情　報　学　基　礎　55－6
デジタル・ドキュメント　19－6
（１９９９．７．１６）

大規模テストコレクション構築について：NTCIR-1の訓練用検索課題の分析

栗山和子　神門典子
学術情報センター 研究開発部
{kuriyama,kando}@rd.nacsis.ac.jp

概要. 本稿では、評価用ツールとしてのテストコレクションにおける検索課題の性質について考察する。検索課題の望ましい性質として、「自然さ」と「難易度のバランス」があげられる。自然さとは、検索課題の内容が現実の検索過程においてシステムに与えられる検索要求と同様に自然なものでなければならないということである。「難易度のバランス」とは、検索課題が易しいすぎるものばかりでも難しすぎるものばかりでも、テストコレクション全体の性質が偏ったものになるので、難易度のバランスがとれているのが望ましいということである。

NTCIR-1では、検索課題を自然なものとするため、分野の研究者から収集している。本稿では、検索課題の難易度について、NTCIR-1の訓練用検索課題を用いて、検索課題そのものについて分析し、予備テストの評価結果との関連を調べた。

その結果、検索課題ごとの平均精度の中央値と、検索課題中の検索要求文の文字数、検索要求文中の単語がの出現する正解文書数、機能分類のそれぞれとには明らかな関連性は見られないものの、機能分類によるグループ分けは検索課題の難易度を予測するために、ある程度の参考になることがわかった。また、提出結果の平均精度の度数分布から、易しい検索課題、難しい検索課題というグループ分けの中でもその性質は一様ではないことがわかった。

Construction of a Large Scale Test Collection : Analysis of the Training Topics of the NTCIR-1

Kazuko Kuriyama　Noriko Kando
R & Dept., National Center for Science Information Systems (NACSIS)

Abstract. The paper discusses the quality of search topics in test collections, which are used in laboratory-typed testing of information retrieval systems. As a tool for evaluation, search topics in a test collection should be "natural" as search requests submitted by actual users, and balance their "difficulty". In the NACSIS Test Collection for Information Retrieval Systems 1 (NTCIR-1), search topics were collected from researchers of the subject domains, i.e., actual users of the systems providing access to scientific documents like NTICR-1, in order to prepare "natural" requests as much as possible. To estimate "difficulty" of topics, we analysed the results of the pretest, which used the training topics of the NTCIR-1. As results; (1) the average precision of each topic had no explicit relation with the number of characters/words/phrases, the number of relevant documents contain the words in the description, nor the number of relevant documents of each topic; (2) the search function needed to conduct search of the topic sometimes found effective to estimate "difficulty" of the topic; (3) the distribution of average precision over systems revealed that the nature of the topics were heterogeneous in a group of "easy topics" or "difficult topics".

1 はじめに

1.1 NTCIRプロジェクト

著者らは、学術情報センター研究開発部の「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（エンティサイル：NACSIS Test Collection for Information Retrieval systems）の構築を行なっている。その過程において、昨年 11 月から今年 9 月まで、テストコレクション 1 (NTCIR-1)（予備版）を用いた、コンペティション形式のワークショップを開催している [2],[3],[5]。

以前の論文 [5] では、NTCIR-1 を使用して、昨年 12 月に行なった予備テストの結果を用いたプーリング実験の結果とプーリングの有効性および判定者間の正解判定のゆれがシステム評価に影響を及ぼさないことについて報告した。本稿では検索課題の性質について考察する。

1.2 本研究の目的

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。

以前の論文 [5] では、正解文書リストの妥当性を検証するため、以下の問題について、システム評価への影響という観点から考察した。

(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) 正解判定の一致度

(1) について検証するため、NTCIR-1 の訓練用正解セットを使用した予備テスト（昨年 12 月）の提出結果を用いてプーリング実験を行ない、日本語の大規模テストコレクション構築におけるプーリングの有効性を確かめた。(2) については、プーリングに入れる各提出結果からの上位 X件を変化させて、プーリングする件数と提出結果の評価との関係を考察し、プーリング数が提出結果の相対的評価にほとんど影響しないことを示した。(3) について、正解判定は人間が行なう主観的な判断であることから、判定者間の判定のゆれは不可避な問題である。そこで、同一の検索課題に関する判定者の異なる複数の正解文書リストを用いた提出結果の評価を行ない、日本語の大規模テストコレクションの構築過程においても、正解判定のゆれによる提出結果間の評価の大きな違いが見られないことを示した。

本稿では、評価用ツールとしてのテストコレクションにおける検索課題の性質について考察する。検索課題に望ましい性質として、「自然さ」と「難易度のバランス」があげられる。[4] 自然さとは、検索課題の内容が現実の検索過程においてシステムに与えられる検索要求と同様に自然なものでなければならないということである。「難易度のバランス」とは、検索課題が易しいすぎるものばかりでも難しすぎるものばかりでも、評価に使用するテストコレクション全体の性質が偏ったものになるので、難易度のバランスがとれていることが望ましいということである。

難易度のバランスについては、TREC-6(the Sixth Text REtrieval Conference) [7][8] では、まず、正解文書リストを作成する前に、検索課題作成者を含む複数の判定者よって、検索課題を易しい検索課題 (easy topics)、中位の検索課題 (middle topics)、難しい検索課題 (hard topics) に分け、次に、正解文書リストが作成された後で、正解文書リストを用いて評価した参加者の提出結果の精度などによって難易度を数値的に表わし、人間による判定と検索結果の評価による数値的な難易度の相関係数を計算したところ、その間に相関があるとは言えなかったということが報告されている [8]。このようなことから、検索課題を作成する時点で検索課題の難易度を測るのは困難であると言われているが、実用的で公平なテストコレクションの構築という面からは、検索システムによる実験を行なう前に、検索課題の難易度が予測できることが望ましい。

また、BMIR[1] では、機能分類が用いられている。これは、検索課題の「難しさ」を示せる可能性のある試みとして、国際会議においても評価された。

NTCIR-1 では、検索課題を自然なものとするために、検索課題を分野の研究者から収集している。本稿では、検索課題の難易度について、NTCIR-1 の訓練用検索課題を用いて、検索課題そのものについての分析と予備テストの評価結果から考察する。

2 検索課題の分析

検索課題の性質を調べるため、予備テストで用いた訓練用検索課題について分析を行なった。本稿では、事例的な分析をするにとどめ、評価用検索課題とその提出結果に対する分析の布石とした。以下では、検索課題の分析のための難易度の一つの指標として、随時検索タスクの提出結果16セットについての検索課題30件のそれぞれの全適合文書のの平均精度（補間なし）を使用し、検索課題中の検索要求の機能分類、単語数、文字数などから見た検索課題の性質と関連について考察する。

2.1 予備テストの概要

NTCIR-1では、訓練用検索課題については、事務局で予め正解文書リストを作成した。このリストの正解文書の網羅性、システム評価に対する公平性を検証するため、NTCIRワークショップでは、昨年12月2日に予備テストを行なった[3]。

予備テストでは、訓練用検索課題30件に対する検索結果を、ワークショップ参加者から自由参加で提出してもらい、内部で用意した正解文書リストの網羅性を評価し、新たに発見された正解文書を追加した。

この予備テストでは、11チームで合計23セットの検索結果が提出された。23の内訳は、随時検索タスク8チーム16セット、言語横断タスク4チーム5セット、単言語（言語横断検索のためのbaselineとして）1チーム2セットである。本稿では、この23セットのうち、随時検索タスクの提出結果16セットを対象として実験を行なった。

一つの提出結果は、ある検索システムによる検索結果の、30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。システムの検索性能は、この30件の検索課題に対する検索結果の評価尺度の平均によって順位付けた。

2.2 検索課題

NTCIR-1では、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指定の検索要求収集用フォームによって収集された。本稿で使用する訓練用検索課題には大学図書館のレファレンス事例も含まれている。検索課題の詳細については、著者の以前の論文[3]を参照されたい。以下に、検索課題の例を示す。

```
〈検索課題 q=0006〉

〈タイトル〉
知的エージェント
〈/タイトル〉

〈検索要求〉
エージェント機能を利用した知的情報検索
〈/検索要求〉

〈検索要求説明〉
インターネット上の情報資源を対象とした情報検索、収集
に関する研究は、コンピュータネットワークの普及、大衆
化とともに非常に盛んになっている。一方、エージェント
という用語は人工知能をはじめとするいくつかの学問分野
での重要な概念となっている。両者を結びつけることによ
る知的な情報(検索)システムの研究は、(1)最近のトレン
ドであること、(2)エージェントという用語が広義かつ曖
昧であること、(3)既存の分野を横断する研究であること、
などからその現状や全貌を知るのは、しばしば困難である。
エージェント機能を「自律的に検索支援、収集代行を行な
うもの」と定義し、この機能を利用している情報検索シス
テムを正解とする。
〈/検索要求説明〉

〈概念〉
情報検索, 情報収集, インテリジェントエージェ
ント, 知的エージェント, 自律(システム), 情報
収集エージェント, インターネットロボット
〈/概念〉

〈分野〉
1. 電子・情報・制御
〈/分野〉

〈/検索課題〉
```

本稿では、検索課題の分析には、主に〈検索要求〉と〈/検索要求〉で括られた検索要求文を用いる。訓練用検索課題30件の検索要求文は以下の通りである。

0001: 自律移動ロボットについて
0002: 複合名詞の構造解析において、シンボリックな手法と統計的な手法を組み合わせたアプローチを取る研究はないか。
0003: 機械学習におけるサンプル複雑性について論じている文献
0004: モデルベースの文書画像理解について述べた文献
0005: クラスタリングにおける特徴次元リダクション
0006: エージェント機能を利用した知的情報検索
0007: 大規模なデータベースとユーザとのインタラクションにおけるユーザの認知的側面に関して論じている論文
0008: データマイニングの1分野である associative rule のマイニングに関して、Apriori アルゴリズムを改良した研究例などの最新の研究動向を知りたい。
0009: インターネットのトラヒック統計を解析している文献が欲しい。
0010: テキストからのキーワード自動抽出の手法について知りたい。
0011: 連結全域グラフを求めるアルゴリズムについて
0012: データマイニング手法を改良、提案している文献
0013: ループ領域解析アルゴリズム
0014: 故障診断システムについて
0015: テキストからのコロケーションの自動抽出について
0016: 最大共通部分グラフ問題について
0017: パスへの同時送信に対する排他制御について論じている論文
0018: マルチメディアネットワークにおける通信品質保証の実現と課題について述べたものはあるか？
0019: 日本語文の係り受け解析手法について
0020: 日本語文におけるカタカナ外来語の研究

0021: 機械翻訳における構造処理能力の評価
0022: 自然言語からの知識獲得法について知りたい。
0023: 新聞記事データの分析
0024: 機械翻訳システムについて
0025: LFG について
0026: 語彙機能文法について
0027: シソーラスを用いたテキストの連想検索について
0028: ニューラルネットワークの手法、理論、原理などについて記述した文献がほしい。
0029: 物体の位置計測について
0030: データ駆動画像処理システムの提案

2.3 精度からみた検索課題の難易度

各検索課題の正解個数、予備テストの随時検索タスクにおいて提出された 8 チーム 16 セットの提出結果の各検索課題についての平均精度の平均、標準偏差、中央値、難易度を以下の表に示す。0028 については正解文書数が提出結果の文書数の 1000 個よりも多いため、ここでは除いた。NTCIR-1 の正解判定リストでは、正解判定は、正解 (A)、部分的正解 (B)、不正解 (C) の 3 値である。本稿では、A と B の両方を正解とした場合の評価を用いる。

表 2-1. 検索課題の統計的データと難易度

topic	rel	ave	stdev	median	dffc
0001	293	0.3476	0.1629	0.4179	easy
0002	19	0.3663	0.2247	0.3870	easy
0003	14	0.0331	0.0264	0.0329	hard
0004	38	0.2931	0.1476	0.3248	mid
0005	13	0.0557	0.0549	0.0323	hard
0006	72	0.1352	0.1047	0.1370	hard
0007	16	0.0564	0.0444	0.0533	hard
0008	25	0.2212	0.1664	0.2533	mid
0009	8	0.1299	0.1293	0.1176	hard
0010	55	0.2565	0.1023	0.2824	mid
0011	7	0.3553	0.1819	0.4326	easy
0012	70	0.3157	0.2137	0.2056	mid
0013	38	0.0701	0.0265	0.0599	hard
0014	317	0.4888	0.1493	0.5692	easy
0015	20	0.2192	0.0990	0.2463	mid
0016	5	0.5973	0.2119	0.7064	easy
0017	16	0.1645	0.0841	0.1931	mid
0018	167	0.0963	0.0563	0.1142	hard
0019	92	0.3174	0.1799	0.4013	easy
0020	16	0.4434	0.2460	0.4350	easy
0021	11	0.2078	0.1206	0.1799	mid
0022	82	0.2448	0.1521	0.2414	mid
0023	98	0.3286	0.1188	0.3584	easy
0024	158	0.3247	0.1277	0.3449	easy
0025	23	0.5077	0.3277	0.7039	easy
0026	23	0.5253	0.2849	0.6342	easy
0027	23	0.2479	0.1116	0.2834	mid
0029	180	0.1451	0.0671	0.1717	hard
0030	23	0.1615	0.1080	0.1584	hard

topic:検索課題番号、rel:正解個数、ave:平均精度の平均、
stdev:平均精度の標準偏差、median:平均精度の中央値、dffc:難易度
easy:易しい、mid:中位、hard:難しい

平均精度の中央値 (median) で、easy:「易しい」、middle:「中位」hard:「難しい」として、表 2-1 の検索課題をグループ分けした。(表 3-1 参照)

3 分析結果

以下に、各検索課題についての、難易度、平均精度の分布、機能分類、検索要求文中の単語数、検索要求文の文字数を表として示す。

表 3-1. 検索課題の難易度と分析結果

topic	dffc	dstr	func	rel	word	char
0001	easy	A	B	293	4	12
0002	easy	M	F	19	9	52
0003	hard	I	F	14	4	26
0004	mid	A	F	38	7	22
0005	hard	I	F	13	5	21
0006	hard	A	F	72	7	19
0007	hard	I	D	16	6	48
0008	mid	A	F	25	15	60
0009	hard	A	D	8	5	29
0010	mid	A	C	55	6	28
0011	easy	M	F	7	5	21
0012	mid	M	B	70	6	22
0013	hard	I	F	38	5	13
0014	easy	A	B	317	3	12
0015	mid	M	D	20	5	23
0016	easy	M	E	5	6	15
0017	mid	A	F	16	5	27
0018	hard	A	D	167	10	43
0019	easy	M	D	92	6	17
0020	easy	A	C	16	6	18
0021	mid	A	F	11	6	18
0022	mid	M	C	82	4	17
0023	easy	A	A	98	5	22
0024	easy	A	A	158	3	10
0025	easy	M	B	23	1	12
0026	easy	M	B	23	4	5
0027	mid	A	F	23	5	10
0029	hard	A	D	180	4	11
0030	hard	A	F	23	6	16

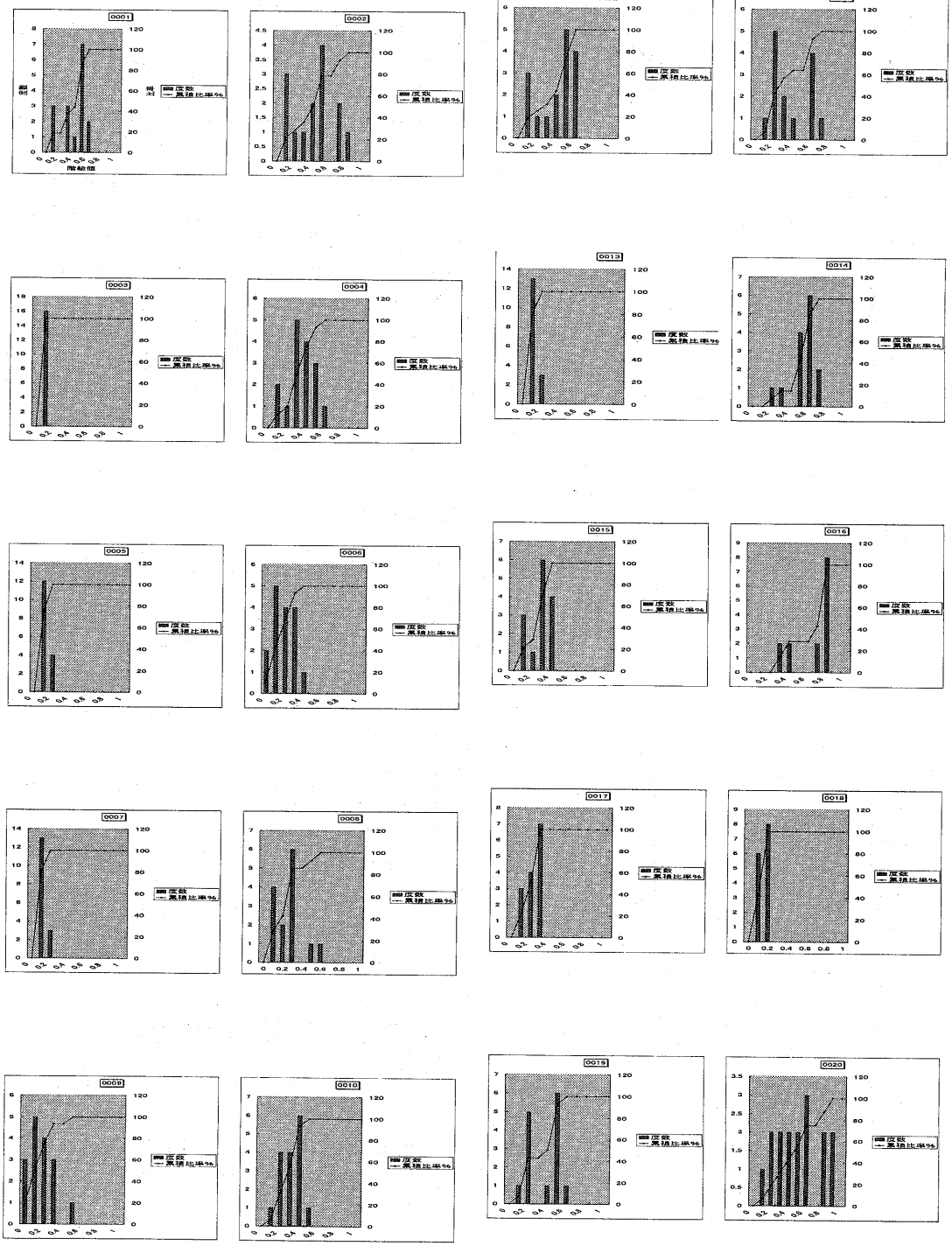
topic:検索課題番号、dffc:難易度、dstr:平均精度の分布、func:機能分類、
rel:正解個数、word:検索要求文中の単語・フレーズ数、char:検索要求文中の文字数、easy:易しい、mid:中位、hard:難しい

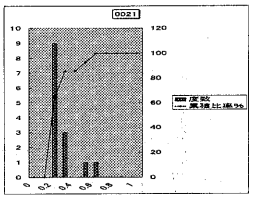
以下では、表 3-1 の各要素について説明する。

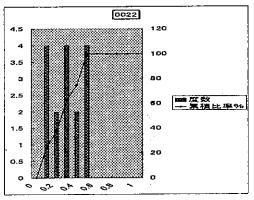
3.1 平均精度の分布

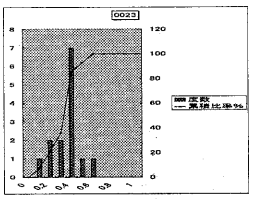
次に、検索課題ごとに各提出結果の平均精度のばらつきとを見るため、平均精度のヒストグラムを以下に示す。階級値は平均精度の 0 から 1 の間の 0.1 刻み、縦軸はその平均精度の提出結果の度数である。

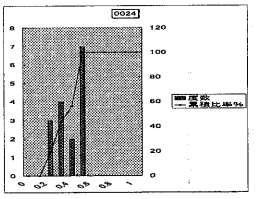
図 3-1. 平均精度のヒストグラム

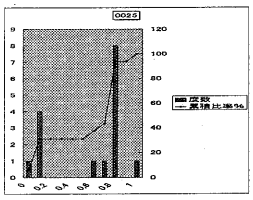


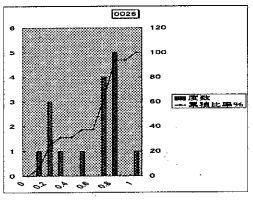


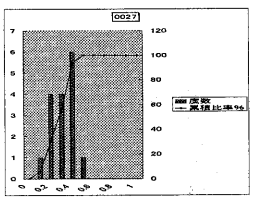


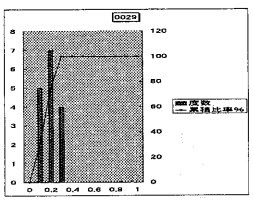












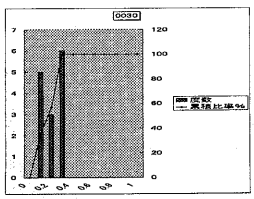


図 3-1 から分布から、分布の形を、A 型：低い階級値と高い階級値の度数が少なく、中位の階級値の度数が多いもの、例としては 0006、I 型：1 つあるいは隣合った 2 つの階級値に全てが含まれるもの、例としては 0005、M 型：低い階級値と高い階級値の度数が多く、中位の階級値の度数が少なく山が 2 つに分かれているもの、例としては 0019、として、検索課題を分けた。(表 3-1 参照)

平均精度によるグループ分け easy,middle,hard との関連を見ると、easy:「易しい」検索課題 11 件は A 型 5 件、M 型 6 件であり、middle:「中位」の検索課題 9 件は A 型 6 件、M 型 3 件であり、hard:「難しい」検索課題 9 件は A 型 3 件、I 型 6 件である。よって、easy は A 型と M 型の 2 つのパターンに同程度分かれ、middle は比較的きれいな分布の A 型が多く、hard は M 型のような検索精度が高い提出結果と低い提出結果に分かれるようなことはない、という傾向はあるが、明らかな相関があるとは言えない。

3.2 機能分類

検索課題を BMIR-J2[1] のファンクション分類に準拠し、以下の 6 つの機能を設定した。BMIR-J2 の「F1: 基本機能」を「F0:基本機能」と「F1:シソーラス機能」に細分した。検索課題中の検索要求文に含まれる語句を用いて正解文書を検索するために必要であるかないか判定をした。判定は 2 名の図書館情報学専攻の大学院生が行なった。

F0: 基本機能
キーワードの存在確認、あるいは、それらの語の存在に関する論理式 (AND や OR など) の充足判定など。

F1: シソーラス機能
キーワードのシソーラスによる展開語の存在確認。および、それらの語の存在に関する論理式 (AND や OR など) の充足判定。

F2: 数値・レンジ機能
数の数え上げや、数値などの範囲を正しく解釈する。数値の大小比較や単位の理解・変換なども含む。

F3: 構文解析機能
複数のキーワードの間の係り受け関係を判断する (構文解析する)。

F4: 内容分析機能
通常の構文解析に必要とされるよりも深い言語知識を利用する。文脈を理化いすることや、言葉の深い意味を理解することを含む。

F5: 知識処理機能
世界知識を利用する。常識的な判断や、蓄積された事実からの推論などを含む。

判定の結果を o/x で、(F0 F1 F2 F3 F4 F5) = ?????? という形で表し、判定のパターンによって検

索課題を以下のようなA～Fという6グループに分類した。

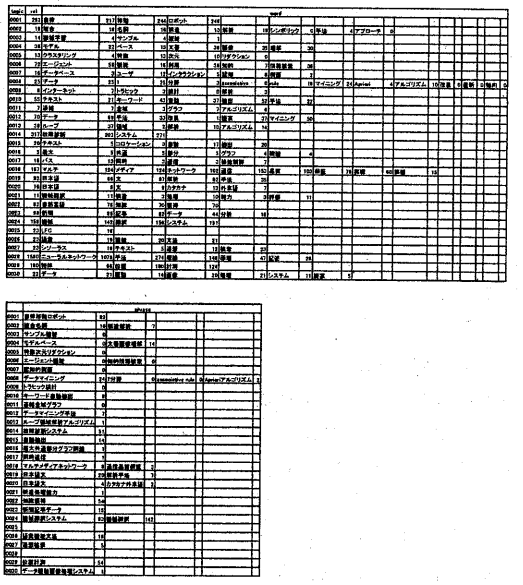
A: 基本機能のみ	oxxxxx
B: シソーラス機能のみ	ooxxxx
C: 構文解析機能のみ	oxxoxx
D: シソーラス機能と構文解析機能	ooxoxx
E: シソーラス機能と内容分析機能	ooxxox
F: シソーラス機能と構文解析機能と内容分析機能	ooxoox

機能分類でグループを分けた。(表3-1参照)
機能分類のグループでは、A,B,C,D,E,Fの順に検索要求に対して必要な処理が多くなることから、その順に難しくなっていると考えることができる。しかし、平均精度と機能分類の関連を考えたとき、機能分類のA(2件)とB(5件)の検索課題は0012のBを除いてeasyの検索課題11件に含まれている。しかし、easyの残りの5件はそれぞれC:1件、D:1件、E:1件、F:2件とばらばらであり、必要な機能が少なければ、易しい検索課題であるとは必ずしも言えない。また、hardの検索課題9件についても同様であり、そのうちの6件はF(12件)の中に含まれ、残りはDに含まれており、hardの検索課題は構文解析が必要であるということが分かるが、構文解析の必要な検索課題はeasyにも含まれているので、その逆は言えない。したがって、機能分類は、検索課題の難易度を予測するためにある程度は参考になるが、明らかな関連があるとは言えない。

3.3 正解文書数中への単語の出現

検索課題中の検索要求文に含まれる単語とフレーズの正解文書中への出現が検索課題の難易度と関連があるかどうか調べる。検索要求文の単語への分割には、日本語形態素解析システムChasen Ver1.5 [6]を使用したプログラムによって行なった。以下に、各検索課題の正解文書数と検索要求文中の単語との出現文書数を示す。

表3-2. 正解文書数と検索要求文中の単語とフレーズの出現文書数



topic:検索課題番号、rel:正解個数、
word:検索要求文に含まれる単語、phrase:検索要求文に含まれるフレーズ

検索要求文中の単語とフレーズには、検索課題のタイトルに含まれている主要な語以外の単語も含まれているため、単純に単語数とその出現文書数を平均精度による難易度のようにグループ分けすることは難しいため、事例をあげて考察する。

主要な単語のうちのいずれか1つが正解文書全体に2分の1未満にしか出現していない検索課題は、0003,0005,0006,0007,0008,0009,0013,0015,
0017,0018の9件であるが、平均精度から見た難易度においては、このうちの8件がhardであり、1件がmiddleである。一方、主要な単語のいずれもが全体の3分の2以上の文書に出現している検索課題は、0002,0004,0010,0011,0016,0022,0023,
0024,0025,0026,0027,0029の12件であり、それぞれeasy 7件 middle 4件 hard 1件である。よって、主要な単語が正解文書数にそのまま出現してこないようなものは検索精度が低くなる可能性があるが、必ずしも難しい検索課題全てに対して同じことが言える訳ではない。

3.4 検索要求文中の文字数

検索要求文中の文字数を示す。ただし、英単語は1単語＝1文字とする。文字数（表3-1参照）によって、C1:1～20個、C2:21～50個、C3:51個以上として検索課題をグループ分けた。(表3-1参照) このグループを平均精度の中央値 (median) による難易度easy,middle,hardと比べてみると、相関がないことがわかった。

4 まとめ

大規模テストコレクションでは、検索システムの評価の公平性と、他のテストコレクションを用いた評価と比較という点から、検索課題は自然で難易度の面でバランスのとれたものであることが望ましい。検索課題を作成する時点で検索課題の難易度を測るのは困難であるが、テストコレクションの構築においては、検索システムによる検索実験を行なう前に、検索課題の難易度が予測できることが望ましい。

NTCIR-1では、検索課題の自然さを、検索課題を分野の研究者が作成することによって実現している。本稿では、検索課題の難易度について、NTCIR-1の訓練用検索課題を用いて、検索課題の特性と予備テストの評価結果との関連を調べた。その結果、検索課題ごとの平均精度と、検索課題の、文字数、単語の出現する正解文書数、機能分類には明らかな関連性は見られないものの、機能分類によるグループ分けは検索課題の難易度を予測するためにある程度の参考にはなることがわかった。また、提出結果の平均精度の度数分布にから、検索課題の易しさ、難しさが一様ではないことがわかった。

今後の予定としては、評価用テストの結果を用いて評価用検索課題の性質と提出結果の評価の詳しい統計的な処理を行ない、バランスのとれた検索課題群の作成の参考としたい。

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] 情報検索システム評価用テストコレクション BMIR-J2 （情報検索システム評価用ベンチマーク Ver.2）利用説明書.

[2] Kando, N. et.al. NTCIR: "NACSIS Test Collection Project". [Poster] the 20th Annual Collection of BCS–IRSG, France, 1998.

[3] 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.

[4] 神門典子. "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, pp.129–136, 1999.

[5] 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.

[6] 松本裕治ほか. 日本語形態素解析システム『茶筌』version 1.5 使用説明書, 奈良先端科学技術大学院大学, http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html, 1997.

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

[8] Voorhees, E.; Harman, D. "Overview of the Sixth Text REtrieval Conference(TREC-6)", NIST Special Publication 500-225.