タスクを考慮した情報検索テストコレクション構築に関する考察

吉岡 真治[1,a)] 神門 典子[,b)]

概要：情報検索システムのためのテストコレクションは、検索課題のセットと各々の課題に対する文書データベース中の適合文書のリストから構成される。このようなテストコレクションでは、ユーザの様々な要求に対応する複数の検索課題を用いることにより、システムの平均的な性能を評価することが可能となる。しかし、質問応答のための情報検索に代表されるような、特定の目的に特化した情報検索システムを評価する場合には、各々の課題が、特定の目的を評価するのにふさわしい課題であるかどうかを判断する必要があると考える。本研究では、課題の適切さを議論するために重要となる指標について議論する。また、質問応答のための情報検索では、質問に対する答が適合文書には必須であるということに注目したテストコレクションの構築方法についても議論を行う。

1. はじめに

異なる情報検索システム間の性能を比較するためには、同一の検索条件のもとで検索を行い、その結果を比較することが重要である。この様な比較を実現するために、文書データベース、そのデータベースに対する検索課題、ならびに、検索課題に対する適合文書集合のリストから構成される情報検索システムのためのテストコレクション [1] が利用されている。

この様なテストコレクションの構築には、多くの手間がかかるために、TREC や NTCIR に代表されるようなワークショップを通じた構築が行われている。この検索課題を作成するにあたっては、情報検索システムの一般的な性能を分析するために、実際に、想定される情報検索システムに問い合わせられるような質問を複数収集し、その平均的な性能を分析することで、広範なタイプの質問に対応可能な情報検索システムの評価を行ってきた。

一方、近年では、質問応答のための情報検索 (IR4QA) [2], [3] や地理時間情報に対する情報検索 [4], [5] の様に、特定のタスクに特化したような情報検索システムのためのテストコレクションが提案されている。このような、特定のタスクに特化した情報検索システムの多くは、一般的な情報検索システムをベースにして、付加的な機能を追加することにより構築されることが多い。

ところが、実際の検索結果の比較時においては、付加的な機能を利用しないシステムでもコンスタントに良い結果を得られる課題や、付加的な機能が不可欠な課題などが混ざりあって存在するため、トータルとしての情報検索システムとしての評価はできるが、個別の付加的な機能を分析するための評価が困難であるという問題があった。

この様な問題点については、日本語の情報検索テストコレクション構築の初期の取り組みにおいて議論がなされている。BMIR-J2 [6] では、複雑な処理が必要とされる検索課題と単純なキーワードによる検索課題などを分類するという方法が提案され、課題ごとのラベルづけがなされていた。しかし、この時代の情報検索システム自体が十分に成熟していなかったこともあり、結果として、単純なキーワード検索の比重が高くなるという形となっていた。現在の情報検索システムの成熟状況を考えると、BMIR-J2 の時代の検索課題の分類基準について、再考する価値があると考える。

本論文では、BMIR-J2 における検索課題に注目した分類ではなく、検索課題と適合文書の間のミスマッチに注目した検索課題の難しさを用いることで、付加的な機能が必要な課題とそうでない課題を分析する枠組を提案する。また、NTCIR9-GeoTime のテストコレクションを対象にして、その分析の具体例を紹介する。

さらに、Factoid 型の質問応答のためのテストコレクションにおいて、答の情報を利用することが、テストコレクションの網羅性を向上させることについて述べ、地理時間

[1] 北海道大学
Hokkaido University, N14 W9, Kita-ku, Sapporo-shi, Hokkaido, 060-0814, Japan
[a)] yoshioka@ist.hokudai.ac.jp
[b)] kando@nii.ac.jp

情報に対する質問応答のような、特定の質問応答のための情報検索システムを評価するためのテストコレクションを作る基準についても議論を行う。

2. IR4QA のためのテストコレクション

2.1 検索課題の難易度と付加的な機能の分析

IR4QA とは、質問応答システムが利用する情報検索システムであり、質問に対する答を含むと考えられる文書を検索することが目的である。この様な用途に対応する形で、文書の適合判定においては、トピックとの適合度という曖昧な判断基準ではなく、答となる記述 (ナゲット) を含むかどうかが、適合性判定の基準となる [7]。そのために、IR4QA のシステムでは、答のタイプ (人名、長さ、時間、…) の推定結果や、手がかり表現などを使ったリランキングなどの様々な付加的な機能が提案され、テストコレクションによる評価が行われている。

上記のような IR4QA のシステムの持つ特徴のため、IR4QA における検索課題では、検索質問の作り方が、課題の難易度に大きな影響を与える。Gey ら [8] は、地理時間情報に関する質問応答の情報検索システムの評価タスクである NTCIR9-GeoTime タスク [5] において、同一のトピックに関する検索課題を異なる表現で作成し、表現と課題の難易度の関係について分析を行っている。

以下に、実際に用いた同一トピックに関する検索課題のバリエーションを示す。

GeoTime-0035

日本語：500 人以上の死者を出したパイプライン事故は、アフリカのどこで、いつ起きましたか？

英語：When and where did a pipeline explosion occur in Africa killing over 500 people?

GeoTime-0036

日本語：5 人以上の死者を出したアフリカで起きたパイプライン事故について、いつ、どこで、起きましたか？

英語：When and where have there been pipeline explosions in an African country with more than 5 fatalities?

GeoTime-0037

日本語：北緯 5 度 52 分 12 秒東経 5 度 45 分の近くで起きた数百人の死亡者を出した事故は、どのような事故ですか？また、それはいつ起きましたか？

英語： What fatal accident occurred near (geographical coordinates 5°52'12"N 5°45'00"E / 5.870°N 5.750°E / 5.870; 5.750), which killed hundreds of people, and when did it occur?

GeoTime-0036 は、GeoTime-0035 と比較して、死者の人数に対する制約が緩く、複数の事故に関する記事が正解となる検索課題であるのに対し、GeoTime-0035 は、制約が存在するため、ただひとつの事故に関する記事を選ぶ必要がある。また、GeoTime-0037 は GeoTime-0035 と同じ質問を緯度・経度情報で与えたものとなっている。

評価実験を行う前の仮説としては、GeoTime-0035 より GeoTime-0036 の方が、500 人以上という数値に関する処理が必要な分、GeoTime-0036 より難しく、GeoTime-0037 は緯度・経度を取り扱うモジュールが必要なため、より難しくなると予想した。

この仮説に対し、英語では、上記の仮説通りの結果が得られたが、日本語では異なる結果を得ることとなった。

(1) 英語では、死亡者の数に制約の厳しい GeoTime-0035 の方が、より制約の緩い GeoTime-0036 よりも難しい事が確認されたが、日本語では確認できなかった。

(2) 緯度・経度のモジュールを必要とするシステムの結果は、他の結果と大きく異なり、適合文書を発見できないようなシステムも存在した。

2.2 適合文書に注目した検索課題の難易度

Gey らの分析では、検索質問の表現に注目して検索課題の難易度を分析していた。これは BMIR-J2 における課題分類と同じアプローチであると考えられる。しかし、本研究では、検索課題の難易度は、検索課題の表現のみによって決まるのではなく、検索課題と適合文書間の比較によって決まると考えて分析を行う。

この考え方に基づき、適合文書に関する分析をしたところ、英語では、0035 と 0036 の適合文書の数は、各々 21 件と 47 件と大きく異なるのに対し、日本語では、共に 10 件であった。これは、文書データベースの違いによる影響が大きいと考える。英語の場合には、New York Times に加え、韓国の Korea Times、中国の新華社通信や日本の毎日新聞の英語版、といった様々な国の文書データが利用されているのに対し、日本語は、毎日新聞のみとなっている。

この結果、日本では、小さな事故が報じられていないため、結果として、死亡者の数の制約を処理することが問題ではなく、アフリカのパイプラインの事故を見つけることができれば良いという課題になり、細かな標記の違いの影響が現れただけと考えられる。

つまり、Gey らの研究や BMIR-J2 の研究のように、検索課題のみに注目した分類だけでは、今回のような検索課題の特徴を分析するのには不十分であり、検索課題と適合文書間の表現の違いに注目した検索課題の難易度の分析 [9] の手法を用いることが適切であると考える。

3. 特定機能を評価するための検索課題の難易度の分析

3.1 検索質問と適合文書のミスマッチに注目した検索課題の難易度の分析

検索課題の難易度に関する議論は多く行われているが、

本研究では、検索質問と適合文書のミスマッチに注目した難易度に関する指標 [10] を用いる。

本手法では、Boolean 式の形でも表現された検索質問 (ただし、単純に Boolean 式を満たしている文書集合全体が適合文書ではない) について、Boolean 式を満たす文書と適合文書の関係を分析する手法で、次の 2 つの指標により、課題の特徴を分析する。

$$CoverageRel = \frac{|R \cap S|}{|R|}$$
$$FocusAppropriateness = \frac{|R \cap S|}{|S|}$$

ただし、R と S は、各々、適合文書、Boolean 式を満たす文書の集合である。

$CoverageRel$(CR) と $FocusAppropriateness$(FA) は、各々、Boolean 式による検索結果の再現率と精度を表す指標であり、両者が共に 1 に近ければ、検索課題と適合文書間のミスマッチは低く、容易な検索課題と考えることができる。

一方、CR が低い場合には、検索式に利用された単語を含まない関連文書が多いことを示し、同義語などの検索語の追加を扱う必要性があると考えられる。これに対し、FA が低い場合には、Bag of Words による検索では、不十分で、適合文書に絞りこむための機能 (例えば、検索意図の推定や、語の係受け関係などを用いる) が必要であることが考えられる。

ただし、一般的な情報検索システムの処理を考えた場合に、必ずしも、これらの指標が低いから検索課題が難しくなるわけではない。例えば、CR が低い場合でも、初期検索の上位で適合文書を多く見つけることができた場合には、疑似適合文書フィードバックによる検索語拡張などでも、十分に検索性能の改善が期待できる。

3.2 二つの指標に注目した検索課題の分析

NTCIR9-Geotime のテストコレクションに対し、先に示した二つの指標を用いて、具体的に分析を行う。

NTCIR9-Geotime では、検索課題を表すための Boolean 式は与えられていないため、何らかの基準で Boolean 式を作成する必要がある。今回の NTCIR9-Geotime の課題では、何らかの固有名詞について、場所や時間的な側面からの質問をしているものが多かったことから固有名詞を利用した Boolean 式を作成した。また、多くの検索モデルにおいて、単語を含む文書数 (DF) が多い単語よりも、少ない単語に重きをおいて、ランキングが行われることに注目し、必要に応じて、DF の少ない単語を追加することとした。

表 1 に、日本語の検索課題に対する各課題ごとの Boolean 式 (and で結合) と各々の Boolean 式に対する CR と FA を示す。この指標を計算する際には、「国際子ども図書館」といった形態素解析器によっては、複数の形態素に分割するような単語についても、形態素に分割せず、1 単語として扱っている。また、初期の検索語から作成した Boolean 式に対して CR,FA の値が低い検索課題については、適合文書を参考に、より適切と考えられる検索式を作成し、その値を併記した。これらの検索式を比較することにより、どのような機能が必要とされていたかを推測することができると考えている。

これらの検索課題の難易度と、地理時間情報を処理するための特別の機能の必要性に関する議論を行うために、これらの特別の機能を有していない検索システムの結果との関連性を議論する。NTCIR9-Geotime では、オーガナイザがベースラインシステムとして、地理時間情報を処理するための特別の機能を付加していない確率モデルと疑似関連文書フィードバックに基づく情報検索システムによる検索結果を提出している [11]。

本研究では、まず、第一次の近似として、このベースラインシステムの性能が、各参加者が提出した他のシステムの平均的な性能よりも低い場合には、地理時間情報を処理するための特別の機能が重要であり、そうでない課題を重要性の低い課題と考えることにした。

表 2 に、提出された全ての検索結果に対する平均の平均精度 (AP) と nDCG の値を示す。課題番号が太字になっている課題が、ベースラインシステムの指標が、平均よりも 10%以上高い課題であり、下線を引いた課題がベースラインシステムの指標が、平均よりも 10%以上低い課題である。

ベースラインシステムでは、疑似関連文書フィードバックを使っていることから、初期検索における関連文書を多く含むような課題では、性能が向上し、関連文書を含まないような場合には、性能が低下すると考えられる。

具体的には、CR と FA が共に 0 となっている課題番号 32,33,37,44,45 では、ベースラインシステムが全て性能が平均以下となっている。逆に、FA が高い課題 47,48 については、性能の向上が確認されている。ただし、課題 49,50 のように、FA が高い場合についても性能が低下している場合が見受けられた。これらの二つの課題で用いた Boolean 式は、固有名詞一つだけのものであるが、ベースラインシステムは、単語連接の情報を扱わないため、検索語が複数の一般名詞に分解されて、うまく検索できなかった可能性もある。

また、図 1 に、ベースラインシステムの平均精度と他のシステムの平均の平均精度の比の順に課題を並べた場合の CR、AF の値をプロットしたグラフを示す。このグラフからは、課題の難易度と CR、AF の値について、相関を確認することができない。例えば、CR の値が十分高い課題 27 や FA の値が十分高い課題 31 のベースラインシステムの性能が悪い。一方、課題 38 のように、CR,FA が低いに

も関わらず、ベースラインシステムの性能の高い課題も存在している。

表 2 平均精度と nDCG

課題	平均値		ベースライン	
番号	AP	nDCG	AP	nDCG
26	0.4347	0.7734	0.4910	0.8032
27	0.4386	0.6993	0.0008	0.0627
28	0.2389	0.4766	0.1306	0.3893
29	0.3974	0.7057	0.1987	0.4725
30	0.324	0.4875	0.3929	0.5912
31	0.2905	0.5447	0.0000	0.0040
32	0.1418	0.2812	0.0004	0.0145
33	0.1718	0.4212	0.0046	0.1042
34	0.3513	0.5822	0.5549	0.8246
35	0.4708	0.6847	0.0008	0.0550
36	0.4639	0.6828	0.0010	0.0743
37	0.1722	0.2109	0.0000	0.0000
38	0.3031	0.582	0.4462	0.7164
39	0.4692	0.7126	0.6550	0.8334
40	0.6391	0.8164	0.0742	0.4151
41	0.7091	0.8677	0.7715	0.8818
42	0.1851	0.4335	0.2439	0.5548
43	0.2151	0.5178	0.1441	0.4523
44	0.1311	0.2647	0.0056	0.1190
45	0.158	0.3911	0.0000	0.0000
46	0.5835	0.7444	0.4873	0.7443
47	0.7012	0.7252	0.7746	0.8461
48	0.5146	0.7457	0.6931	0.8232
49	0.5868	0.7893	0.0669	0.3916
50	0.458	0.677	0.0500	0.2197

3.3 考察

現時点の分析では、作成した Boolean 式の妥当性も含めて、さらなる検討を進める必要があるが、下記のことが主張できると考えている。

(1) FA の高いような検索課題については、特別の機能を組み込まない情報検索システムでも、疑似適合文書フィードバックなどを用いることにより、容易に適合文書を発見することができる。そのため、これらの検索課題については、特別な機能の性能評価というよりは、その機能の副作用のチェックという側面が強くなると考えられる。

(2) 初期検索において、与えられたキーワードから適合文書を見つけることが困難であるような FA や CR が 0 となっている課題においては、特別の機能を組み込まない情報検索システムは、その性能を著しく低下させる。

必要とされる機能については、必ずしも、地理時間情報に関する機能だけではなく、表 1 に示した Boolean 式にあるように、表記のぶれに対応するといったものも存在する。

システム全体の性能を比較するという観点では、このテストコレクションの役割は存在するが、地理時間情報に関連した機能を評価するという観点からは、利用する検索課題の選択を行うという方法も考えられる。

4. IR4QA のテストコレクション作成時における回答の利用

IR4QA の検索タスクでは、適合文書判定において、回答に関するナゲットを含むことが、文書を適合と判断するための必須条件となっている。このことから、回答に関するナゲットが既知であるならば、その情報を用いることにより、より網羅的な文書の検索が行えると考えられる。また、このような形で作成した検索結果の網羅性が十分高いのであれば、ワークショップ型のテストコレクションの作成に変わる簡易的なテストコレクションの作成方法につながると考えられる。

この考え方に基づき、NTCIR9-GeoTime タスク [12] では、あらかじめ、手作業により作成した回答を、元の検索課題に付け加えた検索式を作成し、より網羅的な適合文書の収集を試みた。具体的には、「世界で最も長いつり橋はどこにありますか？また、いつ開通しましたか？」という検索質問に対し、その答である「明石海峡大橋」「兵庫県」「1998 年 4 月 5 日」を検索質問に追加し、「世界で最も長いつり橋はどこにありますか？また、いつ開通しましたか？明石海峡大橋 兵庫県 1998 年 4 月 5 日」を入力として、検索結果を得る。ただし、実際のタスクでは、回答に関する情報を用いて手作業で検索式の修正を行ったチーム (OKSAT)[13] が存在したために、オーガナイザーによる回答を用いた検索結果と OKSAT による回答を用いた検索結果を合わせたものを「**回答を用いた検索結果**」と呼び、それ以外の通常の検索質問を用いた検索結果と比較を行うこととした。

NTCIR9-GeoTime では、各検索課題について、各々のチームから最大 3 つの検索結果を受け取り、その上位 100 件をプーリングして、適合文書の判定を行っている。

回答を用いた検索結果は、検索質問と適合文書の間のミスマッチが小さいことが期待されることから、従来の検索システムでは、発見することが難しい検索課題についても、適合文書を発見できることが期待される。また、答の情報を持つ文書に対するスコアが上昇することが期待されるため、適合文書を網羅的に発見することが期待できる。

この事を確認するために、以下の二つの指標を調べることとした。

- 「回答を用いた検索結果」によってのみ得られたユニーク適合文書数
  検索式中の単語と適合文書の間のミスマッチが大きい場合に、一般の情報検索システムでは適合文書を網

表 1 課題ごとの Boolean 式に利用した単語と二つの指標

課題番号	利用した単語	適合文書数	ブーリアン式を満たした文書数	*CoverageRel* (CR)	*FocusAppropriateness* (FA)
26	スペースシャトル, コロンビア, 事故	115(21)	297	0.95(0.90)	0.37(0.06)
27	コンコルド, 飛行	5(3)	46	0.80(1.00)	0.09(0.07)
28	ワシントン, 連続狙撃事件	11(2)	53	0.91(1.00)	0.19(0.04)
29	ユーロ, 流通	143(10)	183	0.59(0.90)	0.46(0.05)
30	スティーブ・フォセット, 世界一周	2(2)	17	1.00(1.00)	0.12(0.12)
31	開通, つり橋	90(22)	60	0.36(0.73)	0.53(0.27)
32	アメリカ軍, ロープウェー	19(5)	0	0.00(0.00)	—(—)
	［米軍, ロープウエー］	19(5)	24	0.89(1.00)	0.71(0.21)
33	砒素, 死亡	143(92)	2	0.00(0.00)	0.00(0.00)
	［ヒ素, 死亡］	143(92)	303	0.57(0.64)	0.27(0.19)
34	全日空機, ハイジャック	24(14)	141	0.96(0.93)	0.16(0.09)
35	パイプライン, 死者	10(8)	31	0.40(0.38)	0.13(0.10)
36	パイプライン, 死者	10(9)	31	0.40(0.33)	0.13(0.10)
37	死亡者, 事故	9(6)	127	0.00(0.00)	0.00(0.00)
38	植民地, 中国, ヨーロッパ	43(21)	45	0.02(0.05)	0.02(0.02)
	［返還, 植民地, 中国］	43(21)	89	0.35(0.52)	0.17(0.12)
39	原子力潜水艦, 沈没	64(18)	263	0.55(0.56)	0.13(0.04)
40	コンコルド, 墜落	41(14)	63	0.95(1.00)	0.62(0.22)
41	パナマ運河, 返還	21(10)	37	0.95(1.00)	0.54(0.27)
42	中東, 国王	64(26)	270	0.41(0.62)	0.10(0.06)
43	ニューイングランド・ペイトリオッツ	9(9)	10	0.22(0.22)	0.20(0.20)
	［ペイトリオッツ, スーパーボウル)	9(9)	69	0.89(0.89)	0.12(0.12)
44	南アメリカ, 死者, 地震	64(27)	0	0.00(0.00)	—(—)
	［死者, 地震］	64(27)	1016	0.58(0.67)	0.04(0.02)
45	ヨーロッパ中央銀行, 設立	48(3)	0	0.00(0.00)	—(—)
	［欧州中央銀行］	48(3)	604	0.98(1.00)	0.08(0.00)
46	ブッシュ, ケリー, 大統領選, 討論会	41(4)	65	0.78(0.75)	0.49(0.05)
47	ケーブルカー, 火災	86(15)	88	1.00(1.00)	0.98(0.17)
48	国際刑事裁判所, 発効	70(12)	51	0.60(0.83)	0.82(0.20)
49	国際子ども図書館	65(10)	82	1.00(1.00)	0.79(0.12)
50	CAFTA	2(1)	3	1.00(1.00)	0.67(0.33)

() 内の数字は、完全適合のみの数値で、() なしの数値は部分適合を含む。
［］内の単語と対応するデータは、適合文書を考慮して、検索語の追加・削除や、同義語への変更などを行った参考データ

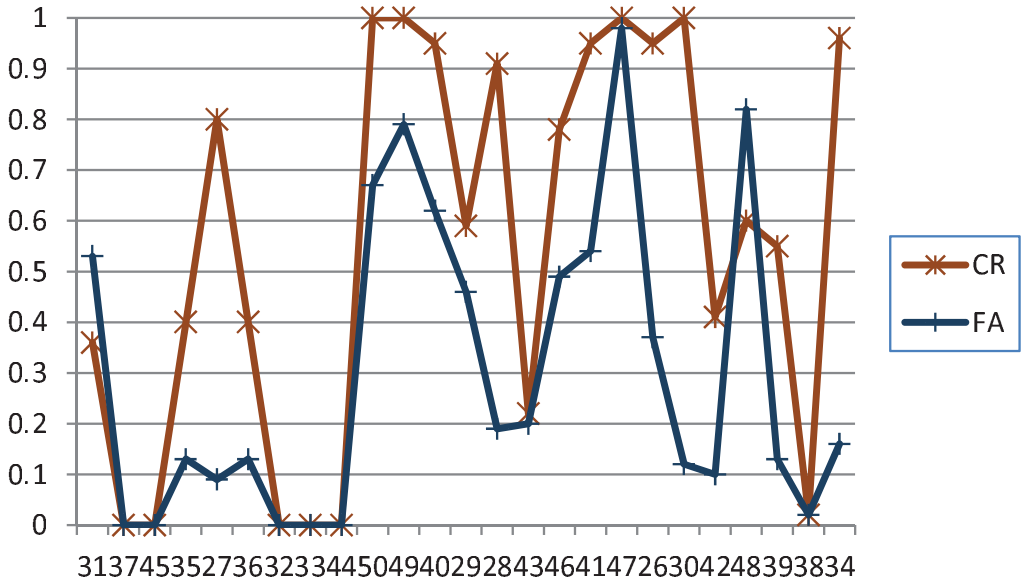

図 1 ベースラインの検索結果と CR・FA の関係

羅的に見つけることが難しくなる。この値が大きい課題は、回答を用いることの有用性が高い検索課題と言える。

- 「回答を用いた検索結果」によって発見された適合文書数
  回答を用いた検索結果により、網羅的に適合文書が発見されることが保証されるのであれば、回答を利用した検索結果のみでテストコレクションを作った場合でも、その有用性が高いことが期待される。

表 3 に「回答を用いた検索結果」のみが見つけた文書数を示す (() なしの数は、部分適合を含み、() 付の数は、完全適合のみの値である)。注目したいのは、検索課題と適合文書のミスマッチが大きい CR=0 の 5 課題中 4 課題 (33,37,44,45) でユニーク文書を発見している点である。これらのことから、回答を利用することにより、特別の機能を必要とするような難易度の課題に対して、特に、適合文書の網羅性の向上に寄与することが確認された。

次に、「回答を用いた検索結果」のみを使ってプールを作成した場合の網羅率を表 4 に示す。適合文書数が 100 件以上の課題における網羅率は低いが、これは、プールを 100 件に限っていることが原因の一つである。実際に、判定を行いながら、プールを増やすといった対策をとることで、一定レベル以上の網羅率を持ったテストコレクションが作成できる可能性を示していると考えている。

表 3 ユニーク適合文書数

検索課題 ID	適合文書数	ユニーク文書数
26	115(21)	9(4)
27	5(3)	0(0)
28	11(2)	0(0)
29	139(10)	1(0)
30	2(2)	0(0)
31	90(22)	24(6)
32	19(5)	0(0)
33	135(92)	16(9)
34	24(14)	0(0)
35	10(8)	0(0)
36	10(9)	0(0)
37	9(6)	1(0)
38	43(21)	0(0)
39	64(18)	3(1)
40	41(14)	0(0)
41	21(10)	0(0)
42	64(26)	3(0)
43	9(9)	0(0)
44	64(27)	19(5)
45	47(3)	14(0)
46	41(4)	3(0)
47	86(15)	0(0)
48	70(12)	0(0)
49	65(10)	0(0)
50	2(1)	0(0)

5. おわりに

本論文では、質問応答のための情報検索といった特定のタスクを考慮した情報検索テストコレクション構築についての考察を行った。特に、一般的な情報検索システムにおいて、容易に、適合文書を探し出すことができる課題の特徴について議論を行い、課題を作成する際に考慮すべき事項を提案した。また、質問応答のためのテストコレクションを作成する場合に、回答の情報を使う利点について説明し、回答を使うことで、一定以上の網羅性を担保したテストコレクションが作成できる可能性を示した。

表 4 適合文書網羅率

検索課題 ID	適合文書数	網羅率
26	115(21)	0.73(1.0)
27	5(3)	1.0(1.0)
28	11(2)	1.0(1.0)
29	139(10)	0.62(1.0)
30	2(2)	1.0(1.0)
31	90(22)	0.98(1.0)
32	19(5)	0.79(1.0)
33	135(92)	0.69(0.73)
34	24(14)	1.0(1.0)
35	10(8)	1.0(1.0)
36	10(9)	0.9(0.89)
37	9(6)	1.0(1.0)
38	43(21)	1.0(1.0)
39	64(18)	0.98(1.0)
40	41(14)	1.0(1.0)
41	21(10)	1.0(1.0)
42	64(26)	0.92(0.88)
43	9(9)	1.0(1.0)
44	64(27)	1.0(1.0)
45	47(3)	0.98(1.0)
46	41(4)	1.0(1.0)
47	86(15)	0.99(1.0)
48	70(12)	0.81(1.0)
49	65(10)	1.0(1.0)
50	2(1)	1.0(1.0)

今後は、他のテストコレクションについても、同様の分析を行い、本提案の妥当性について検討していきたいと考えている。

謝辞 本研究の一部は、NII 共同研究により行われた。また、NTCIR9-GeoTime タスクのテストコレクション作成者であるオーガナイザと参加者に感謝の意を表す。

参考文献

[1] 岸田和明：情報検索技術とテストコレクション, 情報処理, Vol. 41, No. 8, pp. 898–901 (2000).

[2] Greenwood, M. A.(ed.): *Proceedings of the 2nd workshop on Information Retrieval for Question Answering* (2008). http://www.aclweb.org/anthology/W/W08/W08-18.pdf.

[3] Sakai, T., Kando, N., Lin, C.-J., Mitamura, T., Shima, H., Ji, D., Chen, K.-H. and Nyberg, E.: Overview of NTCIR-7 ACLIA IR4QA Task, *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, pp. 63–93 (2010).

[4] *GIR '10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, New York, NY, USA, ACM (2010).

[5] Gey, F., Larson, R., Kando, N., Machado-Fisher, J. and Sakai, T.: NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search, *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, pp. 147–153 (2010).

[6] Sakai, T., Kitani, T., Ogawa, Y., Ishikawa, T., Kimoto, H., Keshi, I., Toyoura, J., Fukushima, T., Matsui, K., Ueda, Y., Tokunaga, T., Tsuruoka, H., Nakawatase, H., Agata, T. and Kando, N.: BMIR-J2: a test collection for evaluation of Japanese information retrieval systems, *SIGIR Forum*, Vol. 33, No. 1, pp. 13–17 (1999).

[7] Sakai, T., Shima, H., Kando, N., Song, R., Lin, C.-J., Mitamura, T., Sugimoto, M. and Lee, C.-W.: Overview of NTCIR-8 ACLIA IR4QA, *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, pp. 63–93 (2010).

[8] Gey, F., Larson, R. R., Machado, J. and Yoshioka, M.: A Micro-analysis of Topic Variation for a Geotemporal Query, *Proceedings of the 4th International Workshop on Evaluating Information Access (EVIA), A Satellite Workshop of NTCIR-9, December 6, 2011 Tokyo Japan*, pp. 9–13 (2011).

[9] Yoshioka, M.: Evaluating Topic Difficulties from the Viewpoint of Query Term Expansion, *Information Retrieval Technology: Third Asia Information Retrieval Symposium, AIRS 2006 Singapore, October 16-18, 2006, Proceedings* (Ng, H. T., kew Leong, M., yen Kan, M. and Ji, D., eds.), Springer-Verlag GmbH, pp. 390–403 (2006). LNCS4182.

[10] Yoshioka, M.: Towards Construction of Evaluation Framework for Query Expansion, *Information Retrieval Technology: Second Asia Information Retrieval Symposium, AIRS 2005 Jeju Island, Korea, October 2005 Proceedings* (Lee, G. G., Yamada, A., Meng, H. and Myaeng, S. H., eds.), Springer-Verlag GmbH, pp. 647–652 (2005). LNCS3689.

[11] Larson, R. R.: Probabilistic Text Retrieval for NTCIR9 GeoTime, *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, pp. 33–37 (2011).

[12] Gey, F., Larson, R., Kando, N., Machado-Fisher, J. and Yoshioka, M.: NTCIR9-GeoTime Overview - Evaluating Geographic and Temporal Search: Round 2, *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, pp. 9–17 (2011).

[13] Sato, T.: NTCIR-9 GeoTime at Osaka Kyoiku University - Toward Automatic Extraction of Place/Time Terms, *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, pp. 59–63 (2011).