは、主として、(1)テストコレクション(文書数と検索課題数)の規模と(2)正解判定の妥当性の問題として捉えられる[4, 9, 11-13]。

(1)の規模については、TREC のようなコンペティション型会議を通じて大規模なプーリングを行なうことにより、現実の検索システムに匹敵する規模の、あるいは現実の検索システムへの外延が統計的に可能な規模の文書数を持ったテストコレクションの構築が可能となった。また、現実の検索システムは多様な文書を扱うため、多様な文書あるいは複数のテストコレクションを用いた評価が必要である[9,11]。検索課題については、統計的には、複数のシステム間の比較には、少なくとも 30 以上の事象(検索課題)が必要である。現実の検索システムは多様な検索要求を多数処理しており、多数の多様な検索課題が望ましい[11]。

NTCIR-1 では 30 万件以上の多分野の文書を含み、検索課題は 300 件を最終目標とした。

(2)の正解判定の妥当性については、いままで、以下のような問題が提起されてきた。すなわち、

検索課題が自然か、妥当か
正解文書候補の収集が網羅的か、公平か
正解判定の基準、過程が不明
正解判定は2値より多いのが自然、 など。

検索課題と正解判定は、対になったものである。情報要求はその人の置かれている状況や携わっているタスクに関わり[14]、特定の文書が検索要求

に適合するか判断するレレバンス判定は、人によって、また同一人でも状況によって異なり、システムとの対話の中でしだいに推移していく。また、判定者の、対象文献に対する認識、検索目的、知識に依存する[15]。テストコレクション構築のための正解判定でもこのような背景状況を想定する必要がある。

そこで、NTCIR-1 では、検索課題には、詳細な検索要求説明を付し、判定基準、用語の説明だけでなく、検索の目的、背景知識、状況なども可能な限り記述することとした。

このような詳細な検索課題の記述は、正解判定の一貫性の向上に貢献し、また、対話型検索システムの評価で利用者の問題状況の擬似的プロファイルとしても利用可能[16]である。また、SGML 型タグによって構造化し、短い問合せ、長い問合せなど多様な場を想定した利用が可能である。

「自然な」、「現実の」については、その種類の文書を日常、検索したり、利用している人が作成することを原則とした。

2.3 テストコレクション 1 (NTCIR-1)の概要

上記の議論を踏まえ、テストコレクション 1 (NTCIR-1) を構築した。これは、文書、検索課題、正解文書リスト、タグ付きコーパスからなる。

2.3.1 検索対象文書(Document) (図1参照):

NTCIR-1 は、「学会発表データベース」データの一部を使用する。これは、日本国内 65 学会の協力を得て、その全国大会、研究会などの発表論文要旨を集めたものである。約 33 万件を選定し、一般に検索に用いる「標題」、「著者」、「学会」、「発表年月日」、「発表要旨」、「著者キーワード」を抽出した。「発表要旨」と「キーワード」は各著者がこのデータベース用に作成したデータで、キーワードは自由語である。対応するシソーラスなどの統制語彙はない。文書の半数以上は日英対訳である。JE コレクション(日英文書全体)、E コレクション (英語の標題と抄録を持つレコードの英語部分) 、J コレクション (日本語の標題と抄録があるレコードの日本語部分) がある。

「ヨミ」、「単語の区切り」など学会から提出された元データにはなく、学術情報センターで独自に付加した情報は NTCIR-1 からは除外した。

2.3.2 検索課題(Search topics)(図 2 参照):

検索課題は、利用者の検索要求を一定の書式で記述したものである(図2参照)。検索課題は、分野の研究者(大学院生以上)から、インタビューあるいは図3のフォームによって収集した。訓練用検索課題には大学図書館のレファレンス事例もある。

```
<REC>
<ACCN>gakkai-0000011144</ACCN>
<TITL TYPE="kanji">電子原稿・電子出版・電子図書館-「SGML 実験誌」の作成実験を通して</TITL>
<TITE TYPE="alpha">Electronic manuscripts, electronic publishing and electronic library </TITE>
<AUPK TYPE="kanji">根岸 正光</AUPK>
<AUPE TYPE="alpha">Negishi, Masamitsu</AUPE>
<CONF TYPE="kanji">研究発表会(情報学基礎)</CONF>
<CNFE TYPE="alpha">The Special Interest Group Notes of IPSJ</CNFE>
<CNFD>1991. 11. 19</CNFD>
<ABST TYPE="kanji"><ABST.P>電子出版というキーワードを中心に、文献の執筆、編集、印刷、流通の過程の電子化について、その現状を整理して今後の動向を検討する。とくに、電子出版に関する国際規格である SGML(Standard Generalized Markup Language)に対するわが国での動きに注目し、学術情報センターにおける「SGML 実験誌」およびその全文 CD-ROM 版の作成実験を通じて得られた知見を報告する。また電子図書館について、その諸形態を展望する。出版文化に依拠するこの種の社会システムの場合、技術的な問題というのは、その技術の社会的な受容・浸透の問題であり、この観点から標準化の重要性を論じる。</ABST.P></ABST>
<ABSE TYPE="alpha"><ABSE.P>Current situation on electronic processing in preparation, editing, printing and distribution of documents is summarized and its future trend is discussed, with focus on the concept: "Electronic publishing. "Movements in the country concerning an international standard on electronic publishing, SGML (Standard Generalized Markup Language), are assumed to be important, and the results from an experiment at NACSIS to publish "SGML Experimental Journal" and to make its full-text CD-ROM version are reported. Various forms of "Electronic library" are also investigated. The author puts emphasis on standardization, as technological problems for those social systems based on cultural settings of publication of the country, are the problems of acceptance and penetration of the technology in the society.</ABSE.P></ABSE>
<KYWD TYPE="kanji">電子出版 // 電子図書館 // 電子原稿 // SGML // 学術情報センター // 全文データベース</KYWD>
<KYWE TYPE="alpha">Electronic publishing // Electronic library // Electronic manuscripts // SGML // NACSIS // Full text databases</KYWE>
<SOCN TYPE="kanji">情報処理学会</SOCN>
<SOCE TYPE="alpha">Information Processing Society of Japan</SOCE>
</REC>
```

図 1:NTCIR-1 の文書レコードの例

形式は、初期 TREC に準じ、<タイトル>、<検索要求>、<検索要求説明>、<概念>、<分野>の項目からなる。これは図書館員やサーチャへの文献調査依頼用の書式にプレサーチインタビューの内容などを組み込んだものであるとも考えられる。

<タイトル>は、検索要求のニックネームであり、検索要求を構成する概念が必ずしもすべて含まれるわけではない。しかし、インタネットサーチエンジンなどにしばしば投入される「非常に短い問合せ」を想定した検索実験にも利用できる。

<検索要求>は、利用者の検索要求を記述した

検索課題 q=0005〉
〈タイトル〉
特徴次元リダクション
〈/タイトル〉
〈検索要求〉
クラスタリングにおける特徴次元リダクション
〈/検索要求〉
〈検索要求説明〉
オブジェクトのクラスタリングを行なうとき、オブジェクトを特徴ベクトルで表現することが望まれる。アプリケーションによっては、オブジェクトの次元は数千、数万となることがある。このような場合、事前に次元を落とすことが必要になる。正解文書は、特徴次元リダクションの方法について、理論面から、または実験によって、提案、比較などを行なっているもの。画像処理などの実験の操作の一部として特徴次元リダクションを用いているだけでは要求を満たさない。
〈/検索要求説明〉
〈概念〉
特徴選択, 主成分分析, 情報の粒度, 幾何クラスタリング
〈/概念〉
〈分野〉
1.電子・情報・制御
〈/分野〉
〈/検索課題〉

図 2NTCIR-1 の検索課題の例

自然言語の文または句で、自動システムに投入する「問合せ(検索式)」となる。特定的な長いものから内容語が１～２語の短いものまで多様である。現実の問合せは多様だからである。記述の原則は、分野の専門知識がある人が、検索要求を表現するのに必要な概念をすべて含むこととした。

<検索要求説明>は、検索要求の背景説明、検索の目的、正解判定基準、用語の定義など、背景情報を提供する。対話型システムの評価において、擬似的な利用者プロファイルとしても用いることができる[16]。分野の専門家でなくてもある程度理解できる記述とした。

<分野>は、検索課題を分類する大まかな目安であって、正解判定には利用しない。

個数は、最終的には、各分野取り揃え合計 300 個を目標とした。ワークショップでは、語の重み付けにあらかじめ正解が分かっている訓練用セットが必要なシステムもあることから、訓練用課題 (30 個) と、評価用 (53 個) を設けた。

検索課題は、予備検索を行ない、正解文書が 5 件以上あるものを選択した。分析者グループで内容、書式を検討し、必要な場合は意図を明確にするため分析者が書き直した。用語用字は分野の慣習を配慮した。分析者は専門分野をもつ大学院生以上であり、検索課題の提供者でもある。

検索実験では、検索課題のどの部分を使用することもできる。ワークショップでは、1 チームが複数の検索結果を提出できるが、結果提出時には検索課題中で使用した項目を報告する。システム間比較を容易にするため、検索式を自動構築するシステムの場合は、〈検索要求〉のみを使用した検索を必須とした。

テストコレクション用検索要求についてのお願い

情報検索システムの評価用テストコレクションを構築するために、「検索要求」を収集しています。「情報要求」とは、「○○について論じている論文が ほしい」「XXについて知りたい」というように、検索したい文献の内容を具 体的に述べたものです。 ご協力いただける方は、以下のフォームを埋めて送信してください。記入にあたっては 記入例 を参考になさってください。よろしくお願いいたします。

なお、ご記入いただいた検索要求については、日本国内の学会発表要旨を集めた「学会発表データベース」のデータの一部を使用して実際に検索し、 その結果(タイトル・抄録など)は、後日お知らせいたします （結果をお知らせするまで時間がかかることがあります)。 不要の方は、その旨ご記入ください。

テストコレクション用検索要求記入フォーム

※(1),(8)は必須です。他の項目は必須ではありませんが、できるだけご記入ください。

1. 検索要求を自然言語の文でご記入ください。

2. 検索要求に関する補足説明(背景や用語の解説など)をご記入ください。(分野に詳しくない者が検索する際の理解を助けるためのものです。)

3. どのような文献が必要ですか。文献の内容、アプローチ、手法など、 検索要求を満たしていると判断するための条件・要素をご記入ください。

4. また、最低限どのような条件・要素を満たせば、完全とはいえないまでも(部分的には)検索要求を満たしているといえますか。

5. どんな目的で文献を探していますか。(例:これから新しい研究を始めるのでそのトピックの現状を知りたい、D論の準備、研究テーマを決めるた め、同じ研究手法を使っている論文をみたい、最新の研究動を知りたい、 新しい手法・モデルを知りたい、など) 6. 検索要求中の主要な概念を表す用語を、同義語や類義語 を含めてご記入ください。

7. 検索要求の分野を選んで下さい。(複数回答可)

1. 電子・情報・制御	2. 化学
3. 建築・土木・造園	4. 生物学・農学
5. 理学	6. 工学
7. 医学・歯学	8. 人文・社会

8. お名前・ご所属等をお知らせください(こちらからの連絡にのみ使います。外部に公表することはありません)。

＊ 氏名: ＊ 所属: ＊ E-mail-address:

9. その他、ご質問等がありましたら、ご自由にお書きください。

以上です。ご協力ありがとうございました。

※ご記入いただいた内容や検索結果等について、後日、問い合わせをさせてい ただくことがありますので、ご協力いただけると幸いです。

※当プロジェクトで開発しているテストコレクション(データベース、検索要求群、 各検索要求のレレバント文献リストのセット)は、広く、情報検索の研究者が使えるようにしたいと考えています。 そのため、 ご記入いただいた内容は、情報検索システム評価用テストコレクション の一部として情報検索研究者に研究目的での使用に限って公開する場合が あります(情報要求作成者のお名前・所属・連絡先などは公開しません)ので、 ご了承ください。外部への公開を不可とする場合は、その旨ご記入ください。

図 3. 検索要求収集用フォーム

2.3.3 正解判定(正解文書の網羅的リスト):

正解文書候補の収集は、プーリング法とした。訓練用検索課題については、プロジェクト内部で3種類の異なる検索システムを用いてパラメタや索引手法を変えた合計約 30 種類の検索結果から上位文献を収集して候補とした。これを補うために、対話型のシステムで網羅性の高い検索を行なって候補に追加した。これに対して分析者が正解判定を行ない、正解リスト ver. 1 を作成した。さらにワークショップで予備テストとして、参加者から訓練用検索課題に対する検索結果を任意で提出を求め、そこで新たに見つかった正解文書を追加して、正解リストを補完し ver. 2 とした。

評価用検索課題については、ワークショップ参加者から提出される検索結果の上位を集めて正解候補とする。

正解判定は分析者が行なった。正解判定は、検索要求に「適合(A)」、「部分的適合(B)」、「不適合(C)」の3段階である。A,B,C 以外は、正解文書候補以外の文献で、分析者が直接判定していないが、さまざまな方法で検索しても検索されなかったことから、「不適合」であると想定している。

専門分野を考慮して主分析者を決め、2 名のクロスチェックに基づき主分析者が最終判定を行なった。自作の検索課題については作成者が主分析者となる。分析者以外から収集した検索課題では、判断に迷う場合は検索課題提供者などから参考意見や背景情報を得た。Ver.2 の正解文書数と判定数を表 1 に示した。

BMIR の書式に基づき、適合文書と判定に迷った不適合文書には判定の根拠を短くコメントとして付した。

表1.訓練用検索課題
正解文書数(Ver.2)

Topic#	正解文書数 A	B	計	C	判定済文書数計
0001	283	10	293	2617	2910
0002	6	13	19	7153	7172
0003	5	9	14	4827	4841
0004	38	0	38	7862	7900
0005	11	2	13	3844	3857
0006	69	3	72	3355	3427
0007	3	13	16	4482	4498
0008	8	17	25	2509	2534
0009	6	2	8	3769	3777
0010	52	3	55	3747	3802
0011	7	0	7	2547	2554
0012	46	24	70	4919	4989
0013	38	0	38	4527	4565
0014	317	0	317	3607	3924
0015	18	2	20	3663	3683
0016	5	0	5	5055	5060
0017	10	6	16	4387	4403
0018	148	19	167	1874	2041
0019	89	3	92	4604	4696
0020	16	0	16	5287	5303
0021	8	3	11	6275	6286
0022	78	4	82	3600	3682
0023	98	0	98	6513	6611
0024	158	0	158	5159	5317
0025	19	4	23	4226	4249
0026	19	4	23	4226	4249
0027	14	9	23	3442	3465
0028	1458	132	1590	6175	7765
0029	138	42	180	4619	4799
0030	15	8	23	4532	4555
sum					136914

A:検索要求に適合、B:部分的適合、C:不適

2.3.4 タグ付きコーパス:

NTCIR-1 の文書データの一部に、語構成要素まで考慮した詳細な形態素タグを付与した[17]。 日本語では語間に空白などの自明の区切り記号がないという自動索引・検索式解析における日本語固有の問題を検討する基礎的データを意図している。

2.4 ワークショップ(コンペティション)[18]

2.4.1 目的と意義:

上述の NTCIR-1 の予備版を用いて、1998 年 11 月からコンペティション型ワークショップを開催している[18]。ワークショップの目的は、(1)参加者から正解文書候補を収集して大規模テストコレクションを構築、(2)各種技法の効果を共通の基盤で比較、(3)意見交換の場を提供することによって日本語情報検索研究を促進することである。

大規模テストコレクション構築では、複数の検索システムによる正解文書候補の収集(プーリング)が必要であるが、コンペティション型ワークショップはその最も有効な手段の一つであり、参加者にとっても、同じ基盤でシステム間の比較、検討、議論に参加できるという利点がある。データフュージョン研究用のデータの蓄積もできる。

このワークショップの主な関心事は、数値によるシステムの順位付けではなく、どのような技術を用いるとどのような効果があるかというシステムやアプローチの相互比較と特徴づけである。新しいアイディアを試し、議論し、様々な試みの中から、今後の有望な方向性を参加者全員の共同作業の中で見出していこうとするものである。そのため、システムの特徴を記述する詳細な「システム説明フォーム」[19]提出、成果報告会での論文形式の報告なども行なう。

2.4.2 タスクと日程:

ワークショップのタスクは以下の3種である。参加者は1つ以上のタスクを遂行する。

随時検索タスク(ad hoc IR task): 特定のデータベースに対して、新しく配布される評価用検索課題の検索を行ない、その検索性能を調べる。

言語横断検索タスク (cross-lingual IR task):日本語の検索課題を用いて、英語の文献を検索する。

用語抽出・役割分析タスク: 標題及び抄録から用語を抽出し、抄録の主要論述における「対象」「手法」「主操作」を表わす用語を識別する。

主な日程は下記のとおりである。

1998年11月初旬：文書データと訓練用課題とその正解文書リストを配布
12月2日：予備テストの結果提出
訓練用課題に対する検索結果を提出
1999年2月8日：評価用検索課題の配布
3月1日：検索結果提出
8月 30-31日：成果報告会、会議録発行
9月末 ：NTCIR-1本格版配布開始

参加者数は、随時検索 23、言語横断検索 16、用語抽出10チームの合計31チームである。国別では国内19、海外8、混成4である。

3.予備テストにおけるプーリングとシステム評価への影響の分析

3.1 予備テストの概要

NTCIRワークショップでは、12月2日に予備テストを行なった。目的は次のとおりである。

(1)内部で正解文書候補のプーリングした訓練用検索課題の正解リスト ver.1 の網羅性を評価し、補完をする

(2)事務局、参加者ともに、検索結果提出の手順を確認し、作業量を見積もる

これは必須ではなく、結果非公開とし自由参加とした。10チームから23の検索結果が提出された。この検索結果について追加の正解判定作業を行ない、新しく見つかった正解文書は、訓練用検索課題の正解文書リストに追加し、Ver.2 としてワークショップ参加者に公開した。

以下に、この訓練用検索課題30件について、プーリング方法の効果とシステム評価への影響を分析したので、その結果を示す。

3.2 訓練用検索課題の正解文書候補収集法

訓練用検索課題の正解文書は、正解文書候補の収集法によって、下記のサブセットにわけられる。

Auto (A): プロジェクト内部での自動検索式構築によるプーリング。検索エンジンは3種。約30回の検索結果の上位をプール

Interactive (I) :図書館情報学専攻の大学院生が対話型検索システムを用いて網羅性の高い検索を実行

Pretest (P) : ワークショップの予備テストとして10チームが提出した23組の検索結果の上位1000件。正解判定を行なったのは上位100件ずつ。表中の数字は、上位1000件に対するもの。

上記のAとIの和が予備テスト前の内部で用意した正解文書リストVer.1である。

3.3 正解文書数の分析 (表2参照)

ワークショップの随時検索用の訓練用検索課題30個に関する正解文書リスト Ver.2 のサブセット別の正解文書数示した (表2参照) 。「適合」と「部分的適合」を合せて、「正解」とした。

A, I, P の相互の重なりは大きい。予備テストによって、新たに追加された正解文書 (表中、P-

表2.訓練用検索課題(q0001～q0030)
プーリング範囲別にみた正解文書数(随時検索タスク)

	Ver.2 (After Pretest)					#of unique relevant documents		
	Ver.1(inside pooling)							
Topic#	Auto	Interac	total	Pretes	total	A-on	I-only	P-onl
0001	235	289	290	283	293	1	9	3
0002	18	19	19	19	19	0	0	0
0003	14	11	14	14	14	0	0	0
0004	33	37	37	37	38	0	1	1
0005	12	2	12	13	13	0	0	1
0006	54	61	68	71	72	0	1	4
0007	12	0	12	15	16	1	1	4
0008	25	0	25	25	25	0	0	0
0009	6	8	8	8	8	0	0	0
0010	23	51	53	53	55	0	1	2
0011	7	3	7	7	7	0	0	0
0012	58	68	70	67	70	0	2	0
0013	23	31	37	37	38	0	0	1
0014	225	283	287	310	317	1	7	30
0015	20	20	20	20	20	0	0	0
0016	5	5	5	5	5	0	0	0
0017	15	13	15	15	16	1	1	1
0018	140	0	140	155	167	12	8	27
0019	89	90	91	92	92	0	0	1
0020	16	16	16	16	16	0	0	0
0021	11	11	11	11	11	0	0	0
0022	70	68	81	82	82	0	0	1
0023	79	98	98	98	98	0	0	0
0024	145	156	157	157	158	0	1	1
0025	23	23	23	23	23	0	0	0
0026	21	23	23	23	23	0	0	0
0027	20	23	23	23	23	0	0	0
0028	257	1586	1586	1018	1590	0	566	4
0029	128	113	159	162	180	3	17	21
0030	15	23	23	22	23	0	1	0
total	1799	3131	3410	2881	3512	19	616	102
%	51.2%	89.2%	97.1%	82.0%	100.0%	0.5%	17.5%	2.9%
total*	1542	1545	1824	1863	1922	19	50	98
%	80.2%	80.4%	94.9%	96.9%	100.0%	1.0%	2.6%	5.1%

total*:total number of relevant documents except q0028

only）は全体で 2.9%、対話型の貢献が非常に大きかった q0028 を除くと 5.1%であった。

内部のプーリングは、使用システム数が３種類のみであったが、概ね良好であった。対話型の検索は、0028、0001 など特定の検索課題で特に効果を示し、17.5%のユニークな正解文書を見つけた。予備テストの検索結果で、さらに網羅性を高めることができた。

3.4 システムの性能評価に対する影響(表３参照)

正解文書の網羅性が検索システムの検索性能評価に及ぼす影響を見るため、正解文書 Ver. 2 のサブセット毎の評価結果を比べた。予備テストとして提出された随時検索タスクの検索結果８チーム 16 件のうち、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件を選び、平均精度(補完なし)の値で順位づけた結果を表３に示した。

A, B, C, ... H はそれぞれ検索結果を表わす。

表 3 に示すように、概ね、どのサブセットを用いて評価を行なっても相対的な順位づけにはほとんど影響がなかった。また、対話型で収集された正解文書のサブセットが特に対話型のシステムに有利であるというような検索法による偏りもこの結果からは、見られなかった。

今回は文書配布後 1 ヶ月足らずで実施した予備テストの限られた検索結果の分析であり、参加システム数も少なかった。本テストでも、プーリングの効果など継続して検討する予定である。

4.まとめ

大規模なテストコレクションの構築では、プーリングによる正解文書候補の選択が前提となる。しかしながら、「異なるモデルに基づく検索システムは異なる正解文書を検索する」は、より正確には、「異なるモデルに基づく検索システムは、概ね似たような正解文書を検索するが、まったく異なる不正解文書を検索する」[20]。すなわち、プーリング参加システムが増えれば増えるほど、正解文書候補の収集は網羅性が高まるが、同時に判定対象としなければならない不正解文書数は飛躍的に増加する。各システムから正解文書候補のプールに入れる文書数の設定をどのようにするかが重要なポイントとなる。

それに対し、検索性能の高いシステムの検索結果を優先することによって正解文書候補数の増加を押さえ、効率よく正解文書リストを作成する Move-to-Front プーリング法がある[21]。TREC では、この Move-to-Front 法は、研究としては妥当な手法でも、コンペティションとしては公平でないという誤解を生む可能性があるので採用は見合わせ、プールする検索結果数の上限、検索結果から採用する文献数は、各参加者で同一としている。

また、上位一定数のプールと分析者間の判定の不一致度に関する推測に基づいて、大規模テストコレクションの正解文書リストの網羅性とそれによるシステム評価の妥当性に関する疑義も提起されている。[22]。それに対し、分析者間の正解判定の一貫性、一致度は低くても、どの分析者による判定結果を用いてシステムの評価を行なっても、評価結果に有意な差は見られなかったという報告もある[23]。その原因については十分な分析がなされていないが、考えられる要因としては、従来からいわれているように、多数の多様な検索課題を用いることによって、個々の判定の不一致や網羅性などさまざまな要因のシステム性能評価への影響を相殺しているということである。

NTCIR においても、引き続き、正解文書候補の収集法と判定法などについて分析を進め、評価として妥当、公平であり、網羅性のあるテストコレクション構築法を考えていく必要がある。

表3 プーリング範囲のシステム評価への影響

run-ID			A	B	C	D	E	F	G	H
topic used for query			short	short	short	long	all	all	short	long
methods		method	auto	auto	auto	auto	inter	auto	inter	auto
ver.2	ver.1	Auto	1	2	3	5	4	6	8	7
		Interac	1	3	2	5	4	6	8	7
		total	1	2	3	5	4	6	8	7
	Pretest		1	2	3	5	4	6	8	7
	total *		1	2	2	4	4	6	7	7

short: description(検索要求)のみ all: 全項目
long: narrative(検索要求説明)を含むもの: t,d,n,c(D)、d,n(E,F)
total: 第2位と3位、4位と5位、7位と8位は、平均精度の差が0.01未満

このNTCIRプロジェクトについては、TREC主催者・参加者、BMIR-WGメンバなど多くの情報検索研究者からアドバイスを得ている。今回のテストコレクション構築の経験も共有化をはかり、今後の評価手法に関する議論に貢献したい。

検索課題の一部は、韓国で現在開発中の韓国語テストコレクションと交換し、最終的に、韓英日3カ国語の言語横断検索が可能になる予定である。言語横断検索については、「自然な」検索課題作成と正解判定には、各言語の母国語話者が必要であり、国際協力が必要である。

また、テストコレクション自体の評価については、ワークショップのほか、TRECコレクションとNTCIR-1の正解文書空間の比較についての共同研究を行なう予定である。

テストコレクションを用いた検索実験は、情報検索システムの機能の一部を評価しているにすぎない。情報検索システムは、利用者や利用者をとりまく行動、ユーザインタフェース、費用対効果など多様な側面から評価が可能である。今後は、テストコレクションの対象文書種類の拡大とともに、対話型システムの評価なども対象とし、情報検索システムのより妥当な評価に関する議論の端緒になれば幸いである。

謝辞: 本研究は、日本学術振興会「未来開拓研究」(課題番号 JSPS-RFTF96P00602)による。

文献

1. Kando, N. et al. (1998) "NTCIR : NACSIS Test Collection Project" [Poster] IRSG98, March , Autrans,France (http://www.rd.nacsis.ac.jp/~ntadm/ も参照)
2. 神門典子.(1999)「情報検索システムの評価を巡って：テストコレクションとコンペティションを中心に」1999年情報学シンポジウム講演論文集, p.129-136.
3. http://trec.nist.gov/
4. Sparck Jones, K; van Rijsbergen, CJ. (1975) Report on the need for and provision of an 'ideal' information retrieval test collection' BLRD Report 5266
5. Blair, DB, et al. (1985) "An evaluation of retireval effectiveness for a full-test document retrieval system" Comm of the ACM, 28(3): 289-299
6. Smeaton,A.et al. (1997) "The TREC experiments and their impact on Europe", J of Inf Sci, 23(2):169-174
7. 木本晴夫ほか. (1998) 「日本語情報検索システム評価用テストコレクションの構築」1998年情報学シンポジウム講演論文集, p.103-120
8. 木谷強ほか(1998)「日本語情報検索システム評価用テストコレクションBMIR-J2」情処研報, 98(2):15-22
9. Salton, G. (1992) "The state of retrieval system evaluation" Inf Proc & Manag, 28(4):441-450
10. Kando, N.(1997) "Cross linguistic scholarly information transfer and database services in Japan." ASIS '97
11. Sparck Jones, K; van Rijsbergen, CJ. (1976) "Information retrieval test collections" Journal of Documentation, 32(1):59-72
12. Harman,D. "Panel: Building and using test collections", Proceedings of ACM-SIGIR'96, p.3335-337
13. Sparck Jones , K. (1980) Information Retrieval Experiment. Butterworths,
14. Schamber, L. et al.(1990) "A re-examination of relevance: toward a dynamic situational definition" Inf Proc & Manag, 26 (3):321-43
15. Oddy, RN.(1980) "9. Laboratory tests: automatic systems" In. Information Retrieval Experiment. Ed by K. Sparck Jones, Btterworths, p.156-178
16. Borlund, P et al..(1997) "The development of a method for the evaluation of interactive information retireval systems" J of Doc, 53(3):225-250
17. Kageura, K. et al. (1997) "NACSIS Corpus Project for IR and Terminological research", NLPRS'97, p.493-6
18. http://www.rd.nacsis.ac.jp/~ntcadm/workshop/
19. http://www.rd.nacsis.ac.jp/~ntcadm/workshop/sysdesc-ad.html, ../sysdesccl.html
20. Lee, JH et al., (1997), "Analysis of multiple evidence combination." ACM-SIGIR 97, Philadelphia,p.282-89.
21. Cormack, GV, et al. "Efficient construction of large test collections". Proc. of ACM-SIGIR 98, Melbourne, p.282-289.
22. Zobel, J. (1998) "How reliable are the resutls of large-scale information retrieval experiments?" Proc.of ACM-SIGIR 98, Melbourne, p.307-313.
23. Voorhees, EM. (1998) "Variations in relevance judgments and the measurement of retrieval effectiveness". Proceedings of ACM-SIGIR 98, Melbourne, p.315-323.