検索課題の難易度を考慮したテキスト検索システムの評価

江口浩二† 栗山和子† 神門典子†

本論文では，テキスト検索システム評価用テストコレクションの構築あるいは利用において，考慮すべき重要な要素の１つである検索課題の難易度について様々な観点から分析を行う．第１に，テストコレクションの信頼性の観点から，検索課題の難易度が検索システムの有効性に関する相対的評価に与える影響を分析する．第２に，検索課題難易度の予測可能性を検討する．その目的のもと，文書データベース中の語の頻度情報や人間による判定などに基づいて，検索課題に関する各種特徴量を定義し，それらと検索課題の難易度に関する相関性を分析する．以上に関してテストコレクションNTCIR-1を対象に行った分析の結果，テストコレクションを用いたテキスト検索システムの評価において，検索課題の難易度のレベルに基づいた分類ごとに評価を実施することが有効であると確認された．また，検索課題の難易度は文書データベース中の語の頻度情報に起因することを示唆する分析結果を得た．

Evaluation of Text Retrieval Systems Considering Topic Difficulty

Koji Eguchi,† Kazuko Kuriyama† and Noriko Kando†

This paper analyzes topic difficulty as one of important factors for construction or use of test collections. First of all, we analyze the differences of system ranking affected by the topic difficulty, from the point of view of reliability of test collections. Secondly, we investigate the predictability of topic difficulty. With this objective, this paper defines measures for the various features of the topics, on the basis of term frequencies in the document databases or human judgments, and analyzes the correlation between the topic difficulty and them. Through the results of the investigations using the NTCIR-1 test collection, the topic categorization based on its difficulty turned out to be effective in evaluating text retrieval systems using a test collection. The results also suggest that the topic difficulty depend on the topic term frequencies in the document database.

1. はじめに

近年，いくつかの評価ワークショップが実施され，注目を集めつつある．**評価ワークショップ**（evaluation workshop）とは，複数の参加者による複数のシステムを用いて，ある問題を解決する情報技術の性能を共通の基盤の上で評価することにより，相互の特徴比較を行うことを目指すものである．情報検索システムに関する評価ワークショップとしてはTREC[1]が知られており，共通のテストコレクションを用いて各システムの**検索有効性**（retrieval effectiveness）が比較される．ここで，**テストコレクション**（test collection）とは，(1) 文書データベース，(2) 検索課題の集合，(3) 各検索課題に対する適合文書セットからなる検索実験用データセットのことである．日本語を対象としたテストコレクションとしてはBMIR-J1とBMIR-J2[2]があるが，最近では評価ワークショップとしてNTCIRプロジェクト[3]が実施され，テストコレクションの構築などにおいて成功を収めている．

本論文では，情報検索システム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関して，テストコレクションNTCIR-1[4]を用いた分析結果を報告する．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．検索課題の内容は，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あることが望ましい．NTCIR-1では，検索課題を自然なものとすることを目指し，検索課題の作成を各領域の専門家から収集した．ところで，検索課題がやさしすぎるものや難しすぎるものに偏る場合，テストコレクションが情報検索システムの有効性を十分に評価するに足らず，あの特定の条件下における有効性の評価にしか利用できない可能性が増す．このような問題を避けるため，

† 国立情報学研究所
National Institute of Informatics

検索課題の難易度にバランスがとれていることが望ましい．難易度のバランスについては種々の観点から議論することができるが，本論文では検索課題の相対的な難易度の高いものと低いものが，均等に分布している状態が望ましいととらえる．これを実現するには，個々の検索課題の難易度もしくは複数の検索課題の難易度分布が，参加者による検索実行前に予測できる必要があるが，一般に容易ではない☆．TREC-6では検索課題の難易度に関して基礎的な検討が実施された．その結果，人間が検索課題文を閲覧して判定することによる検索課題の難易度の分類と，参加者が提出した検索結果の評価による数値的な難易度に，相関があるとはいえないことが報告されている[1]．

本研究では，より多様な観点から，NTCIR-1を対象に検索課題の難易度に関する分析を実施する[6)~8)]．まず，テストコレクションの信頼性という観点から，検索課題の難易度が検索システムの相対的評価に与える影響を分析する．次に，NTCIR-1における検索課題，文書データベースおよび適合文書セットに関する種々の特徴量を計量し，それらと検索課題の難易度との相関性を分析することにより，検索課題の難易度の予測可能性について検討する．

2. テストコレクション NTCIR-1

本章では，テストコレクション NTCIR-1[4)]を構成する，(1) 文書データベース，(2) 検索課題，(3) 適合文書セットのそれぞれについて概要を述べる☆☆．

2.1 文書データベース

国立情報学研究所が日本国内の65学協会の協力を得て，全国大会や研究会などの発表論文の要旨を集めた学会発表データベース[11)]から，約33万件の文書を選択し，各文書ごとに特定の項目を抽出したものが用いられた[4)]．約半数の文書は日英対訳であり，各レコードは，表題，著者名，会議録名，学会名，発表年月日，要旨，著者キーワードから成る．

☆ 実際の検索課題作成の戦略として，NTCIR-1およびNTCIR-2では，単一のシステムによる検索結果の上位100件以内に5件以上の適合文書が含まれていることを，検索課題の条件にしている．TREC-8では，上位25件以内に1〜20件の適合文書が含まれており，かつ，それらを利用した適合フィードバックによる検索結果の上位100件以内に10件以上の適合文書が含まれていることを，条件にしている[5)]．これらの閾値に理論的な根拠は示されていない．

☆☆ テストコレクション NTCIR-1には，情報検索システムの評価を目的とした (1)，(2)，(3) に加えて，自然言語処理の基礎的データを提供することを目的としたタグ付きコーパス[9)]が含まれているが，本論文では分析の対象としない．また，テストコレクション NTCIR-2[10)]を用いた分析については稿を改めて報告する．

```
⟨TOPIC q=0035⟩
⟨TITLE⟩
電子図書館
⟨/TITLE⟩
⟨DESCRIPTION⟩
分散環境における電子図書館についての研究はないか．
⟨/DESCRIPTION⟩
⟨NARRATIVE⟩
様々な人がネットワークを利用するようになり，ネットワークを介した情報提供サービスも数多く実現してきている．電子図書館もその１つでネットワークを通じて遠くにある電子化された出版物や画像を検索したり閲覧するというサービスが行われてきている．ネットワーク上の利用者や資源は基本的に分散して存在するものであり，電子図書館に保存される資料も複数の場所に分散してることも考えられる．このように，電子図書館を分散環境で利用するために必要な技術について述べている論文が欲しい．ネットワークを通じての電子図書館の利用について知りたいので，所蔵品を電子化して検索できるシステムを設置しましたという論文は要求を満たさない．新しい研究を始めるにあたり，このトピックの現状を知りたい．
⟨/NARRATIVE⟩
⟨CONCEPT⟩
⟨J.CONCEPT⟩
a. 電子図書館,
b. 分散環境，ネットワーク
⟨/J.CONCEPT⟩
⟨E.CONCEPT⟩
a. Digital Library, Electronic Library, Virtual Library,
b. Distributed System, Distributed Environment, Network
⟨/E.CONCEPT⟩
⟨A.CONCEPT⟩
c. Z39.50
⟨/A.CONCEPT⟩
⟨/CONCEPT⟩
⟨FIELD⟩
1. 電子・情報・制御
⟨/FIELD⟩
⟨/TOPIC⟩
```

図 1 検索課題の例
Fig. 1 A sample topic.

2.2 検索課題

検索課題（search topic）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-1では，訓練用30課題，評価用53課題が作成された．これらは，各領域の専門家（大学院生以上の研究者）から収集したものである．**図 1** に検索課題の例を示す．検索課題は主に，検索要求文（description），検索要求説明（narrative），タイトル（title），概念語リスト（concept）☆☆☆，分野（field）から成る．検索要求文は，利用者の検索要求を1文で記述したものである．検索要求説明は，背景説明・検索の目的・適合判定基準・用語の定義などを含み，検索要求を第三者が理解することを促す．タイトルは検索課題を数語で表現したものであり，概念語リストは検索課題における重要

☆☆☆ 図1における⟨J.CONCEPT⟩は日本語で記述された概念語リスト，⟨E.CONCEPT⟩は英語で記述された概念語リスト，また，⟨A.CONCEPT⟩は頭字語の概念語リストを示している．

な概念に関する同義語・類義語のリストである．検索実験では以上のいずれかの項目を処理してクエリを自動作成してもよい（以下，非対話型システム）が，人間が検索課題の記述を参照しながら対話的にクエリを入力してもよい（以下，対話型システム）．ただし，結果提出に際しては検索課題のどの項目を使用したか，対話型システムと非対話型システムのいずれであるかを報告する必要がある．

2.3 適合文書セット

日本語の検索要求に対して日本語および英語の適合文書を検索する「随時検索タスク」と日本語の検索要求に対して英語の適合文書を検索する「言語横断検索タスク」の2つのタスクに対して，ワークショップの参加者が各自のシステムによる検索結果を提出し，それらに基づく評価が実施された．検索課題ごとに，各システムの検索結果文書リスト（以下，提出結果と呼ぶ）の上位一定数の和集合に加え，別途に実施した再現率重視の対話型検索の結果に対しても，**適合判定**（relevance judgment）を実施することで，網羅的な適合文書セットを収集することを目指す[12]．適合判定に際しては，2名のクロスチェックに基づく最終判定が行われた．また，判定は検索要求に「適合」,「部分的適合」,「不適合」の3段階で実施された☆．ここで，部分的適合とは検索課題に記述された検索要求の一部に関してのみ適合であることを意味する．

3. 検索課題難易度がシステム順位に与える影響の分析

3.1 検索課題難易度の定義

実際の検索課題の難易度を特定するために，提出結果ごとの検索有効性を示す**非補間平均精度**（non-interpolated average precision）の中央値に基づいて検索課題を分類するものとする．このとき，2.3節に述べた随時検索タスクにおいて，検索課題中の検索要求文のみを用いた26の非対話型システムに関する検索結果に基づき，評価用検索課題の分類を実施した☆☆．対話型システムか非対話型システムか，あるいは，検索課題中のどの項目を使用したかによって，検索有効性の分布の傾向が異なることを避けるためである．

検索課題ごとの提出結果リストに関して，次の各種統計値を求めた．

(1) 適合と判定された文書の総数（$|REL|$），

(2) 提出結果リストにおける非補間平均精度の分布に関する平均値（ave），標準偏差（$stdev$），中央値（med），歪度（$skew$），尖度（$kurt$）．

特に，上記の非補間平均精度の中央値を検索課題の難易度の指標と見なし，これの値の昇順に検索課題を並べかえ，さらに検索課題を3つの難易度レベルに等分割した．各レベルは前述の中央値の降順に「hard」,「middle」,「easy」とし，これらを**検索課題難易度**（topic difficulty）と呼び，*diff*と表記する．

3.2 検索課題難易度レベルごとのシステム順位の比較

あるシステムは平均的な難易度を持つ検索課題に対して有効な検索処理を実現するが，難易度の高い検索課題に対しては有効でないことがありうる．逆に，他のシステムは，ある種の難易度の高い検索課題に対して，特に有効な検索処理を実現できるかもしれない．そこで，本論文では，システム順位が検索課題難易度に影響されるかどうかを確認するため，検索課題難易度のレベルごとにシステム順位を求め，それらの相関性について分析する．

3.1節で定義した3段階の検索課題難易度レベル*diff*の各々に対して，非補間平均精度の平均値に基づいたシステム順位について調べる．3.1節で述べた，26のシステム順位を分析の対象とする．**表1**にランキング上位のみの抜粋を示す．なお，表中のrun IDは，ワークショップ参加者の特定のシステムによる提出結果を指す．また，同表に，非補間平均精度の平均値（ave）および，ランキングにおいて1位だけ順位が上がるに応じた非補間平均精度の増加百分率（%increase）を併記する．

検索課題難易度ごとのシステム順位の相関性について順位相関係数Kendallのτを用いて分析する☆☆☆．検索課題難易度ごとのKendallのτおよび有意水準αの算出結果を**表2**に示す．表2のとおり，すべての検索課題難易度レベルの組合せについて0.7から0.9程度の有意な相関が見られたことから，検索課題難易度が異なる場合でもシステム順位に有意な異なりは生じないことが示唆される．しかしながら，表1から分かるとおり，検索課題難易度ごとの各システム順位の上位において順位の入れ替わりが観察されたが，統計的に有意であるとされている平均精度の平均値の増加率[14]5%を超えて順位が入れ替わる例が見られた．このことから，検索課題難易度は，システムの相対的評

☆ 3章および5章においては,「適合」あるいは「部分的適合」と判定された文書を適合文書と見なして分析を行った．

☆☆ 訓練用検索課題の予備的分析については文献8)を参照されたい．

☆☆☆ システム順位の比較分析にKendallのτを用いる例は，他の研究（たとえば文献13)）においても見られる．

表 1 検索課題難易度レベルごとのシステム順位の抜粋
Table 1 System ranks of top runs for three topic difficulty levels.

	easy			middle			hard			all		
rank	run ID	*ave*	%increase	run ID	*ave*	%increase	run ID	*ave*	%increase	run ID	*ave*	%increase
1	K32002	0.65	2.4	R2D22	0.33	6.1	jscb1	0.19	59.5	jscb1	0.38	8.4
2	jscb1	0.63	0.3	jscb1	0.31	9.9	K32001	0.12	2.7	K32002	0.35	0.7
3	K32001	0.63	5.4	K32001	0.29	0.6	K32002	0.11	3.5	R2D22	0.35	0.6
4	R2D22	0.60	2.2	K32002	0.28	2.6	R2D22	0.11	7.3	K32001	0.35	7.3
5	R2D24	0.58	4.4	R2D21	0.28	0.5	R2D24	0.10	13.1	R2D24	0.32	3.9
6	R2D21	0.56	2.9	R2D24	0.28	8.8	BKJJBIDS	0.09	0.8	R2D21	0.31	5.8
7	BKJJBIDS	0.54	1.8	NTE151	0.25	5.5	R2D21	0.09	9.2	BKJJBIDS	0.29	2.3
8	R2D23	0.53	1.8	BKJJBIDS	0.24	0.7	R2D23	0.08	1.3	R2D23	0.29	4.8
9	CRL12	0.52	0.1	R2D23	0.24	4.4	FX1	0.08	10.6	CRL14	0.27	2.1
10	CRL8	0.52	1.1	CRL14	0.23	4.6	CRL14	0.07	9.9	CRL13	0.27	0.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表 2 検索課題難易度レベルごとのシステム順位に関する Kendall の順位相関係数
Table 2 Kendall's rank correlation coefficients between system ranking for three topic difficulty levels.

		easy	middle	hard	all
easy	τ		*0.809*	*0.717*	*0.914*
	α		0.000	0.000	0.000
middle	τ			*0.698*	*0.883*
	α			0.000	0.000
hard	τ				*0.766*
	α				0.000
all	τ				
	α				

τ：Kendall の順位相関係数，α：両側有意水準，
強調：相関係数が 1%水準で有意（両側）．

価に一定の影響を与えると見なすことができる．

4. 検索課題の特徴量

本章では，検索課題の難易度に関する予測可能性を検討するため，人間による判定や文書データベース中の語の頻度情報などに基づいて，検索課題に関する各種特徴量を定義する．

4.1 機 能 分 類

機能分類（function-based topic categorization）とは，ある検索課題を充足する検索結果を獲得するに必要とされる検索システムの機能に基づき，検索課題を分類したものである．検索課題を BMIR-J2 の機能分類[2)]に準拠し，以下の 6 種の機能を設定した．ただし，BMIR-J2 における「基本機能」を，本論文では「F0. 基本機能」と「F1. シソーラス機能」に細分した[8)]．

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拡張語の存在確認，および，それらの語の存在に関する論理式の充足判定．

F2. 数値・レンジ機能： 数の数え上げや数値の範囲に関する正確な解釈．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受け関係についての判断（構文解析）．

F4. 内容解析機能： 通常の構文解析に必要とされるよりも深い言語知識の利用．文脈を理解することや，言葉の深い意味を理解することを含む．

F5. 知識処理機能： 世界知識の利用．常識的な判断や蓄積された事実からの推論などを含む．

2 名の図書館情報学を専攻する大学院生により，NTCIR-1 の検索課題に対して機能分類の判定を実施した☆．判定の結果，該当する機能の有無を，それぞれ「1」,「0」で表し，**表 3** に示す．このような機能の有無のパターンによって検索課題を同図に示す 6 つのカテゴリに分類した．このとき，A, B, ···, F の順に必要とされる処理が多くなり，一般に困難とされる要素技術が加わることから，この順に検索要求に適切な検索の実行が困難になると考えられる．したがって，この分類を，人間により判定された検索課題の難易度に関する一指標と見なす．

4.2 検索課題文の特徴量

本論文では，検索課題の特徴を示す以下の検索課題文の各種特徴量に着目する．

(1) 検索課題文の特徴語数（$\#term$），文字数（$\#char$），
(2) 検索課題文の特徴語に関する文書データベース中の語頻度，
(3) 検索課題文の特徴語に関する文書データベース

☆ BMIR-J2 における機能分類[2)]と同様，本論文においても判定者は検索課題文のみを閲覧して機能分類を判定するものとする．

表 3 機能分類に基づく検索課題の分類結果
Table 3 The results of the function-based topic categorization.

機能に基づく検索課題のカテゴリ	F0	F1	F2	F3	F4	F5
A. 基本機能のみ：	1	0	0	0	0	0
B. シソーラス機能のみ：	1	1	0	0	0	0
C. 構文解析機能のみ：	1	0	0	1	0	0
D. シソーラス機能と構文解析機能：	1	1	0	1	0	0
E. シソーラス機能と内容解析機能：	1	1	0	0	1	0
F. シソーラス機能と構文解析機能と内容解析機能：	1	1	0	1	1	0

中の文書頻度.

以下に上記の特徴量を選択した理由を述べる．(1) については，検索課題文に特徴語が多く含まれるほど，有効な検索が可能になることを期待している．(2) および (3) については，一般に検索課題文を構成する特徴語が文書データベースに出現する頻度が少ないほど有効な検索が容易になるという直観による．

なお，検索課題文に対して形態素解析☆を実行して求めた形態素群に対して，いくつかの連接ルール[16]を適用して複合語を求めた．この結果，名詞あるいは未知語と判定された形態素と複合語を，検索課題文の特徴語と見なし，以下，**検索課題語**（topic terms）と呼ぶ．検索課題語の語数を $\#term$ とした☆☆．なお，4.3 節に述べる検索課題文の特徴語 tm は，上記の検索課題語を示す．

検索課題語に関する文書データベースならびに適合文書セット中の語頻度および文書頻度については，情報検索研究の成果の 1 つである TF-IDF 法[17]における発想を参考にした．ここで，TF-IDF 法では，特定の語に関する文書集合における出現頻度（以下，**語頻度**，term frequency）と特定の語を含む文書の出現頻度（以下，**文書頻度**，document frequency）が用いられ，これらを組み合わせることにより文書集合中の語の重み付けを実現する手法である．これを検索課題文の特徴量の計算に適用するが，詳細については，4.3 節，4.4 節および 4.5 節にて後述する．

4.3 検索課題文の特徴語に関する語頻度

検索課題 tp に対して，以下のように $tf\_rel(tp)$, $tf\_db(tp)$, $tf\_rat(tp)$ を定義した．ただし，TT は検索課題語集合，tm は TT の要素すなわち検索課題語である．REL，DB は，それぞれ適合文書セット，文書データベースを示す．また，$tf(tm, A)$ は，文書セット A における語 tm の出現頻度を示す．

$$tf\_rel(tp) = \frac{1}{|TT|} \sum_{tm \in TT} tf(tm, REL) \quad (1)$$

$$tf\_db(tp) = \frac{1}{|TT|} \sum_{tm \in TT} tf(tm, DB) \quad (2)$$

$$tf\_rat(tp) = \frac{1}{|TT|} \sum_{tm \in TT} \frac{tf(tm, REL)}{tf(tm, DB)} \quad (3)$$

適合文書セット中に検索語が出現するほど，あるいはそれが文書データベース中に出現しないほど，tf_rat は大きな値を持つ．

tf_rel，tf_db，tf_rat のそれぞれに関して，すべての検索課題にわたっての平均を求める．

4.4 検索課題文の特徴語に関する文書頻度

検索課題 tp に対して，以下のように $df\_rel(tp)$, $df\_db(tp)$, $df\_rat(tp)$ を定義した．ただし，文書セット A 中における語 tm を含む文書の出現頻度を $df(tm, A)$ で示す．

$$df\_rel(tp) = \frac{1}{|TT|} \sum_{tm \in TT} df(tm, REL) \quad (4)$$

$$df\_db(tp) = \frac{1}{|TT|} \sum_{tm \in TT} df(tm, DB) \quad (5)$$

$$df\_rat(tp) = \frac{1}{|TT|} \sum_{tm \in TT} \frac{df(tm, REL)}{df(tm, DB)} \quad (6)$$

df_rel，df_db，df_rat のそれぞれに関して，すべての検索課題にわたっての平均を求める．

4.5 TF-IDF

TF-IDF 法[17]における発想を検索課題文の特徴量の計算に適用した．4.3 節および 4.4 節で定義した特徴量を組み合わせて，以下のように $ltf\_db(tp)$, $idf\_db(tp)$ を定義した．

$$ltf\_db(tp) = \frac{1}{|TT|} \sum_{tm \in TT} ltf(tm, DB) \quad (7)$$

$$idf\_db(tp) = \frac{1}{|TT|} \sum_{tm \in TT} idf(tm, DB) \quad (8)$$

$$tfidf\_db(tp) = \frac{1}{|TT|} \sum_{tm \in TT} tf(tm, DB) \cdot idf(tm, DB) \quad (9)$$

☆ 日本語形態素解析には，『茶筅』[15]を利用した．

☆☆ ただし，後ほど式 (12) に示す idf などの計算のため，文書データベース中において検索課題語が出現する頻度が 0 である場合，その語あるいは複合語は語数に含めなかった．

表 4 検索課題難易度と検索課題の各種特徴量に関する Kendall の順位相関係数
Table 4 Kendall's rank correlation coefficients between the topic difficulty and feature quantities of the topics.

	diff	*func*	*|REL|*	*ave*	*stdev*	*med*	*skew*	*kurt*	*#term*	*#char*	*tf_rel*	*df_rel*	*tf_db*	*df_db*	*tf_rat*	*df_rat*	*tfidf_db*	*ltfidf_db*
diff		0.094	0.087	*-0.798*	*-0.688*	*-0.824*	*0.655*	0.227	-0.068	-0.014	-0.142	-0.063	*0.296*	*0.333*	*-0.421*	*-0.362*	*0.312*	*-0.291*
		0.443	0.424	0.000	0.000	0.000	0.000	0.035	0.548	0.902	0.189	0.562	0.006	0.002	0.000	0.001	0.004	0.007
func			-0.032	-0.090	-0.023	-0.114	0.110	0.006	0.029	-0.133	0.035	-0.024	0.015	-0.026	0.028	-0.002	0.081	-0.058
			0.771	0.401	0.829	0.287	0.307	0.954	0.795	0.224	0.746	0.822	0.888	0.810	0.802	0.987	0.449	0.589
|REL|				-0.119	-0.167	-0.119	0.064	-0.062	0.113	0.091	*0.639*	*0.780*	0.122	0.195	0.170	0.139	0.065	-0.116
				0.211	0.080	0.214	0.504	0.514	0.258	0.348	0.000	0.000	0.200	0.040	0.086	0.163	0.494	0.222
ave					*0.795*	*0.901*	*-0.592*	-0.181	0.060	0.004	0.110	0.047	-0.193	*-0.266*	*0.424*	*0.389*	-0.203	0.196
					0.000	0.000	0.000	0.055	0.541	0.969	0.247	0.618	0.041	0.005	0.000	0.000	0.032	0.038
stdev						*0.736*	*-0.430*	-0.182	0.045	-0.007	0.071	0.012	-0.147	*-0.246*	*0.336*	*0.330*	-0.134	0.221
						0.000	0.000	0.054	0.648	0.939	0.452	0.902	0.119	0.009	0.001	0.001	0.156	0.019
med							*-0.669*	-0.200	0.058	0.009	0.115	0.041	-0.201	*-0.274*	*0.389*	*0.358*	-0.214	0.221
							0.000	0.035	0.556	0.926	0.225	0.667	0.034	0.004	0.000	0.000	0.024	0.019
skew								*0.244*	-0.015	-0.019	-0.114	-0.050	0.174	0.186	*-0.339*	*-0.288*	0.199	-0.160
								0.010	0.883	0.847	0.228	0.597	0.066	0.050	0.001	0.003	0.036	0.091
kurt									-0.080	0.082	-0.127	-0.140	-0.126	-0.112	-0.116	-0.161	-0.052	0.094
									0.416	0.393	0.179	0.139	0.182	0.237	0.237	0.102	0.581	0.319
#term										*0.609*	0.121	0.125	0.171	0.148	-0.022	0.018	0.148	-0.093
										0.000	0.221	0.207	0.084	0.135	0.827	0.863	0.135	0.349
#char											0.012	0.042	-0.010	0.017	-0.061	-0.021	-0.005	-0.020
											0.902	0.667	0.920	0.860	0.545	0.834	0.957	0.835
tf_rel												*0.797*	0.160	0.140	*0.264*	0.196	0.144	-0.024
												0.000	0.090	0.139	0.007	0.047	0.127	0.800
df_rel													0.149	0.190	0.231	0.189	0.116	-0.075
													0.116	0.045	0.019	0.056	0.223	0.429
tf_db														*0.803*	-0.209	-0.152	*0.714*	*-0.338*
														0.000	0.034	0.123	0.000	0.000
df_db															-0.207	-0.158	*0.569*	*-0.417*
															0.035	0.109	0.000	0.000
tf_rat																*0.871*	*-0.274*	-0.033
																0.000	0.005	0.740
df_rat																	-0.195	-0.025
																	0.048	0.798
tfidf_db																		-0.232
																		0.014
ltfidf_db																		

各セルの上段：Kendall の τ，下段：両側有意水準 α，強調：相関係数が 1% 水準で有意（両側），下線：相関係数が 5% 水準で有意（両側）．

$$ltfidf\_db(tp) = \frac{1}{|TT|} \sum_{tm \in TT} ltf(tm, DB) \cdot idf(tm, DB) \quad (10)$$

ただし，

$$ltf(tm, A) = \log(tf(tm, A)) + 1.0 \quad (11)$$

$$idf(tm, A) = \log(N/df(tm, A)) \quad (12)$$

ltf_db，idf_db，$tfidf\_db$，$ltfidf\_db$ のそれぞれに関して，すべての検索課題にわたっての平均を求める．

5. 検索課題難易度と検索課題の各種特徴量に関する相関分析

実際の提出結果の各種統計値とそれに基づいた検索課題難易度，機能分類，その他検索課題の各種特徴量に関する相関性を分析した．本論文では，順位に基づく Kendall の相関係数 τ を用いた．通常，Pearson の相関係数が用いられることが多いが，本研究の目的に関しては 4 章で述べた各種特徴量の絶対値よりも，それらの相対的な順位関係の方がより重要と考え，Kendall の相関係数を用いることとした☆．算出された順位相関係数とその両側有意水準を**表 4** に示す．ただし，相関係数の算出の際に，*diff* については easy, middle，hard の順にそれぞれ 1，2，3 の値を割り当てた☆☆．*func* については 4.1 節でも述べたとおり，A, B, ···, F の順に有効な検索に必要とされる処理が多くなり，一般に困難とされる要素技術が加わるため，この順に検索の難易度が高くなるという考えのもと，それぞれ 1, 2, ···, 6 の値を割り当てた．表 4 から以下の事実が確認された．

(1) 提出結果リストにおける非補間平均精度の分布に関する歪度 *skew* および尖度 *kurt* は，ともに検索課題難易度 *diff* と明らかな正の相関があった．また，標準偏差 *stdev* は，検索課題難易度 *diff* と明らかな負の相関があった．このことから，検索課題の難易度が高くなるほど，提

☆ 他の分析手法の適用あるいは新たな分析手法の提案は今後の課題とする．たとえば，検索課題難易度レベルごとに各統計量に有意差があるかどうかについて分散分析を実施することなども検討に値する．

☆☆ ただし，本論文の分析で用いた Kendall の相関係数 τ は順位に基づくため，これら *diff* の数値そのものに意味はない．後述の *func* についても同様である．

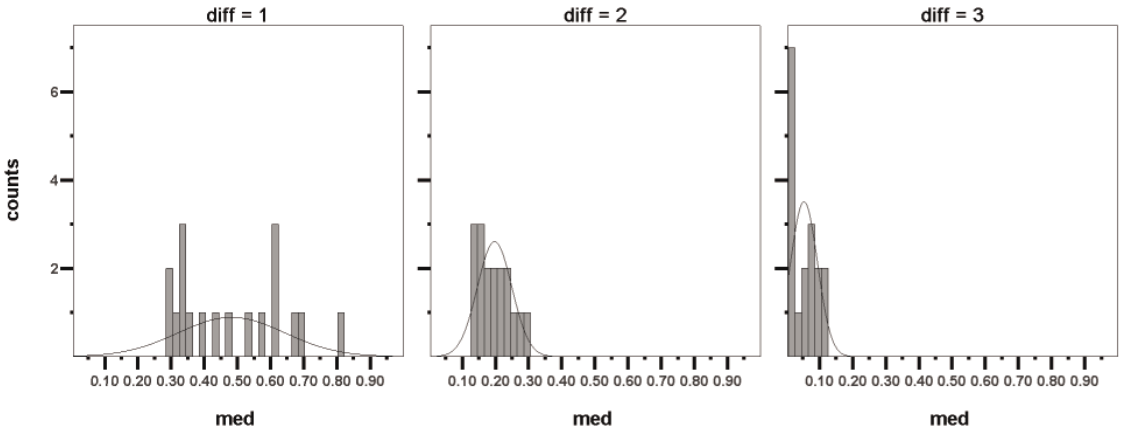

図 2 提出結果の非補間平均精度の中央値に関する検索課題難易度レベルごとのヒストグラム
Fig. 2 Histograms of medians of the non-interpolated average precision based on submitted results for three topic difficulty levels.

出結果の平均精度の分布は，低平均精度領域に偏るだけでなく，尖ったものになることが確認された．以上は**図 2** からも観察される．

(2) 検索課題語に関する文書データベース中の語頻度 tf_db と文書頻度 df_db との間で，順位相関係数が約 0.80 と大きく，統計的検定の結果からも明らかな正の相関があった．また，それぞれの変形である ltf_db と idf_db については明らかな負の相関があり，同じく相関の度合いは大きかった☆．したがって，tf_db と df_db（あるいは ltf_db と idf_db）は，統計的に互いに独立な特徴量であるとはいいがたい．以下，df_db をもって，これら検索課題語の文書データベースに対する頻度情報を代表する特徴量とした．ところで，df_db と検索課題難易度 *diff* とは，順位相関係数が約 0.33 とそれほど大きくないものの，統計的検定の結果から明らかな正の相関が認められた．このことは**図 3** からも観察される．このことは，文書データベース中に検索課題中の特徴語を含む文書が多いほど，検索が難しいことを示唆する．

(3) 検索課題語に関する適合文書セット中の語頻度 tf_rel と文書頻度 df_rel は，ともに提出結果に基づく検索課題難易度 *diff* とは明らかな相関性が認められなかった．一方，検索課題語に関する，適合文書セットと文書データベース中の語頻度の比率 tf_rat と文書頻度の比率 df_rat については，いずれも提出結果に基づく検索課題

☆ この結果は，紙面の都合により表 4 では省略した．

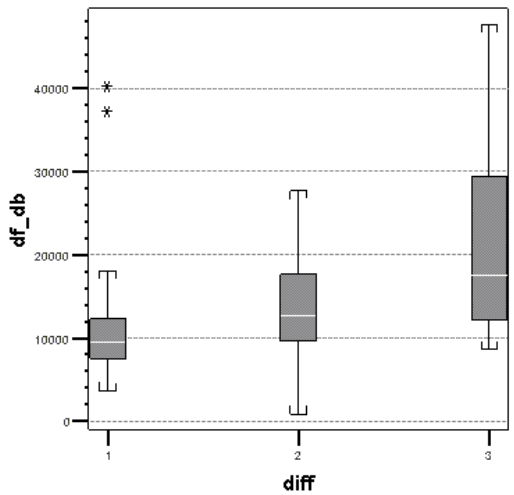

図 3 df_db と *diff* の関係を示す箱ヒゲ図
Fig. 3 Box-and-whisker graph presenting the relation between df_db and *diff*.

難易度 *diff* と明らかな相関性が認められた．両者はいずれも適合文書セットにおける頻度が高く，文書データベースにおける頻度が低ければ，大きい値をとる特徴量である．しかしながら，tf_rat あるいは df_rat を求めるには，適合文書セットが必要となるため，検索課題の難易度の分布を予測するという目的にはそぐわない．

(4) 提出結果に基づく検索課題難易度 *diff* と人間により判定された機能分類に基づく難易度の基準 $func$ とは明らかな相関性が見られなかった．さらに，各機能ごとに検索課題難易度 *diff* と他の各種特徴量の相関性を，各検索課題難易度レベルごとに機能分類 $func$ と他の各種特徴量の相

関性を分析したが，特に明らかな事実は確認できなかった．

(5) 検索課題語数 $\#term$，検索課題文の文字数 $\#char$，適合文書数 $|REL|$ は，いずれも検索課題難易度 $diff$ とは明らかな相関性が確認されなかった．

6. おわりに

テストコレクションの信頼性の観点から，検索課題の難易度が検索システムの相対的評価に与える影響を分析した．また，検索課題の難易度の予測可能性を吟味するため，NTCIR-1 における検索課題，文書データベースおよび適合文書セットに関する種々の特徴量を求め，それらの相関性に関する分析を行った．分析結果により，検索課題の難易度という観点で次に示すいくつかの事実が明らかになった．

- 提出結果に基づいて検索課題難易度を定義し，それらを 3 段階のレベルに分け，それぞれのレベルごとに非補間平均精度に基づいたシステムのランキングを行った．レベルごとの順位の相関を分析したところ，すべての組合せについて 0.7 から 0.9 程度の有意な相関が見られたことから，検索課題難易度が異なる場合でもシステム順位に有意な異なりは生じないことが示唆された．しかしながら，個々の順位を観察すると，無視できない順位の入れ替わりが見られ，検索課題難易度はシステムの相対的評価に一定の影響を与えうることが確認された．
- 検索課題難易度が高くなるほど，多様な情報検索手法による提出結果の非補間平均精度分布は，低平均精度領域に偏るだけでなく，尖ったものとなることが観察された．
- 提出結果に基づく検索課題難易度と，人間により判定された難易度と見なしうる機能分類とは，明らかな相関が見られなかった．
- 検索課題難易度と，検索課題文を構成する特徴語の文書データベースにおける頻度情報には，度合いは大きくはないものの明らかな相関が認められた．

実用的な観点から，検索課題の難易度あるいは検索課題セットの難易度の分布を予測するには，よりいっそうの検討が必要である．ところで，本論文では，複数の検索課題語について各々が文書データベース中に出現する頻度の平均をとった．しかしながら，すべての検索課題語が等しく検索に有効であるとは限らないことが，Kwok 氏により示されている[18]．これを鑑みて，何らかの基準に従って特に重要な検索課題語に着目し，頻度情報を算出することを考えている．また，本論文では，検索課題難易度と検索課題文の各種特徴量との相関性という観点から，検索課題難易度の予測可能性を吟味したが，各種特徴量の組合せ，他の分析手法の適用あるいは新たな分析手法の提案などが，今後の課題として検討に値する．

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) Voorhees, E. and Harman, D.K.: Overview of the Sixth Text REtrieval Conference (TREC-6), *Proc. 6th Text REtrieval Conference* (*TREC-6*), NIST Special Publication 500-240, pp.1–24 (1997).
2) 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).
3) National Institute of Informatics: NTCIR Project, ⟨http://research.nii.ac.jp/ntcir/⟩.
4) Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H. and Hidaka, S.: Overview of IR tasks at the First NTCIR Workshop, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.11–44 (1999).
5) Hawking, D., Voorhees, E., Craswell, N. and Bailey, P.: Overview of the TREC-8 Web Track, *Proc. 8th Text REtrieval Conference* (*TREC-8*), NIST Special Publication 500-246, pp.131–149 (2000).
6) Eguchi, K., Kuriyama, K. and Kando, N.: Analysis of the Topic Difficulty for NTCIR (NACSIS Test Collection for Information Retrieval Systems), *Proc. 3rd International Conference of Asian Digital Library* (*ICADL 2000*), pp.231–238 (2000).
7) 江口浩二，栗山和子，神門典子：テストコレクションにおける検索課題の難易度予測への挑戦，情報処理学会研究報告，No.2001-FI-63, pp.17–24 (2001).
8) 栗山和子，神門典子：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，情報処理学会研究報告，No.99-FI-55, pp.41–

48 (1999).
9) Kageura, K., Yoshioka, M., Takeuchi, K., Koyama, T., Tsuji, K., Yoshikane, F. and Okada, M.: Overview of TMREC Tasks, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.415–416 (1999).
10) Eguchi, K., Kando, N. and Adachi, J.(Eds.): *Proc. 2nd NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, National Institute of Informatics (2001). ISBN: 4-924600-96-2.
11) National Institute of Informatics: NACSIS-IR, ⟨http://www.nii.ac.jp/ir/ir-e.html⟩.
12) Kuriyama, K., Kando, N., Nozue, T. and Eguchi, K.: Pooling for a Large-Scale Test Collection : An Analysis of the Search Results from the First NTCIR Workshop, *Information Retrieval*, Vol.5, No.1, pp.41–59 (2002).
13) Voorhees, E.M.: Evaluation by Highly Relevant Documents, *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.74–82 (2001).
14) 岸田和明：検索実験における評価指標としてのMean Average Precision の性質，情報処理学会研究報告，No.2001-FI-63, pp.97–104 (2001).
15) 松本裕治，北内 啓，山下達雄，今一 修，今村友明：日本語形態素解析システム『茶筅』version 1.5 使用説明書 (1997).
16) Kando, N., Kageura, K., Yoshioka, M. and Oyama, K.: Phrase Processing Methods for Japanese Text Retrieval, *SIGIR Forum*, Vol.32, No.2, pp.23–28 (1998).
17) Salton, G.: *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley (1989).
18) Kwok, K.L.: A New Method of Weighting Query Terms for Ad-hoc Retrieval, *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.187–195 (1996).

(平成 13 年 9 月 25 日受付)
(平成 13 年 11 月 26 日採録)

（担当編集委員 原田 隆史）

江口 浩二（正会員）

1993 年同志社大学工学部電子工学科卒業．1999 年関西大学大学院工学研究科博士課程修了．博士（工学）．同年学術情報センター助手．2000 年国立情報学研究所助手，現在に至る．情報検索，Web 情報管理の研究に従事．電子情報通信学会，ACM 各会員．

栗山 和子（正会員）

1993 年図書館情報大学大学院図書館情報学研究科修了．1996 年筑波大学大学院工学研究科博士課程修了．博士（工学）．同年，同大学準研究員．1998 年学術情報センター（現，国立情報学研究所）リサーチ・アソシエイト．2001 年国立情報学研究所 COE 研究員，現在に至る．数式処理，情報検索の研究に従事．日本数式処理学会，日本応用数理学会，ACM（SIGSAM，SIGIR）各会員．

神門 典子（正会員）

1994 年慶應義塾大学文学研究科博士課程修了．博士（図書館・情報学）．同年学術情報センター助手．1995 年米国シラキウス大学情報学部客員研究員，1996〜1997 年デンマーク王立図書館情報大学客員研究員．1998 年学術情報センター助教授．2000 年国立情報学研究所助教授，現在に至る．テキスト構造を用いた検索と情報活用支援，言語横断検索，情報検索システムの評価等の研究に従事．ACM-SIGIR，BCS-IRSG，ASIS&T，言語処理学会，日本図書館情報学会各会員．