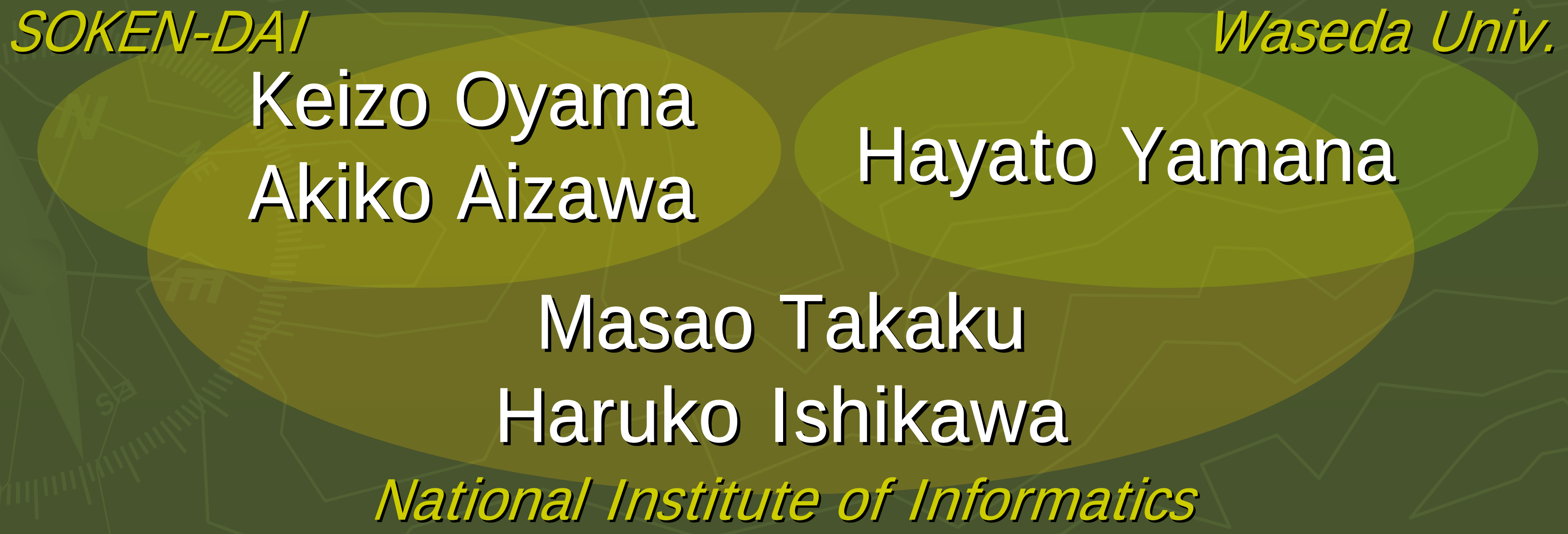
Overview of the NTCIR-5 WEB Navigational Retrieval Subtask 2 (Navi-2)



SOKEN-DAI

Keizo Oyama
Akiko Aizawa

Waseda Univ.

Hayato Yamana

Masao Takaku
Haruko Ishikawa

National Institute of Informatics

NTCIR WEB Tasks

- NTCIR-3 WEB
  - Survey retrieval (Topic retrieval, Similarity retrieval)
  - Target Retrieval
  - Optional
    - Search results classification
    - Speech-driven retrieval
- NTCIR-4 WEB
  - Informational retrieval
  - Navigational retrieval (Navi-1)
  - Geographical information retrieval (pilot subtask)
  - Topical classification (pilot subtask)
- NTCIR-5 WEB
  - **Navigational retrieval (Navi-2)**
  - Query term expansion (pilot subtask)

What is *Navigational Retrieval*?

"Web IR for helping a user to visit a specific web page of something"

- ► Navi-1 & Navi-2: "Known Item Search"
  - The user knows the item to some degree.
  - The user searches for one or a few representative pages of the item.
  - The user may/may not know the page.

Product, organization, store, person, facility, natural thing, event … existing in the real world

Information service, blog, data file, document, online shop … existing in the cyber space

What does “representative” mean?

- The provider of the page must be responsible for or authoritative about the entity.
- Content of the page must cover strongly related information comprehensively and is preferred to contain least irrelevant information.

- ✓ A partial *frame* page is usually regarded as imperfect and hence cannot be a representative page.
- ✓ Both of an entry page consisting of only a movie, etc. and a fully informative top page may be regarded as representative pages.
- ✓ An entry page without content but just for redirecting to another representative page may be regarded as a representative page.

NW1000G-04: Document data set

- Web pages: 1.36TB (1.5 × 1012 bytes)
  - raw: as were crawled
  - euc: EUC character code
  - cooked: extracted text data
  - segmented: segmented word data
- List data
  - sitelist: site ID+site name ... 389,875 sites
  - doclist: doc ID+URL ... 95,870,352 pages
  - linklist: (doc ID+URL)^2 ... 1,290,150,449 links
- Crawled sites:
  - Mainly in JP domain
    - not necessarily in Japanese
  - Some from other domains
    - sites judged as including pages in Japanese

Search topics

- 17 topic creators
- 400 topics out of 891
- 842 optional topics

A: Knows the item in detail.
B: Knows the outline of the item.
C: Knows the item to the extent the item can be identified from others.
D: Knows existence of the item but little about it.

H: Events
Z: Others

```
<TOPIC>
<NUM>1251</NUM>
<TYPE>3</TYPE>
<CATEGORY>D</CATEGORY>
<TITLE>三鷹，ジブリ，記念館</TITLE>
<DESC>東京三鷹市にあるジブリアニメの記念館につ
    いて調べたい。</DESC>
<NARR>
<TERM>ジブリとは「となりのトトロ」や「魔女の宅急
    便」など有名アニメを制作してるアニメ制作会社
    である。</TERM>
<BACK>東京の三鷹市にジブリアニメの資料を集め
    た記念館があると聞き、詳しく知りたいと思った。
    </BACK>
<RELE>ジブリ美術館の公式ページを適合とする。
    </RELE>
</NARR>
<USER SPECIALTY="C">大学学部4年，男性，検索
    歴5年</USER>
</TOPIC>
```

Relevance assessment (1)

- Pooling
  - Top 20 from each run
    - Necessarily assessed; in the order of ranks and URLs
  - All docs from each run
  - Docs linked from them
    - Optionally assessed
- Judgment bases
  - Content
    - text; images, etc. if still available
  - Link source/target pages
    - anchor, frame, meta refresh, …
  - URL
  - Current web pages

Relevance assessment (2)

- Relevance judgment
  - A: Relevant
    - Representative page of the search target item
    - Can be more than one
  - B: Partially relevant
    - Information need must be almost satisfied
      - Provider is a little inappropriate
      - Subject is a little inappropriate but the relevant page can be reached easily

  Rigid = A; Relaxed = A + B
- Additional judgment
  - Undistinguishability
    - Representative page of an irrelevant item that cannot be eliminated with TITLE and DESC parts of topics
  - Duplication (only for rel. and partially rel. docs)

Relevance assessment (3)

- What kind of pages can be relevant?
  - Relevant
    1. Just a representative page containing required content
    2. Entry page with movie or animation redirecting to 1.
    3. Frameset page consisting of component pages
       (in most cases, each component page cannot be relevant or even partially relevant)
    - Page provider must be appropriate for all cases
  - Partially relevant
    1. Pages one level upper/lower with link to rel. page
    2. Directory page on the search target with link to rel. page
    3. Satisfactory content but not by representative provider
    4. Almost relevant but having unacceptable URL
    - Page provider must be reliable
    - Link to rel. page (if required) must be easy-to-find

Relevance assessment (1)

- Pooling
  - Top 20 from each run ------ 456+1484
    - Necessarily assessed; in the order of ranks and URLs
  - All docs from each run ------ 104+ 532
  - Docs linked from them ------ 63+ 225
    - Optionally assessed
- Judgment bases
  - Content
    - text; images, etc. if still available
  - Link source/target pages
    - anchor, frame, meta refresh, …
  - URL
  - Current web pages

Summary of participation

- Participated groups
  - Kansai Lab., NEC Corp.
  - Research and Development Strategy Dept.; Justsystem Corp.
  - Sato Lab., Osaka Kyoiku Univ.
  - Software Engineering Center; Univ. of Aizu
  - Univ. of Tsukuba, Nagoya Univ., Toyohashi Univ. of Technology
  - Organizers
- Technologies attempted
  - System structures
    - Agent-type distributed system based on VSM and term-partitioning
    - Doc-partitioned index based on VSM
  - Information sources and scoring methods
    - Full text, Title part --- Boolean+TF-IDF, VSM, Probabilistic model
    - Anchor text --- Boolean, Boolean+TF-IDF, Language model
    - Link structure --- PageRank, Anchor count, In-link count, Out-link count
    - URL --- Same site, Same network domain
  - Score merging methods
    - Product
    - Weighted harmonic mean
- Runs --- 63 in total
  - 44 used anchor text; 39 used link information; 4 used URL informaton

System evaluation methods

- Number of topics used: 269
  - At least one relevant document exists in the corpus
- Evaluation measure
  - DCG: Discounted Cumulative Gain (cut-off at 10)

    (Ga, Gb) = (3, 0), (3, 2) and (3, 3)

    - Correction considering duplication/redundancy is necessary
  - WRR: Weighted Reciprocal Rank (cut-off at 10)

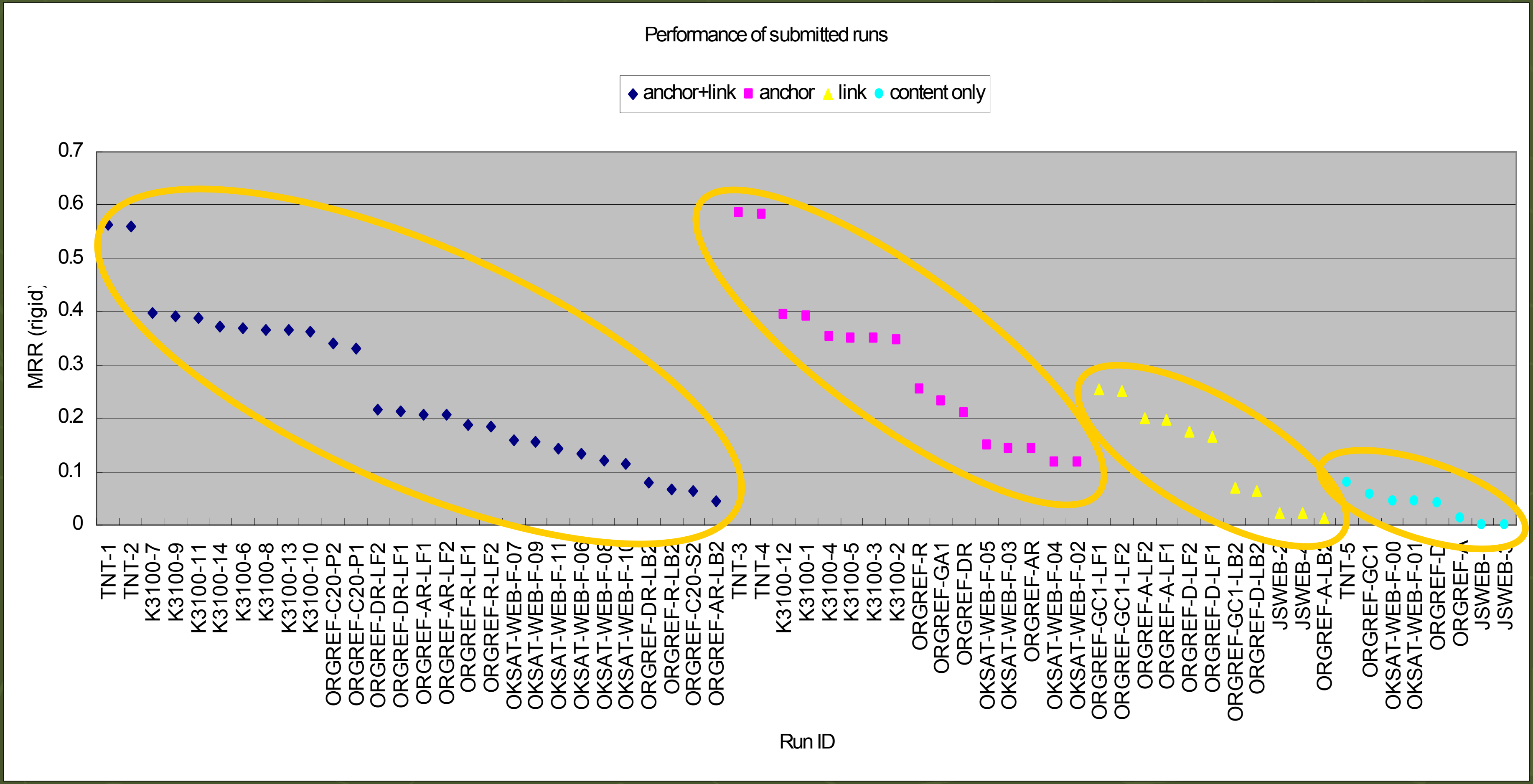
    $(\delta a, \delta b)$ = (1, 0) and (1, 1), $(\beta a, \beta b) = (\infty, \infty)$

    - Equivalent to MRR with rigid/relaxed relevance level

Evaluation results

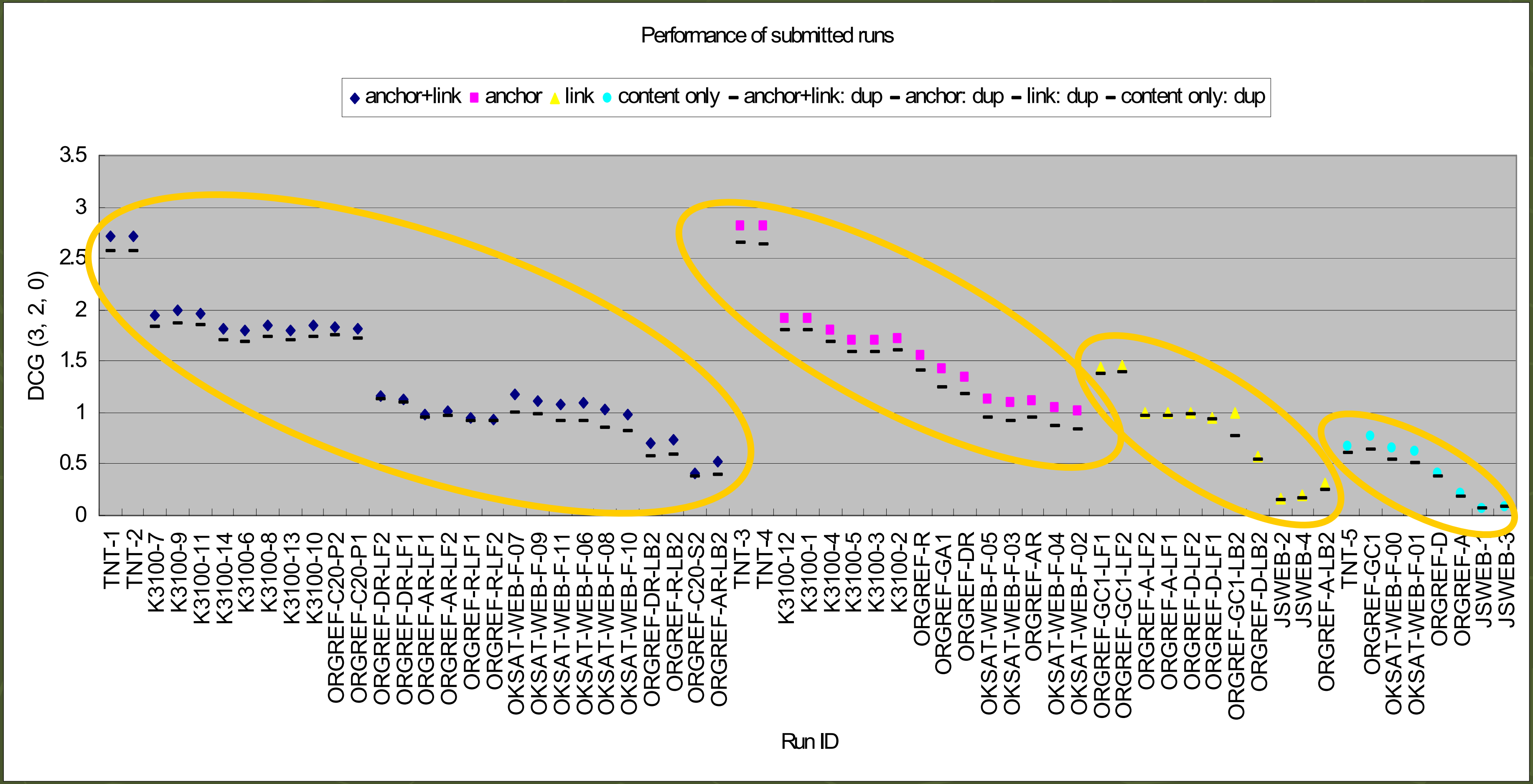
WRR (1,0) (=MRR rigid)

- Reflecting the rank of relevant doc. appearing first
- Maybe weighting too much on highly ranked pages; i.e. evaluation by rather easily retrieved documents/topics



DCG (3,2,0) + dup. correction

- Reflecting compound aspects: ranks, grade, multiple rel. docs
- Probably representing user’s satisfaction better than WRR



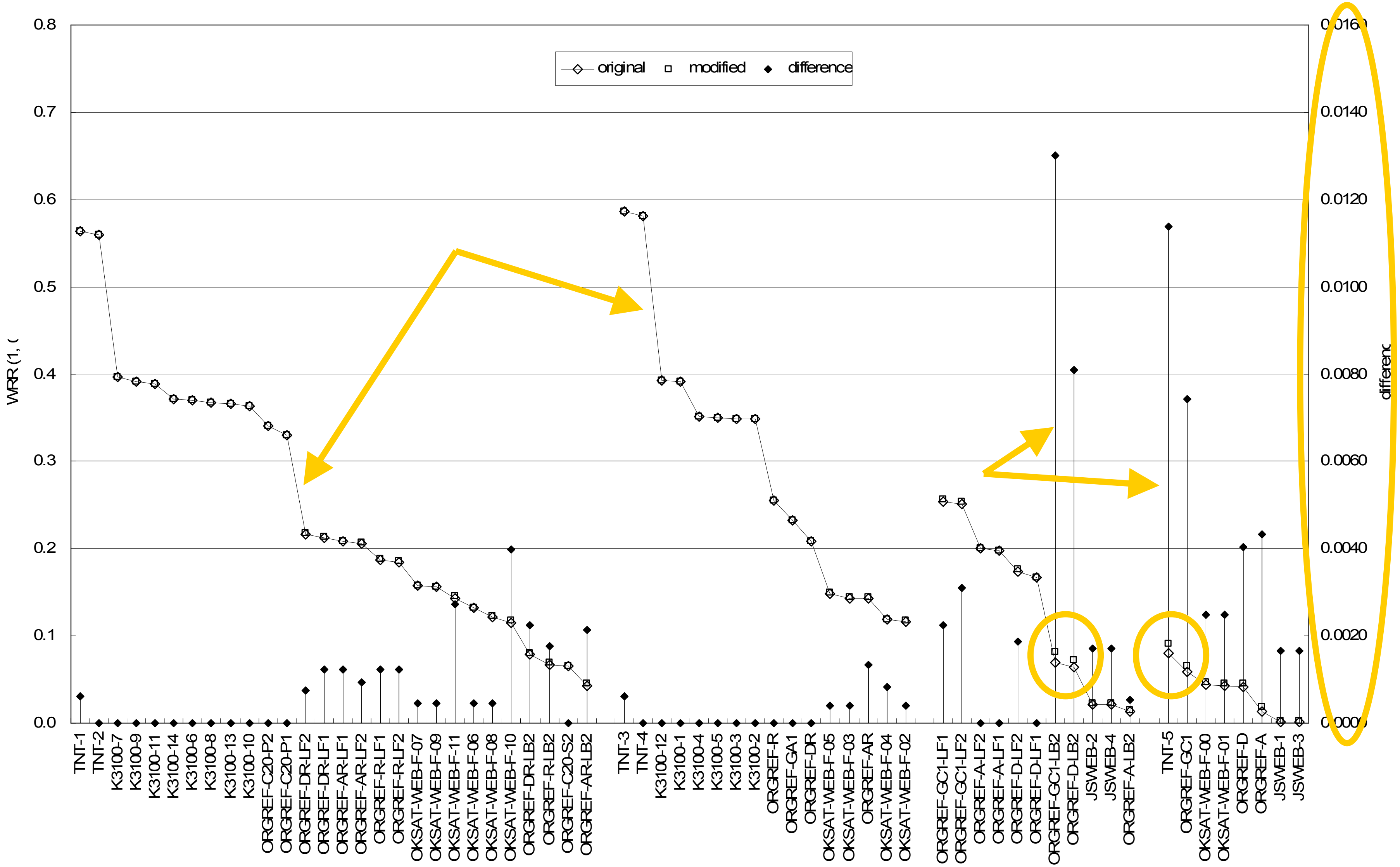
Analysis of web-specific effects (1)

- Duplicate documents
  - Causes
    - Aliased sites: www.abc.co.jp, abc.co.jp, www.abc.jp ...
    - Directory entry pages: x/, x/index.html, x/Default.asp
    - Ordinal duplication
  - Correction methods
    - Ignore duplicates (moderate correction)
    - Regard duplicates as non-relevant (large correction)
  - Tendencies
    - Effects vary among runs
    - No outstanding trends among information sources used
    - Correction result seems to converge to relative rank of WRR
  - Link-related redundancy is intrinsic to web documents; its effect is estimated to be double or more

Analysis of web-specific effects (2)

- FRAMESET structure
  - Frameset is used for:
    - Showing a static menu or a logo image
    - Hiding the actual URLs
    - Just layout, …
  - Relevance assessment
    - Frameset pages are the first candidate
    - Frame component pages can rarely be relevant or partially relevant
  - Problem
    - No effective content is contained in most frameset pages
  - Correction method
    - Make frame component pages referenced only from the frameset page inherit the relevance
  - Tendencies
    - Effects vary among runs
    - Effective mainly for non-anchor-text-based systems

FRAMESET effects on WRR (1,0)



Some observations

- Several anchor-base systems performed best
- Link-base method or URL-base method made no improvement on anchor-base systems
- Several link-base systems performed fairly
- Proper handling of FRAMESET will improve non-anchor-base systems
- Duplication affects to DCG notably; link-related redundancy is estimated to affect much more

Conclusion

- Task settings
  - Huge document data set : 1.36TB
    - Not only the participants but also the organizers suffered from the amount
    - Systems were not tuned before formal run submission
      ➔Please check the participants presentation
- Consistent tendencies with Navi-1:
  - High performance can be achieved only by exploiting anchor text
  - Highly-ranked systems have achieved satisfactory performance for most topics
  - Systems attempting to use local information (i.e. content, local links, etc.) only still stay at relatively low performance

Future works

- Evaluate systems considering link-related redundancy
- Verify stability of evaluation measures
- Check comprehensiveness of assessment results
- Study on evaluation measures reflecting users' overall cost
- Analyze topic-by-topic behavior of each system
- Analyze relations among topic types, search target categories, styles and structures of relevant pages, and effective techniques

Acknowledgements

- Supported by
  - Grants-in-Aid for Scientific Research on Priority Areas of “Informatics” (#13224087), MEXT
- Cooperated in web crawling with
  - e-Society Foundation Software project, MEXT
- Thanking to
  - People who helped preparing web doc data
  - Topic creators and assessors
  - All the participants to Navi-2