

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

江口 浩二[†] 栗山 和子[†] 神門 典子[†]

[†]国立情報学研究所

〒101-8430 東京都千代田区一ツ橋 2-1-2

e-mail: {eguchi,kuriyama,kando}@nii.ac.jp

あらまし 本論文では、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一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なう。第一に、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。第二に、検索課題難易度の予測可能性を検討する。そのため、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 を対象に行なった検討結果を報告する。

Challenge to Predict Topic Difficulty in Test Collections

Koji EGUCHI[†] Kazuko KURIYAMA[†] Noriko KANDO[†]

[†]National Institute of Informatics

2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

e-mail: {eguchi,kuriyama,kando}@nii.ac.jp

Abstract 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 terms frequencies in the document databases or human judgments, and analyzes the correlation between the topic difficulty and them. This paper reports the results of the investigations using the NTCIR-1 test collection.

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上で評価することにより、相互の特徴比較を行な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正解判定リス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] があったが、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、成功を治めている。

本報告では、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 [5, 6, 7]。さて、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」が挙げられる。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を旨とし、検索課題の作成を各領域の専門家（大学院生以上の研究者）から収集した。また、検索課題が易し過ぎるものや難し過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、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。それを実現するには、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、検索実行前に予測することができることが求められる。TREC-6 [1] 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が、その結果、人間の判断による検索課題の難易度の分類と検索結果の評価による数値的な難易度に、相関があるとは言えないことが報告されている。

本研究では、より多様な観点から、検索課題の難易度に関する分析を NTCIR-1 を対象にして実施した。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した。その結果、検索課題の難易度は、システム順位に一定の影響を与えることが確認された。また、NTCIR-1 における検索課題、文書データベース、および正解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した。その結果、検索課題の難易度の予測は容易でないものの、文書データベースにおける検索課題文中の語の頻度情報が、検索課題の難易度と有意な相関性を示すことなどが確認された。

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 を構成する (1) 文書データベース、(2) 検索課題、(3) 正解判定リストについて概要を述べる¹ [4]。

¹ テストコレクション NTCIR-1 には、(1),(2),(3) に加えて、タグ付きコーパスは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 [8] が含まれているが、本論文では分析の対象としない。なお、テストコレクション NTCIR-2 については、本論文では分析に至っていないが、概要については論文集 [9] を参照されたい。

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、全国大会や研究会等の発表論文の要旨を集めた学会発表データベース [10] 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る [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。

2.2 検索課題

検索課題 (search topics) 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り、訓練用 30 課題、評価用 53 課題を作成した。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。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)、概念リスト (concept) から成る。検索要求は、利用者の検索要求を記述した自然言語文であり、検索要求説明は、背景説明・検索の目的・正解判定基準・用語の定義などを含み、検索要求の背景情報を提供する。タイトルは、検索課題を簡潔に表現したものであり、概念リストは検索課題を表す重要な概念、キーワードのリストである。検索実験では以上のいずれの項目を使用しても良く、検索課題の記述を参照しながら対話的にクエリの作成と入力を行なっても良いが、結果提出に際しては検索課題のどの項目を使用したか、対話型検索であるかどうかを報告する必要がある。図 1 に検索課題の例を示す。

2.3 正解判定

日本語の検索要求に対して日本語および英語の正解文書を検索する「随時検索」と日本語の検索要求に対して英語の正解文書を検索する「言語横断検索」の二つのタスクに対して、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セット（以下、提出結果と呼ぶ）の上位一定数の和集合に加えて、再現率重視の対話型検索の結果に対して正解判定を実施することで、網羅的な正解文書セットを収集する [11]。正解判定に際しては、二名のクロスチェックに基づく最終判定が行なわれた。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。

```

<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。電子図書館もその一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なわれてきている。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い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

3 検索課題難易度がシステム順位に与える影響

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、3.1 に後述する通り、提出結果ごとの検索有効性 (retrieval effectiveness) 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レベル分けした。このとき「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から提出された検索結果に基づき、評価用検索課題²を対象としてレベル分けを実施した。対話型検索か非対話型検索か、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セットに関して、次の各種

² 訓練用検索課題の予備的分析については [7] を参照されたい。

統計値を求めた。

- 正解と判定された文書の総数 ($|REL|$) ,
- 提出結果セッ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、更に検索課題を三つのレベルに等分割した。各レベルは、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に対するシステムランキング比較

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得る。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。本論文では、システムランキング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、検索課題難易度が引き起こすシステムランキングの異なりについて分析する。

3.1 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ランキングについて調べる。3.1 で述べた、26 のシステムのランキング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。また、表 1 には、ランキングにおいて一位順位が上がるに応じた、非補間平均精度の増加率を併記する。

検索課題難易度ごとのシステムランキングについて順位相関係数 Kendall の τ と有意水準 α を算出し、表 2 に示す。結果として、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ランキング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 からわかる通り、検索課題難易度ごとの各システムランキングの上位において順位の入替わりが観察されたが、経験的に有意であるとみなされている増加率 5% を越えて順位が入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価に一定の影響を与えるとみなすことができる。

表 1: 検索課題レベルごとの非補間平均精度に基づくシステムの順位

rank	easy			middle			hard			all		
	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

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

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) .

4 検索課題の特徴量

本節では, 検索課題難易度の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 [2] の機能分類に準拠し, 以下の 6 種の機能を設定した . ただし, 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 [7] .

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受け関係を判断する (構文解析を行なう) .

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を利用する . 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 .

2 名の図書館情報学を専攻する大学院生により, NTCIR-1 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「o」あるいは「x」で, 例えば, $(F0, F1, F2, F3, F4, F5) = 000000$ のように表現し, 機能のパターンによって検索課題を図 2 のような六つのカテゴリに分類した . このとき, A, B, C, D, E, F の順に必要とされる処理が多くなることから, 一般的にはその順に検索要求に適切な検索の実行が困難であると考えられることができる . 従って, 人間により判定された難易度に関する一指標になり得る .

4.2 検索課題文の特徴量

検索課題の特徴を示すであろう, 以下の検索課題文の各種特徴量に着目する .

- 検索課題文の特徴語数 ($\#term$), 文字数 ($\#char$),

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:	ooxoox
D. シソーラス機能と構文解析機能:	ooxoox
E. シソーラス機能と内容解析機能:	ooxoox
F. シソーラス機能と構文解析機能と内容解析機能:	ooxoox

図 2: 機能分類に基づく検索課題の分類結果

- 検索課題文の特徴語に関する語頻度,
- 検索課題文の特徴語に関する文書頻度.

なお, 検索課題文に対して形態素解析³を実行して求めた形態素群に対して, いくつかの接続ルール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みなし, 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⁴. なお, 4.3 に述べる検索課題文の特徴語 tm は, 前述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・正解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一つである TF-IDF 法 [13] 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 (term frequency: tf , 以下, 語頻度) と特定の語を含む文書の出現頻度 (document frequency: df , 以下, 文書頻度) 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, 4.4, 4.5 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, それぞれ正解文書セット, 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.

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

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

³ 日本語形態素解析には『茶筌』 [12] を利用した.

⁴ ただし, 後ほど式 (12) に示す idf 等の計算のため, 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.

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

正解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.

全ての検索課題に対して, tf_{rel} , tf_{db} , tf_{rat} の各総和平均を求める.

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.

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

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

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

全ての検索課題に対して, df_{rel} , df_{db} , df_{rat} の各総和平均を求める.

4.5 TF-IDF

TF-IDF 法 [13] 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, 4.4 で定義した特徴量を組み合わせ, 以下のように $ltf_{db}(tp)$, $idf_{db}(tp)$ を定義した.

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

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

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

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

ただし,

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

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

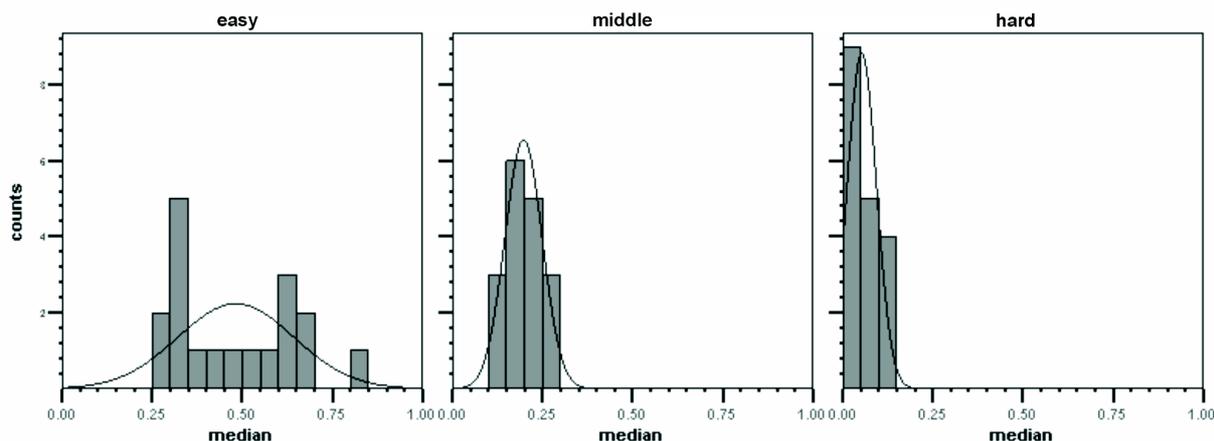


図 3: 検索課題レベルに対する提出結果の非補間平均精度の中央値のヒストグラム

全ての検索課題に対して, ltf_db , idf_db , $tfidf_db$, $lftfidf_db$ の各総和平均を求める.

5 検索課題難易度と各種検索課題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. 相関係数として順位に基づく Kendall の τ を用いた. 本研究の目的に関しては上記の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, より重要と考えるからである. 順位相関係数とその両側有意水準を表 3 に示す. 表 3 から以下の事実が観察された.

- (1) 提出結果セッ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, とともに検索課題の難易度 $diff$ と明らかな正の相関があった. また, 標準偏差 $stdev$ は, 検索課題の難易度と明らかな負の相関があった. 以上は図 3 から観察される. このことから, 検索課題の難易度が高くなるほど, 検索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.
- (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, 同じく相関の度合は大きかった⁵. 従って, tf_db

⁵この結果は, 紙面の都合により表 3 には省略した.

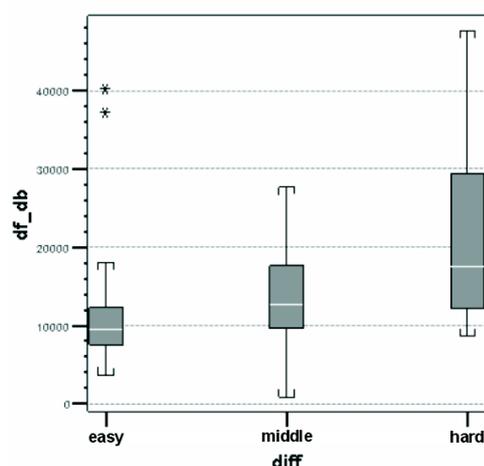


図 4: df_db と $diff$ の関係を示す箱ヒゲ図

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たがたい. 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. ここで, 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 4 から観察される. 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正解文書セット中の語頻度 tf_rel と文書頻度 df_rel は, とともに検索結果の難易度 $diff$ 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, 正解文書セットと文書データ

表 3: 検索課題難易度と各種検索課題特徴量に関する Kendall の順位相関係数

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

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) .

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, いずれも検索結果の難易度と明らかな相関性が認められた. 両者はいずれも正解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, 正解文書セットを必要とするため, 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) 提出結果に基づく実際の検索課題難易度 $diff$ 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. また, 各機能ごとの検索課題難易度 $diff$ と他の各種特徴量の相関性, 各検索課題難易度のレベルごとの機能分類 $func$ 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 $\#char$, 正解文書数 $|REL|$ 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を目的として, NTCIR-1 における検索課題, 文書データベース, および正解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なった. 分析結果により, 検索課題の難易度という観点からいくつかの事実が明らかになった.

- 検索課題難易度を 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順位づけを行なった. レベルごとにランキング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ランキング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みられ、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得ることが確認された。

- 検索課題の難易度が高くなるほど、多様な情報検索手法による検索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。
- 提出結果に基づく実際の難易度と、人間により判定された難易度と見なせる機能分類とは、明らかな相関が見られなかった。
- 検索課題の難易度と、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、より一層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により示されている [14]。従って、何らかの基準に基づいて特に重要と見なされる検索課題語に着目して、頻度情報を算出することが検討に値する。

参考文献

- [1] E. Voorhees and D. K. Harman: "Overview of the sixth Text REtrieval Conference (TREC-6)", Proceedings of the 6th Text REtrieval Conference (TREC-6), NIST Special Publication 500-240, pp. 1-24 (1997).
- [2] 木谷, 小川, 石川, 木本, 中渡瀬, 芥子, 豊浦, 福島, 松井, 上田, 酒井, 徳永, 鶴岡, 安形: "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, DBS114-1, pp. 15-22 (1998).
- [3] "NTCIR Project", (<http://research.nii.ac.jp/ntcir/>).
- [4] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato and S. Hidaka: "Overview of IR tasks at the first NTCIR workshop", Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Tokyo, Japan, pp. 11-44 (1999).
- [5] K. Eguchi, K. Kuriyama and N. Kando: "Analysis of the topic difficulty for NTCIR (NACSIS test collection for information retrieval systems)", Proceedings of the 3rd International Conference of Asian Digital Library (ICADL 2000), pp. 231-238 (2000).
- [6] 江口, 栗山, 神門: "大規模テストコレクション NTCIR-1 における検索課題の難易度に関する分析", 情報処理学会研究報告, No. 2000-FI-59(2000-DD-24), pp. 25-32 (2000).
- [7] 栗山, 神門: "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).
- [8] K. Kageura, M. Yoshioka, K. Takeuchi, T. Koyama, K. Tsuji, F. Yoshikane and M. Okada: "Overview of TMREC tasks", Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Tokyo, Japan, pp. 415-416 (1999).
- [9] K. Eguchi, N. Kando and J. Adachi Eds.: "Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization", National Institute of Informatics (2001).
- [10] "NACSIS-IR", (<http://www.nii.ac.jp/ir/ir-e.html>).
- [11] 栗山, 神門, 野末, 江口: "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 の分析", 学術情報センター紀要, 12, pp. 17-30 (2000).
- [12] 松本, 北内, 山下, 今一, 今村: "日本語形態素解析システム『茶筌』version 1.0 使用説明書" (1997).
- [13] G. Salton: "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer", Addison-Wesley (1989).
- [14] K. L. Kwok: "A new method of weighting query terms for ad-hoc retrieval", Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp. 187-195 (1996).