An Analysis on Topic Features and Difficulties Based on Web Navigational Retrieval Experiments

Masao Takaku[1], Keizo Oyama[2] and Akiko Aizawa[2]

[1] Research Organization of Information and Systems, Japan

[2] National Institute of Informatics, Japan

masao@nii.ac.jp

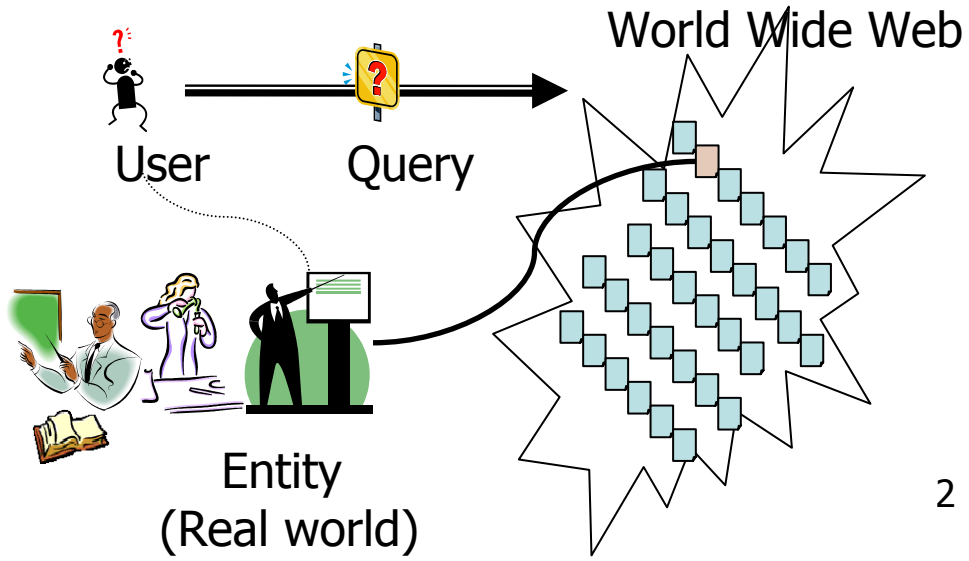
Third Asia Information Retrieval Symposium (AIRS),
Singapore, October 16-18, 2006

NTCIR

NII

What we do

- NTCIR-5 WEB task
  - Web Navigational Retrieval Subtask (Navi-2)
    - = known-item search (Representative pages)
- Analysis on topic difficulties
  - Metadata of topics
  - Query terms
  - Pool size



Navi-2: Web Navigational Retrieval

- Web test collection
  - Dataset: over one terabyte, 100 million pages
  - Topics: 400 topics (Japanese)
  - Relevance judgements (graded) - 3 levels
- 63 search result runs (from 6 teams)
- Evaluation measure
  - Reciprocal Rank (RR) $1/Rank_{first}$
  - Discounted Cumulative Gain (DCG)

$$DCG[i] = \begin{cases} CG[i], & if\ i = 1 \\ DCG[i-1] + G[i]/\log_b(i), & otherwise \end{cases}$$

Topic Example

Figure 1 (p.627)

<NUM>1041</NUM>
<TYPE>1</TYPE><CATEGORY>B</CATEGORY>
<TITLE>**UNESCO**</TITLE>
<DESC>I want visit to the homepage of UNESCO.</DESC>
<NARR>
<BACK>I would like to know about activities of UNESCO</BACK>
<RELE>The top page of the National Federation of UNESCO Association in Japan would be relevant.</RELE>
</NARR>
<USER SPECIALTY="C">Graduate-doctoral course 1st year, female, 5 years experience in searching</USER>

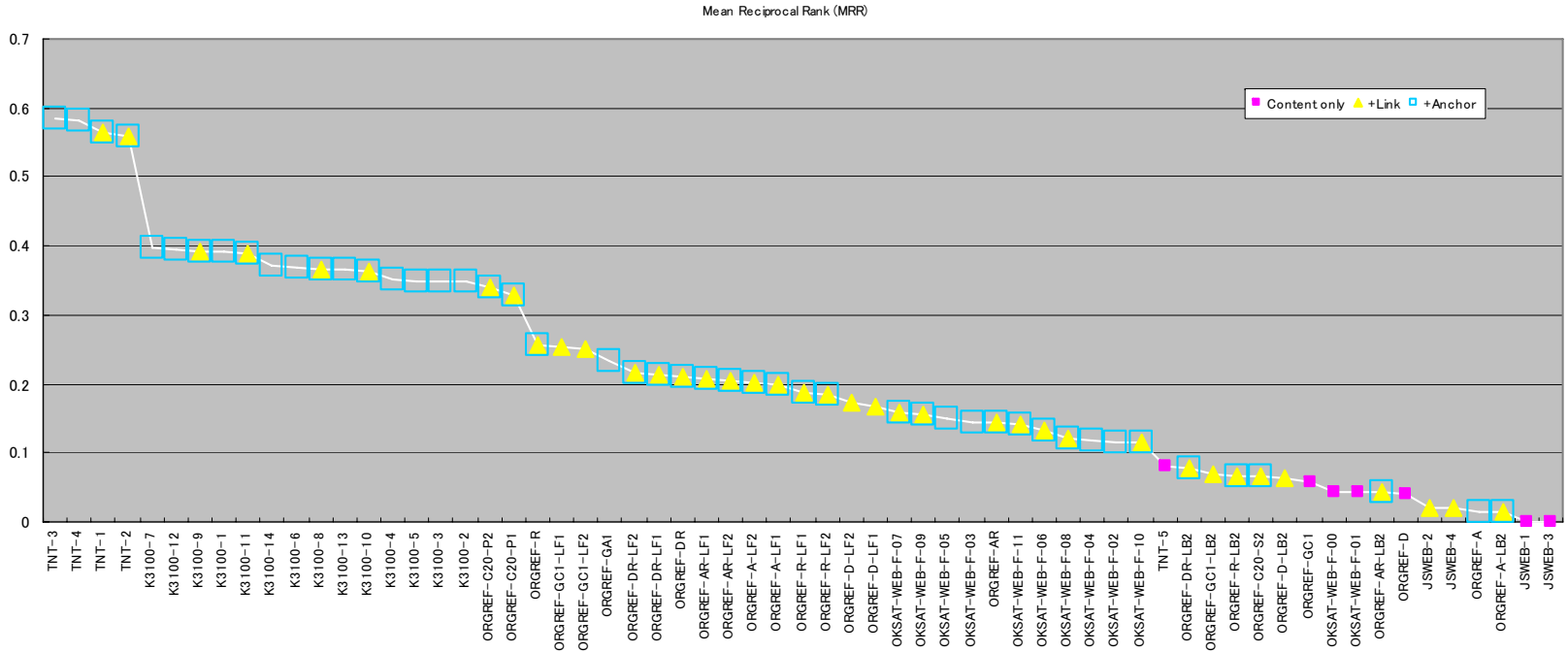
- 1: a phrase represents its target item
- 2: two or more phrases represent the item
- 3: one or more phrases do not specifically represent the item.

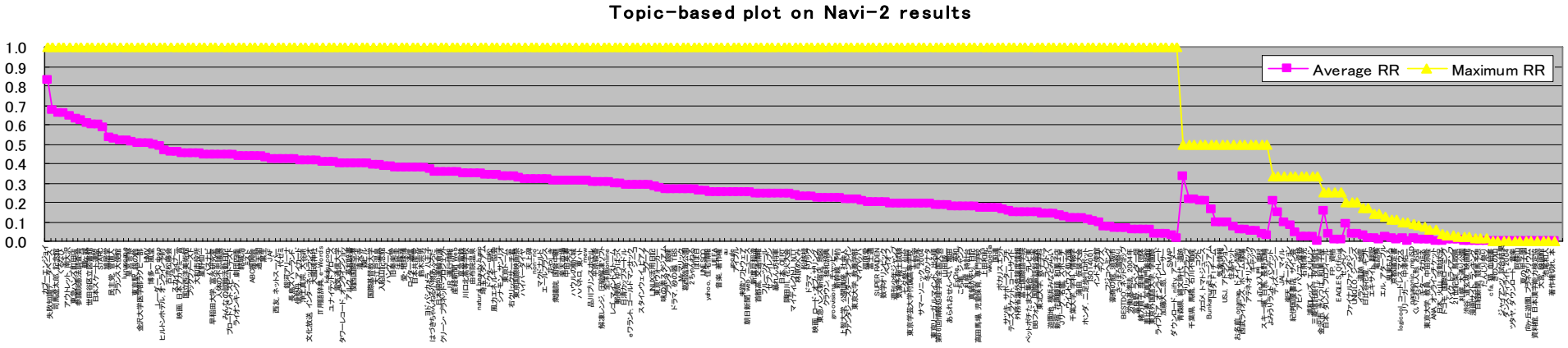
- A: searcher knows the item in detail
- B: searcher knows its outline
- C: searcher knows it to the extent the item can be identified among others
- D: searcher knows only its existence but knows very little about the item

- A: Products / services
- B: Companies / organizations
- C: Persons
- D: Facilities
- E: Sights, historic spots, and natural things
- F: Information resources
- G: Online shops / services
- H: Events

Table 2 (p.627)

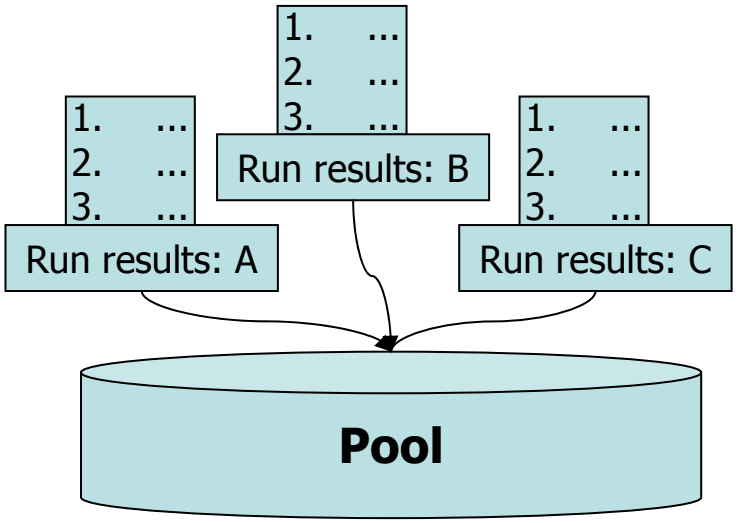
NTCIR WEB Navi-2 results

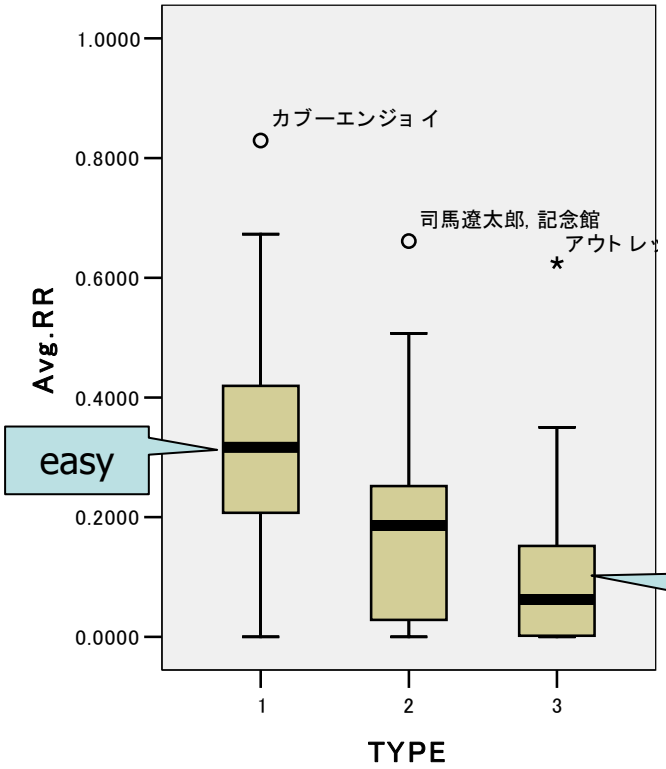


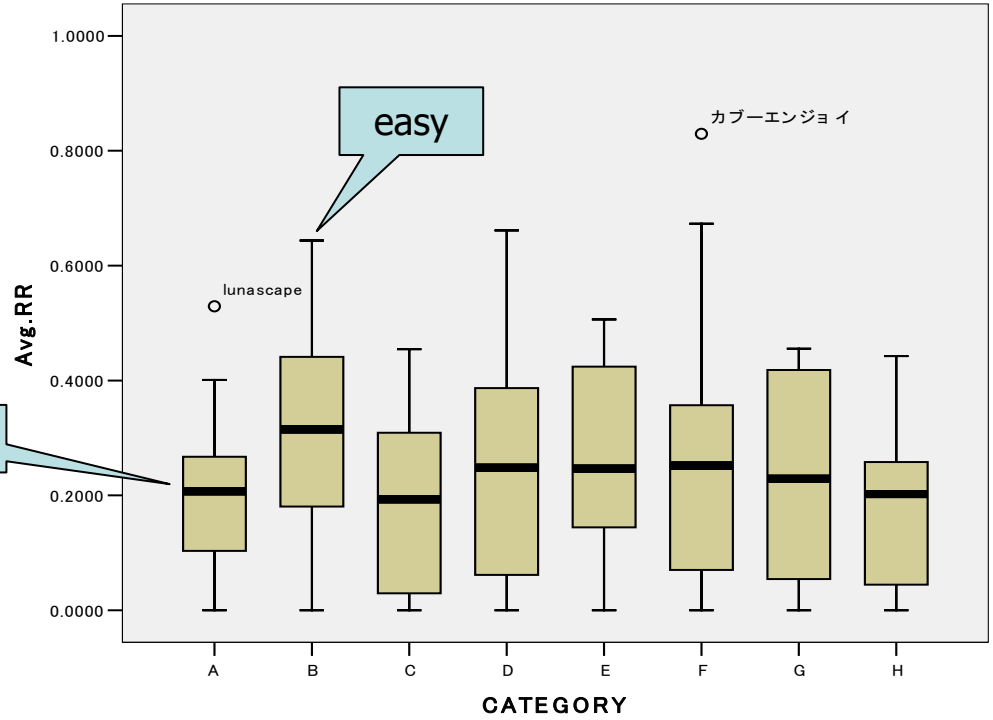


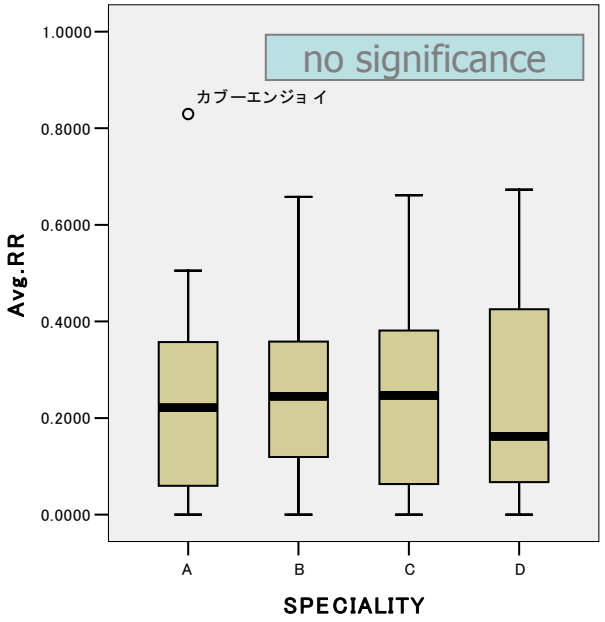
Analysis on topic difficulties

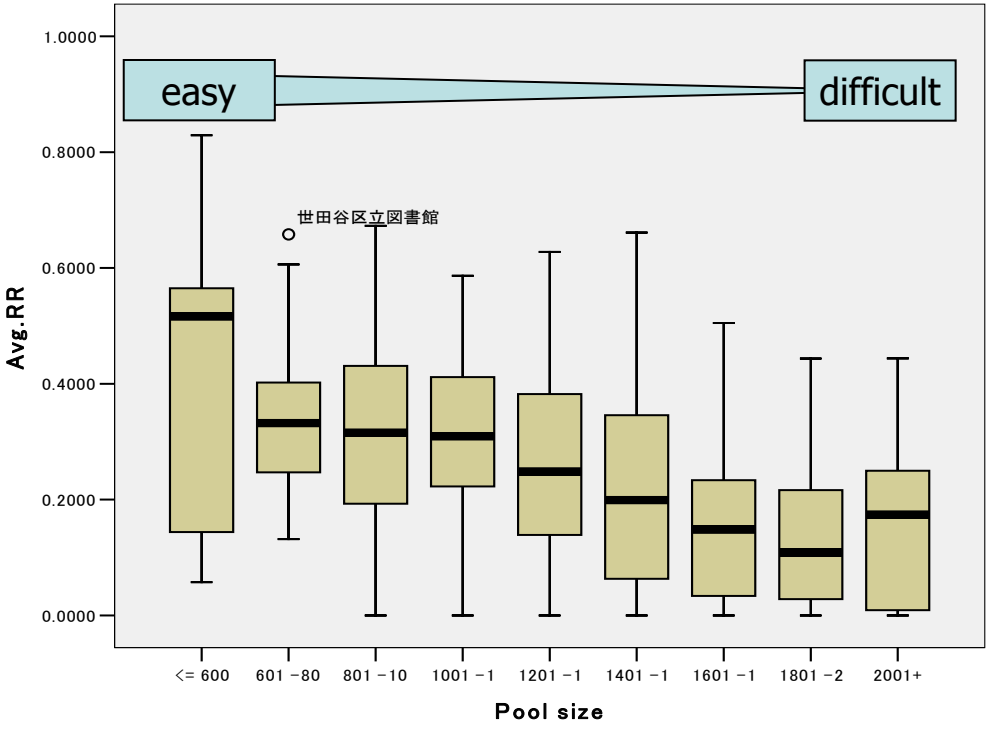
- Selected 14 runs, 269 topics
- Topic group analysis
  - Metadata
    - TYPE, CATEGORY, SPECIALTY
  - Query length
    - phrase-, word-, byte-
  - Query-corpus features
    - IDF, QScope, AvICTF, SCS
  - Pool size
    - the number of unique documents in the pool, which was constructed from the selected run results in which pool depth were at most 100.




1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
Avg.RR
カブーエンジョイ
司馬遼太郎, 記念館
アウトレッ
easy
difficult
1
2
3
TYPE


1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
Avg.RR
easy
カブーエンジョイ
lunascape
difficult
A
B
C
D
E
F
G
H
CATEGORY


1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
Avg.RR
no significance
カブーエンジョイ
A
B
C
D
SPECIALITY


1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
Avg.RR
easy
difficult
世田谷区立図書館
<= 600
601 -80
801 -10
1001 -1
1201 -1
1401 -1
1601 -1
1801 -2
2001+
Pool size

Pearson's correlation coefficients *r*

Query length | Query-corpus feature | Pool size

	Avg.RR	ql_{phrase}	ql_{word}	ql_{byte}	IDF_{min}	QScope	AvICTF	SCS	PoolSize
Avg.RR	1	**-.438**	-.112	**-.256**	-.091	.112	-.077	.095	**-.342**
		.000	.067	.000	.137	.066	.209	.066	.000

Table 6. (p. 630)

- Future works
  - Some features significantly correlate with average RR.
    - TYPE, CATEGORY, Pool size ...
  - Hint for improving result (ranking)
  - Switching algorithms/techniques along with topic difficulties