4P－2

大規模テストコレクション NTCIR-1 の構築 (2)
－ 検索課題の分析 －

神門典子　栗山和子　江口浩二　野末俊比古
学術情報センター 研究開発部

1　はじめに

本発表では、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。「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や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。

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（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した。本発表では、検索課題の難易度について、NTCIR-1 の評価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評価テストでの評価結果から考察する。

2　検索課題の分析

2.1　評価テストの概要

NTCIR ワークショップ [3] では、1999 年 3 月 4 日に評価テストを行ない、評価用検索課題 53 件について、24 チームで合計 121 セットの検索結果が提出された。本研究では、このうち、随時検索タスクの提出結果 18 チーム 47 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、53 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。

2.2　精度から見た検索課題の難易度

表 2-1. 検索課題の統計的データと難易度

topic	rel	ave	stdev	median	difficulty
0031	21	0.2312	0.1596	0.2099	middle
0032	23	0.0383	0.0419	0.0267	hard
0033	162	0.2633	0.2250	0.1757	middle
0034	15	0.1920	0.1646	0.1660	hard
0035	32	0.3054	0.1811	0.3529	easy
0036	14	0.5085	0.2930	0.6583	easy
0037	65	0.0767	0.0659	0.0681	hard
0038	39	0.2361	0.1681	0.2373	middle
0039	16	0.2746	0.1252	0.3028	middle
0040	47	0.2730	0.1652	0.2703	middle
0041	16	0.2487	0.1428	0.2430	middle
0042	22	0.1002	0.0923	0.0794	hard
0043	35	0.5644	0.2769	0.6687	easy
0044	15	0.2202	0.1536	0.2461	middle
0045	18	0.0675	0.1253	0.0206	hard
0046	37	0.2338	0.1703	0.1653	hard
0047	30	0.1076	0.1349	0.0543	hard
0048	34	0.0932	0.1702	0.0147	hard
0049	20	0.0658	0.0755	0.0409	hard
0050	37	0.3805	0.2420	0.3716	easy
0051	20	0.0419	0.0390	0.0410	hard
0052	9	0.3045	0.2543	0.2850	middle
0053	84	0.1769	0.1229	0.1717	middle
0054	584	0.0738	0.0947	0.0555	hard
0055	40	0.2752	0.1377	0.3226	easy
0056	68	0.1619	0.1026	0.1776	middle
0057	187	0.3241	0.1807	0.3770	easy
0058	10	0.5990	0.3332	0.7570	easy
0059	61	0.3262	0.1862	0.3300	easy
0060	10	0.3351	0.2202	0.3039	easy
0061	24	0.2309	0.1592	0.2343	middle
0062	22	0.1489	0.0909	0.1356	hard
0063	43	0.1659	0.1386	0.1280	hard
0064	59	0.4037	0.1897	0.4196	easy
0065	10	0.4654	0.2943	0.5412	easy
0066	33	0.5188	0.3063	0.6047	easy
0067	23	0.4332	0.2229	0.5193	easy
0068	52	0.2084	0.1562	0.1686	middle
0069	12	0.0685	0.0969	0.0218	hard
0070	111	0.3275	0.1730	0.3105	easy
0071	13	0.1688	0.1040	0.1660	hard
0072	21	0.1836	0.1152	0.1823	middle
0073	11	0.1753	0.1648	0.1738	middle
0074	17	0.1809	0.1379	0.1671	middle
0075	14	0.1822	0.1530	0.1724	middle
0076	17	0.2189	0.1565	0.1827	middle
0077	6	0.3756	0.2176	0.3575	easy
0078	6	0.6318	0.3326	0.6810	easy
0079	10	0.5907	0.2831	0.6559	easy
0080	16	0.2011	0.1837	0.1339	hard
0081	9	0.1468	0.1927	0.0698	hard
0082	9	0.4754	0.2441	0.5695	easy
0083	36	0.1572	0.2025	0.0378	hard

各検索課題の正解個数、評価テストの随時検索タスク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、難易度を表 2-1 に示す。「正解」は「正解」と「部分的正解」を合わせて「正解」とする。本発表では、A と B の両方を正解とした場合の評価を用いる。検索課題の難易度は、平均精度の中央値 (median) で、easy:「易しい」、middle:「中位」hard:「難しい」として、グループ分けした。

NTCIR-1:NACSIS Test Collection for Information Retrieval systems-1 (2) – Analysis of the Search Topics –
Noriko Kando, Kazuko Kuriyamal, Koji Eguchi, Toshihiko Nozue
R & D Dept., National Center for Sciecnce Information Systems (NACSIS)

3 分析結果

以下に、各検索課題についての、難易度、機能分類、検索要求文中の単語数、検索要求文の文字数、tfTF、dfDF（後述）を表として示す。

表 3-1. 検索課題の難易度と分析結果

topic	difficulty	func	word	char	tfTF	dfDF
0031	middle	D	12	43	1.78	1.77
0032	hard	D	6	37	0.60	0.45
0033	middle	F	9	39	12.41	11.80
0034	hard	F	10	32	0.38	0.27
0035	easy	D	7	25	6.63	4.76
0036	easy	G	7	27	3.55	2.92
0037	hard	D	6	32	1.59	0.86
0038	middle	D	10	32	4.87	5.48
0039	middle	D	6	26	12.35	12.74
0040	middle	B	8	27	2.98	2.68
0041	middle	F	9	26	1.70	1.23
0042	hard	H	12	34	8.47	100.08
0043	easy	D	8	31	6.24	4.55
0044	middle	H	6	26	0.11	0.09
0045	hard	F	8	25	0.06	0.04
0046	hard	F	8	28	6.16	6.16
0047	hard	F	16	49	0.88	0.88
0048	hard	F	6	19	0.76	0.55
0049	hard	F	6	22	0.68	0.44
0050	easy	F	5	15	13.44	11.31
0051	hard	B	8	41	0.74	0.71
0052	middle	D	2	20	21.67	33.45
0053	middle	F	8	32	1.72	1.46
0054	hard	I	6	20	6.78	6.71
0055	easy	F	17	72	4.17	3.45
0056	middle	D	16	60	4.03	3.16
0057	easy	D	12	53	16.49	15.30
0058	easy	F	5	32	18.82	16.91
0059	easy	F	7	37	4.32	2.53
0060	easy	G	9	27	10.77	8.36
0061	middle	F	12	35	3.18	3.39
0062	hard	G	5	26	4.88	2.94
0063	hard	H	12	30	10.31	9.51
0064	easy	F	8	30	7.91	5.29
0065	easy	F	8	26	0.77	0.39
0066	easy	D	12	32	5.03	4.08
0067	easy	D	13	51	3.53	2.57
0068	middle	D	11	41	8.23	9.87
0069	hard	D	7	32	0.15	0.25
0070	easy	D	7	19	4.74	3.62
0071	hard	D	10	31	3.11	2.51
0072	middle	D	7	30	1.57	0.94
0073	middle	F	9	31	1.39	1.62
0074	middle	D	12	46	2.45	2.09
0075	middle	F	10	37	1.54	1.13
0076	middle	D	7	22	1.66	2.11
0077	easy	D	6	27	2.29	2.77
0078	easy	F	14	44	5.43	3.76
0079	easy	F	11	37	168.14	8.80
0080	hard	D	11	43	0.54	2.95
0081	hard	F	11	37	0.57	2.04
0082	easy	D	5	20	18.96	12.89
0083	hard	F	11	38	3.76	2.33

以下では、表 3-1 の各要素について説明する。

3.1 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、6 つの機能 F0～F5 を設定した [2]。検索課題中の検索要求文に含まれる語句を用いて各機能が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。判定のパターンによって検索課題を A～I という 9 グループに分類した。A:F0, B:F0+F1, C:F0+F3, D:F0+F1+F3, E:F0+F1+F4, F:F0+F1+F3+F4, G:F0+F1+F3+F5, H:F0+F1+F3+F4+F5, I:F0+F1+F2+F3+F5.

表 3-1、および機能分類 func と平均精度の中央値との相関係数から、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かった。

3.2 正解文書中への単語の出現

検索課題中の検索要求文を単語・フレーズに分割し、単語・フレーズ数 word と平均精度の中央値との相関係数を計算してみたところ、特に相関はないことがわかった。

次に、tf:各単語が含まれる正解文書数、df:各単語の正解文書中での出現頻度、DF:各単語が含まれる DB 中の文書数、TF:各単語の DB 中での出現頻度、tfTF:単語ごとの tfTF(%) の検索課題ごとの平均、dfDF:単語ごとの df/DF(%) の検索課題ごとの平均として、tfTF、dfDF のそれぞれと平均精度の中央値との相関係数と計算すると、相関があることがわかった。

4 まとめ

分析の結果、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、検索要求文中の語の含まれる正解文書数とその中での出現頻度の DB 中での割合は、平均精度の中央値、すなわち、難易度と相関があることがわかった。

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] 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”. 99-FI-55-6, pp.41-48, 1999.

[3] Proceedings of the 1st NTCIR Workshop on Research and Development in Japanese Text Retrieval, Tokyo, Aug.30-Sep.1, 1999. (to appear)