Sensitivity of IR systems Evaluation to Topic Difficulty

Koji Eguchi †, **Kazuko Kuriyama** ‡, **Noriko Kando** †

† National Institute of Informatics (NII)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
{eguchi, kando}@nii.ac.jp

‡ Shirayuri College
1-25 Midorigaoka, Chofu-shi, Tokyo 182-8525, Japan
kuriyama@nii.ac.jp

Abstract

The difficulty of the topics or queries is one of important factors in evaluating information retrieval (IR) systems. This paper analyzes the differences of system ranking affected by the topic difficulty using a test collection 'NTCIR-1,' which is constructed for evaluating Japanese IR systems and composed of (1) the topics, (2) the document database, and (3) the lists of relevant judgments. Furthermore, this paper defines measures for the various features on the topics, and analyzes the correlation between them, in order to investigate the predictability of the topic difficulty.

1. Introduction

The 'NTCIR' workshop is one of the evaluation workshops for information retrieval (IR) systems and compares the retrieval effectiveness of each system using a common test collection (National Institute of Informatics, 2001). Here, a test collection means a benchmark set for the experiments on IR systems, and is composed of (1) the document database, (2) the topics, and (3) the list of relevant judgments for each topic.

The topics are to be balanced for topic difficulty. The topic difficulty distribution should not be biased, as it is neither too difficult nor too easy, in order to keep reliability of test collections. This paper investigates the properties of topics in a test collection and the sensitivity of IR systems evaluation to the topic difficulty. We define topic difficulty using a test collection 'NTCIR-1' (Kando et al., 1999), and analyses the differences of system ranking caused by the differences of the topic difficulty.

By the way, test collections are required, from the point of view of reliability, to predict topic difficulty or its distribution for a set of topics. TREC-6 investigated the predictability of the topic difficulty, and reported that the topic categorization by humans from the point of view of topic difficulty does not correlate well with computational difficulty on the basis of the evaluation of search results (Voorhees and Harman, 1997). This paper preliminarily investigate the predictability of the topic difficulty from another point of view. We measures the various feature quantities on the topics, and analyzes the correlation between these and the topic difficulty.

2. Defining Topic Difficulty and Measuring Topic Features

2.1. Definition of Actual Topic Difficulty

In order to identify the actual topic difficulty we categorized the topics on the basis of the median of non-interpolated average precision; this means retrieval effectiveness, and indicates the actual topic difficulty of the submitted result set for each topic.

We analyzed the test topics using only the 26 submitted result set of non-interactive searches, which were done, in the ad hoc IR task, by automatic query construction and by using only the descriptions of the topics[1]. This was to try to avoid the different retrieval effectiveness distributions that result when the query uses an interactive search or a non-interactive search using different parts of the topic. Then, we categorize the topics, ranked according to the medians in ascending order, into the following three categories: "hard," "middle" and "easy," so that each of the categories contains the same number of topics. This categorization is, in this paper, referred to as *actual topic difficulty* and indicated as $diff$.

2.2. Function-based Topic Categorization

We use the function-based topic categorization as one measure of the human-judged topic difficulty.

According to the function-based topic categorization in BMIR-J2 (Sakai et al., 1999), we define the following functions. Here, "the basic function" achieved by keyword search or query expansion using thesauri, as defined in BMIR-J2, is divided into "F0. basic function" and "F1. thesaurus function".

F0. basic function: The relevant texts can be retrieved by simply using words extracted from the query, or by their Boolean expressions.

F1. thesaurus function: The relevant texts can be retrieved using words extracted from the query and their related words expanded by thesauri, or by satisfying their Boolean expression.

F2. numerical range function: The system needs to handle a numeric range description.

[1] While the analysis of actual topic difficulty in this paper, we used only "relevant" judgments but not "partially relevant."

Table 1: The results of the function-based topic categorization.

the function-based topic categorization	F0	F1	F2	F3	F4	F5
A. basic function	1	0	0	0	0	0
B. thesaurus function	1	1	0	0	0	0
C. syntactic function	1	0	0	1	0	0
D. thesaurus function & syntactic function	1	1	0	1	0	0
E. thesaurus function & semantic function	1	1	0	0	1	0
F. thesaurus function, syntactic function	1	1	0	1	1	0

F3. syntactic function: Analysing a syntactic relationship among query words helps the query be understood.

F4. semantic function: A semantic / context analysis is required to understand the query.

F5. the world knowledge function: Common sense / world knowledge is required to process the query. Such information is often missing in the text or the system's lexicon.

Assessors, two graduate school students, performed the judgments for the function-based topic categorization. As the results of the judgments, the existence of each function is represented as "1" in Table 1.

In general, the effective search execution tends to be difficultin the order A, B, C, D, E and F, since the processing required increases in the same order. Thus, the function-based topic categorization can be used as one measure of the human-judged topic difficulty.

2.3. Feature Quantities of the Topics

We consider the nouns and compound words in the topic to be the *topic terms* that typically represent each topic. In order to extract topic terms, we perform a Japanese morphological analysis for each topic, and then obtain the compound words by applying some rules to the morpheme set [2]. Using them, we definethe following features of the topics[3]:

- the frequencies of the topic terms in the document database,
- the frequencies of the documents that include the topic terms in the document database.

The aforementioned term frequencies, the document frequencies and the combination between them are discussed in Subsecs. 2.3.1., 2.3.2. and 2.3.3.

2.3.1. Term Frequencies of The Topic Terms

We definethe following $tf\_db(tp)$ for the topic tp. Here, TT indicates the topic term set which constitute the topic, and tm indicates the topic term that is an element of TT. DB indicates the document database, while $tf(tm, A)$ indicates the frequency of a term tm in the document set A.

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

2.3.2. Document Frequencies of the Topic Terms

We definethe following $df\_db(tp)$ for the topic tp. Here, $df(tm, A)$ indicates the frequency of a document that includes a term tm in the document set A.

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

2.3.3. TF-IDF

We apply the TF-IDF weighting method, which is a well-known result in IR research areas, to compute frequency-based feature quantities of the topics. We definethe following $tfidf\_db(tp)$ and $ltfidf\_db(tp)$. The former is a naive TF-IDF method (Salton, 1989) and the latter is often used in IR systems based on a vector space model (Buckley et al., 1993).

$$tfidf\_db(tp) = \frac{1}{|TT|} \sum_{tm \in TT} tf(tm, DB) \cdot idf(tm, DB), \quad (3)$$

$$ltfidf\_db(tp) = \frac{1}{|TT|} \sum_{tm \in TT} ltf(tm, DB) \cdot idf(tm, DB), \quad (4)$$

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

$$idf\_db(tp) = \frac{1}{|TT|} \sum_{tm \in TT} idf(tm, DB). \quad (6)$$

where

$$ltf(tm, A) = \log(tf(tm, A)) + 1.0, \quad (7)$$

$$idf(tm, A) = \log(N/df(tm, A)). \quad (8)$$

3. Results

3.1. System Ranking Comparison for Different Topic Difficulty Levels

We analyze the differences of system ranking caused by the topic difficulty, since we consider that the system ranking might be affected by the topic difficulty, for example,

[2] We made use of *ChaSen* (Matsumoto et al., 1997) for the Japanese morphological analysis, and the method used in (Kando et al., 1998) for obtaining compound words.

[3] We considered those compound words, nouns or unknown words, whose frequencies in the document database were 0, should not be included into the topic terms because of the computation of idf in Eq.(8). The topic terms will be represented as tm in Subsec. 2.3.1.

Table 2: 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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

one system is good at retrieval for average levels of topic difficulty but bad at retrieval for difficult topics, and another is conversely good at retrieval for a kind of difficult topics.

At first, we investigate system ranking of top runs for three topic difficulty levels $diff$, whose values are "hard," "middle" or "easy" as defined in Subsec. 2.1. Here, the number of system search results is 26, as described in the beginning Section 2.1. The results are presented as Table 2, where ave indicates the mean average precision for each run. Then, we obtained the respective average values of non-interpolated average precision for three topic difficulty levels, and analyzed the rank correlation between them. The resulting Kendall's τ and observed significance level α are presented in Table 3.

Table 3 explains that the differences of system ranking is not significant in spite of the differences of topic difficulty. However, we can find through Table 2 the cases when respective system ranks of top runs are changed in details according to topic difficulty levels, where the improving rates of mean average precision reach 5%, the statistically significant level (Kishida, 2001). This suggests that the topic difficulty can affect the system ranks to some extent.

Table 3: Kendall's rank correlation coefficients between system ranking of submitted results 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's rank correlation coefficient;
α: observed two-sided significance level;
emphasized: correlation coefficients which are significant within 1% two-sided significance level.

3.2. Preliminary Investigatation on Topic Predictability

We investigate the correlation between the actual topic difficulty based on submitted result set, the human-judged topic difficulty based on function-based topic categorization and the other features of the topics. We made use of Kendall's rank correlation coefficients τ to measure the correlations, because we consider that the comparative measures are more important than the absolute ones in this case of correlation analysis.

For computing the correlation, the actual topic difficulty $diff$ has one of the values 1, 2 or 3 respectively corresponded to "easy," "middle" or "hard," and the human-judged difficulty $func$ has one of the values 1, 2, $\cdots$, 6 respectively corresponded to the A, B, $\cdots$, F, shown in Table 1. The values of Kendall's τ and their two-sided significance level are indicated in Table 4, where ave, $stdev$, med, $skew$ and $kurt$ respectively indicate the average, standard derivation, median, skewness and kurtosis of the distribution of non-interpolated average precision using the submitted result set; $\#rel$ and $\#term$ respectively indicate the average number of relevant documents and topic terms. The following evidence can be found through Table 4.

(1) Both of the skewness ($skew$) and kurtosis ($kurt$) of the distribution of average precision have evident positive correlation with $diff$. Conversely, standard derivation ($stdev$) has evident negative correlation with $diff$. Those can be found in Figure 1. Thus, it appears that the more difficult the topics, namely the more they are shifted to the region of low average precision in the distribution of the average precision, the sharper the distribution becomes.

(2) Evident correlation cannot be seen between the two topic difficulty indicators: $diff$, the actual difficulty based on the submitted results, and $func$, the human-judged difficulty using function-based topic categorization. In addition, we investigated the correlation between $diff$ and the other feature quantities within each

Table 4: Kendall's rank correlation coefficients between the topic difficulty and feature quantities of the topics.

	$diff$	$func$	$\#rel$	ave	$stdev$	med	$skew$	$kurt$	$\#term$	tf_db	df_db	$tfidf\_db$	$ltfidf\_db$
$diff$		0.094	0.087	*-0.798*	*-0.688*	*-0.824*	*0.655*	0.227	-0.068	*0.296*	*0.333*	*0.312*	*-0.291*
		0.443	0.424	0.000	0.000	0.000	0.000	0.035	0.548	0.006	0.002	0.004	0.007
$func$			-0.032	-0.090	-0.023	-0.114	0.110	0.006	0.029	0.015	-0.026	0.081	-0.058
			0.771	0.401	0.829	0.287	0.307	0.954	0.795	0.888	0.810	0.449	0.589
$\#rel$				-0.119	-0.167	-0.119	0.064	-0.062	0.113	0.122	0.195	0.065	-0.116
				0.211	0.080	0.214	0.504	0.514	0.258	0.200	0.040	0.494	0.222
ave					*0.795*	*0.901*	*-0.592*	-0.181	0.060	-0.193	*-0.266*	-0.203	0.196
					0.000	0.000	0.000	0.055	0.541	0.041	0.005	0.032	0.038
$stdev$						*0.736*	*-0.430*	-0.182	0.045	-0.147	*-0.246*	-0.134	0.221
						0.000	0.000	0.054	0.648	0.119	0.009	0.156	0.019
med							*-0.669*	-0.200	0.058	-0.201	*-0.274*	-0.214	0.221
							0.000	0.035	0.556	0.034	0.004	0.024	0.019
$skew$								*0.244*	-0.015	0.174	0.186	0.199	-0.160
								0.010	0.883	0.066	0.050	0.036	0.091
$kurt$									-0.080	-0.126	-0.112	-0.052	0.094
									0.416	0.182	0.237	0.581	0.319
$\#term$										0.171	0.148	0.148	-0.093
										0.084	0.135	0.135	0.349
tf_db											*0.803*	*0.714*	*-0.338*
											0.000	0.000	0.000
df_db												*0.569*	*-0.417*
												0.000	0.000
$tfidf\_db$													-0.232
													0.014
$ltfidf\_db$													

upper sections in each cell: Kendall's τ; lower sections: observed two-sided significance level α;
emphasized: correlation coefficients which are significant within 1% two-sided significance level;
underlined: correlation coefficients which are significant within 5% two-sided significance level.

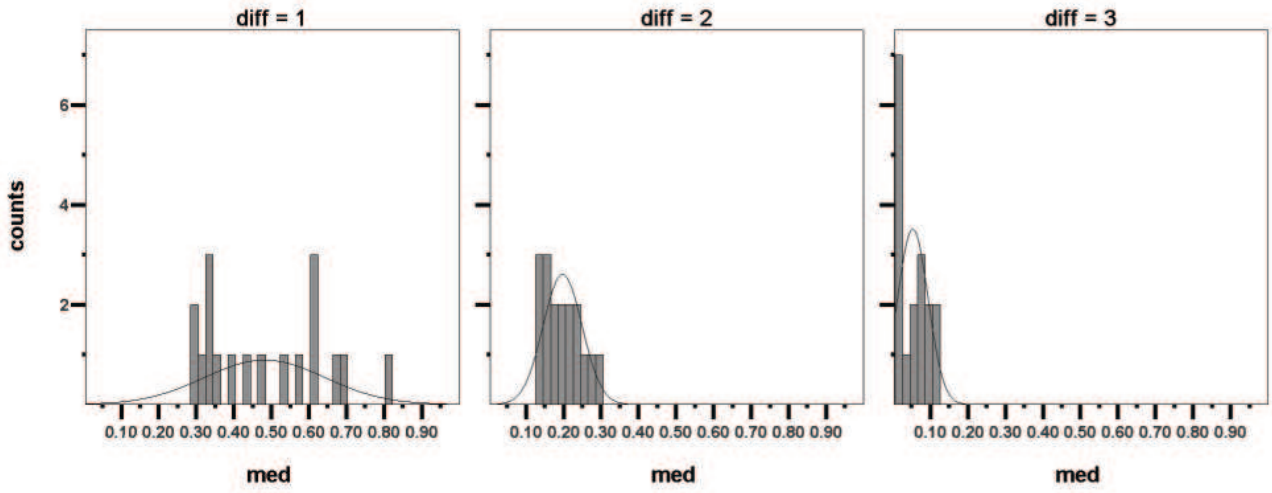

Figure 1: Histograms of medians of the non-interpolated average precision based on submitted results for three topic difficulty levels.

$func$ category, and investigated the correlation between $func$ and the others within each $diff$ category. However, we found no evident correlation in the results.

(3) The average number of relevant documents ($\#rel$) and topic terms ($\#term$) do not have evident correlation with $diff$.

(4) Kendall's τ between tf_db and df_db is larger than 0.80, and exhibits strong positive correlation in the statistical test. In addition, a τ between ltf_db and idf_db exhibits strong negative correlation[4]. Thus, it does not seem to make sense that tf_db and df_db (ltf_db and idf_db) are treated as independent feature quantities, as in Eqs.(3), (4). This can be found through the fact that τ between $diff$ and tf_db (df_db) nearly equal τ between $diff$

[4] This result is omitted in Table 4 because of lack of space.

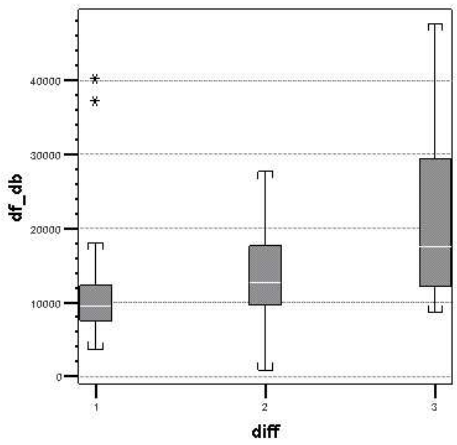

Figure 2: Box-and-whisker graph presenting the relation between df_db and $diff$.

and $tfidf$ ($ltfidf$) in Table 4. Hereafter, we will suppose df_db to be the typical feature quantity of those using frequency information in the document database.

The τ between df_db and $diff$ is more than 0.33, which is not a large value, but exhibits evident positive correlation through the statistical test. It also can be found in Figure 2. This suggests that the more frequently the topic terms appear in the document database, the less effectively the search can be achieved. However, it is not easy to predict the topic difficulty or its distribution using only the frequency information of the topic terms in the document database.

4. Conclusions

Respective system ranks of top runs may be significantly changed to topic difficulty, although the differences of total ranking is not significant as the results of statistical tests. This suggests that the topic difficulty affects the system ranks to some extent.

Evident positive correlation can be found between the topic difficulty and the frequencies of the topic terms in the document database, but it does not have a large correlation coefficient. This suggests that the various IR methods that have been proposed so far tend to depend on the topic / query term frequencies in the document database.

Unfortunately, the evident correlation cannot be seen between two topic difficulty indicators: the actual difficulty based on submitted results and the human-judged difficulty using function-based topic categorization.

Further investigations on various features of the topics and their combinations are required to predict, for practical use, the topic difficulty or its distribution.

Acknowledgements

A part of this research is supported by the "Research for the Future" program (JSPS-RFTF96P00602) of the Japan Society for the Promotion of Science, and the Telecommunications Advancement Foundation, Japan.

5. References

Chris Buckley, Gerard Salton, and James Allan. 1993. Automatic retrieval with locality information using SMART. In D. K. Harman, editor, *Proceedings of the First Text REtrieval Conference (TREC-1)*, pages 59–72. NIST Special Publication 500-207.

Noriko Kando, Kyo Kageura, Masaharu Yoshioka, and Keizo Oyama. 1998. Phrase processing methods for Japanese text retrieval. *SIGIR Forum*, 32(2):23–28.

Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Soichiro Hidaka. 1999. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 11–44, Tokyo, Japan.

Kazuaki Kishida. 2001. Property of mean average precision as performance measure in retrieval experiment. *IPSJ SIG Notes*, (2001-FI-63):97–104. (in Japanese).

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Osamu Imaichi, and Tomoaki Imamura, 1997. *Japanese Morphological Analysis System ChaSen Manual*. (in Japanese).

National Institute of Informatics, 2001. NTCIR Workshop. ⟨http://research.nii.ac.jp/ntcir/workshop/⟩.

Tetsuya Sakai, Tsuyoshi Kitani, Yasushi Ogawa, Tetsuya Ishikawa, Haruo Kimoto, Ikuo Keshi, Jun Toyoura, Toshikazu Fukushima, Kunio Matsui, Yoshihiro Ueda, Takenobu Tokunaga, Hiroshi Tsuruoka, Hidekazu Nakawatase, Teru Agata, and Noriko Kando. 1999. BMIR-J2: A test collection for evaluation of japanese information retrieval systems. *SIGIR Forum*, 33(1):13–17.

Gerard Salton. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.

Ellen Voorhees and Donna Harman. 1997. Overview of the sixth Text REtrieval Conference (TREC-6). In E. Voorhees and D. K. Harman, editors, *Proceedings of the 6th Text REtrieval Conference (TREC-6)*, pages 1–24. NIST Special Publication 500-240.