Navigation Retrieval with Site Anchor Text

Hideki KAWAI, Kenji TATEISHI and Toshikazu FUKUSHIMA

NEC Internet Systems Research Labs.

Introduction

- Navigation Retrieval Task in NTCIR-4 WEB (task B)
  - Searching for one or more "representative Web pages."
  - *Relevancy* and *Representativeness* of document are both important.
- Motivation
  - Verify the efficiency of referential information

➡ Retrieval system which indexes only site anchor text

Two advantages :

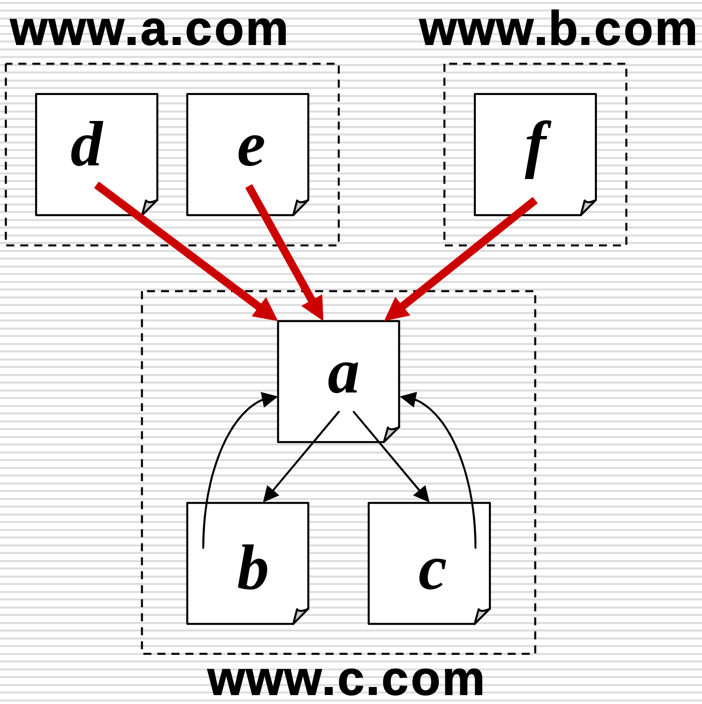
- The index size is very small.
- A user can retrieve uncrawled documents as well as crawled documents.

Site Anchor Text

- Anchor text of links from external Web site
  - Anchor(d,a)+Anchor(e,a)+Anchor(f,a)

Summarizing content and popularity of the Web site

We can calculate *relevancy* and *representativeness*.



Note :

We defined "external Web sites" simply as sites whose domain name is different from the target page.

Retrieval Method

Step1 : Parse the query and search for pages
Step2 : Determine score of each page
Step3 : Sort pages by Score

□ **Score of page *p***

$$\mathrm{Score}(p) = \mathrm{Rep}(p) \times \mathrm{Rel}(p, q)$$

Representativeness of page *p*
derived from link structure

Relevancy of page *p* and query *q*
based on two kinds of measures, *reference consistency* and *specificity of word combination*

Representativeness of page p

□ Derived from link structure

$$\mathrm{Rep}(p) = C \times T$$

C : Citation frequency from external Web sites

T : Likelihood of top page determined by following heuristics:

(H_1) Does the URL of the page consist of only domain name?

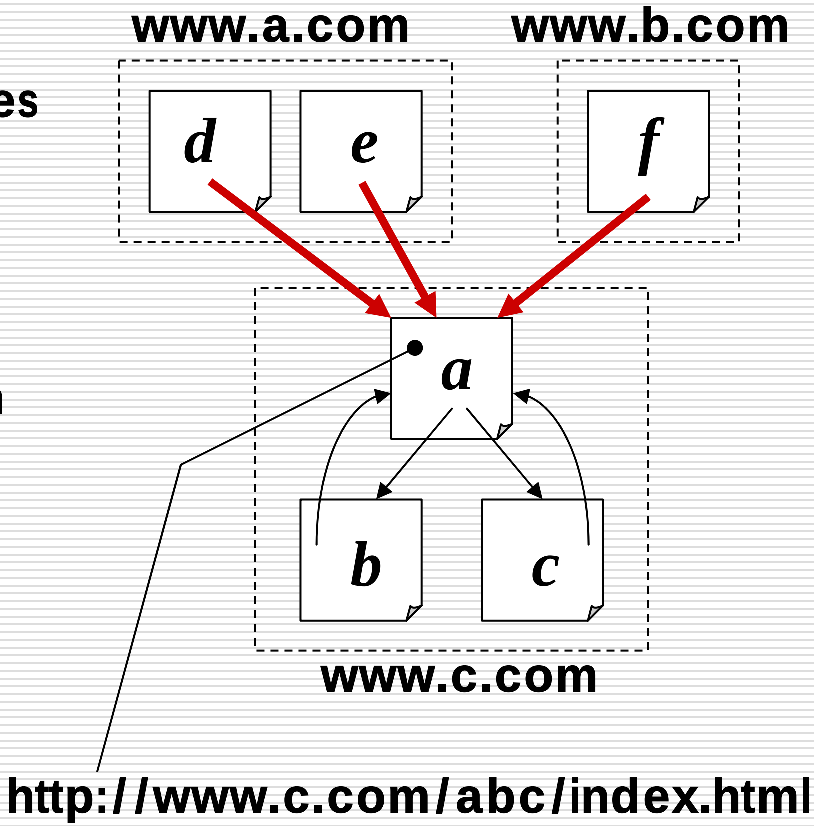
(H_2) Does the file name of the URL contain such a string as "index" or "default"?

(H_3) Does the URL end with a slash "/" ?

$$T = w_1 \times \delta_1 + w_2 \times \delta_2 + w_3 \times \delta_3 + w_4$$

$$\delta_i = \begin{cases} 1 & \text{if } \mathrm{H}_i \text{ is } \mathit{true} \\ 0 & \text{if } \mathrm{H}_i \text{ is } \mathit{false} \end{cases}$$

$$(w_1, w_2, w_3, w_4) = (1000, 100, 10, 1)$$



e.g. $\mathrm{Rep}(a) = 3 \times 101 = 303$

Relevancy of page *p* and query *q*

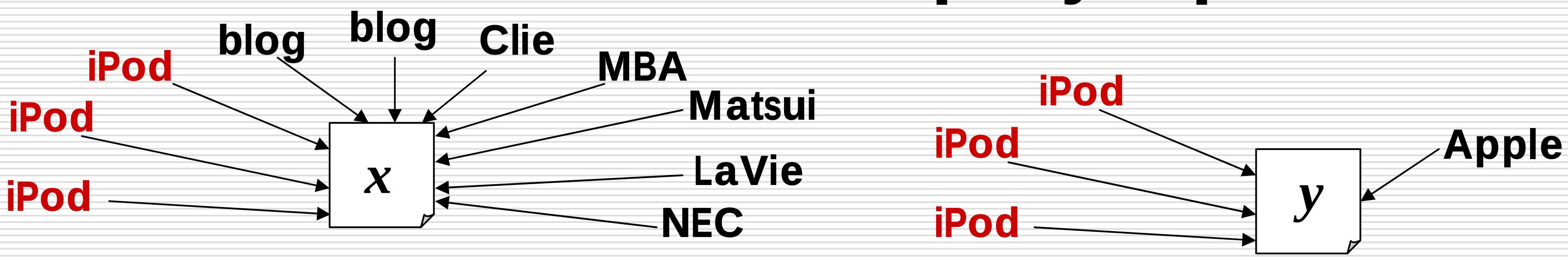
Main concept :

Effective use of limited information to determine the relevancy

- □ Reference consistency
  - How consistently is the page referred by external Web sites?
  - (How sharply does the site focus on a topic?)
- □ Specificity of word combination
  - How specifically are pages identified by given word combination?

Reference consistency

□ Which is relevance for query "i-pod" ?



$$\mathrm{Rel}(p,q) = \sum_{t\in q} kw_t \times \left(\frac{f_t^{\,2}}{N_{sa}} \right)$$

f_t : Frequency of word t in the site anchor text for page p

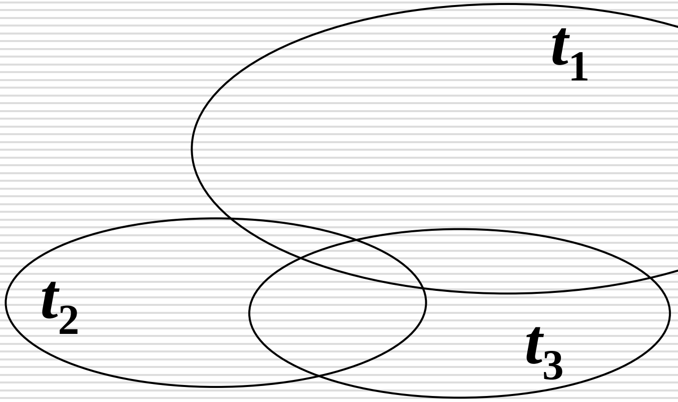
N_{sa} : Amount of site anchor text for page p

kw_t : Weight of the word in query q $\quad kw_i = 2^{(n_q - i)}$

In this case ... $\mathrm{Rel}(x,\text{"}\mathbf{iPod}\text{"}) < \mathrm{Rel}(y,\text{"}\mathbf{iPod}\text{"})$

Specificity of word combination

- How specifically are pages identified by given word combination?



$$\mathrm{Rel}(p,q) = \log \frac{N}{\left|\mathrm{D}(\tau \in p,q)\right|}$$

$\left|\mathrm{D}(\tau \in p,q)\right|$ **: Number of pages that contain keyword group included in both page p and query q**

if $\left|\mathrm{D}(t_1,t_2,t_3)\right| < \left|\mathrm{D}(t_1,t_2)\right| < \left|\mathrm{D}(t_1,t_3)\right| < \left|\mathrm{D}(t_2,t_3)\right|$ **and**

$i \in \mathrm{D}(t_1,t_2,t_3),\ j \in \mathrm{D}(t_1,t_2), k \in \mathrm{D}(t_1,t_3), l \in \mathrm{D}(t_2,t_3)$ **then,**

$\mathrm{Rel}(i,q) > \mathrm{Rel}(j,q) > \mathrm{Rel}(k,q) > \mathrm{Rel}(l,q).$

Note :

Traditional TF-IDF schema tends to be biased toward words with highly specificity (t_2 and t_3), so $\mathrm{Rel}(l, q) > \mathrm{Rel}(j,q)$ or $\mathrm{Rel}(k,q)$ in this case.

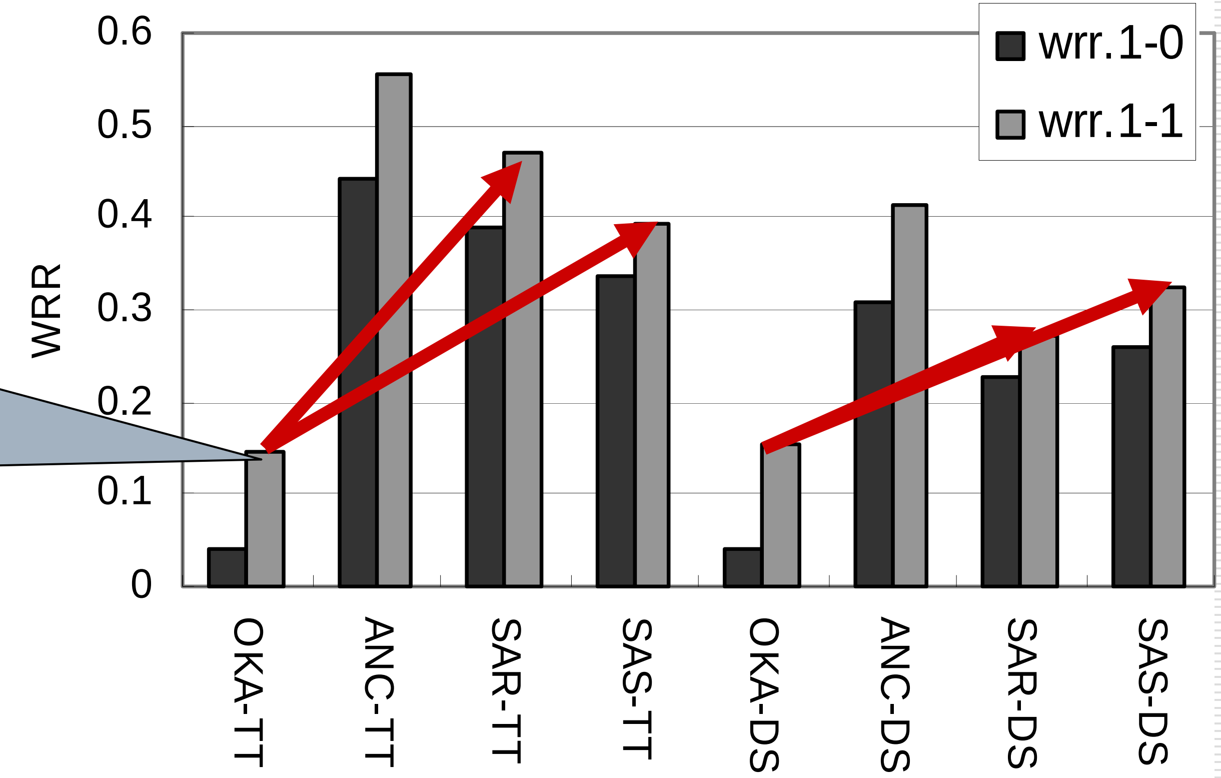
Evaluation

- Document collection :100GB NW100G-01
- Total size of site anchor text : 94MB
- Evaluation scales : WRR (and DCG)
  - "relevant", "partially relevant", "irrelevant"
- Compared with following 4 systems:

ID	Index	Relevancy calculation
OKA	Full text of crawled pages	OKAPI
ANC	Full text of crawled pages	High weight to anchor text
SAR	Site anchor text only	Reference consistency
SAS	Site anchor text only	Specificity of word combination

Result and discussion (1/4)

- Site anchor text retrieval (SAR and SAS) has great advantages over simple full text retrieval (OKA).



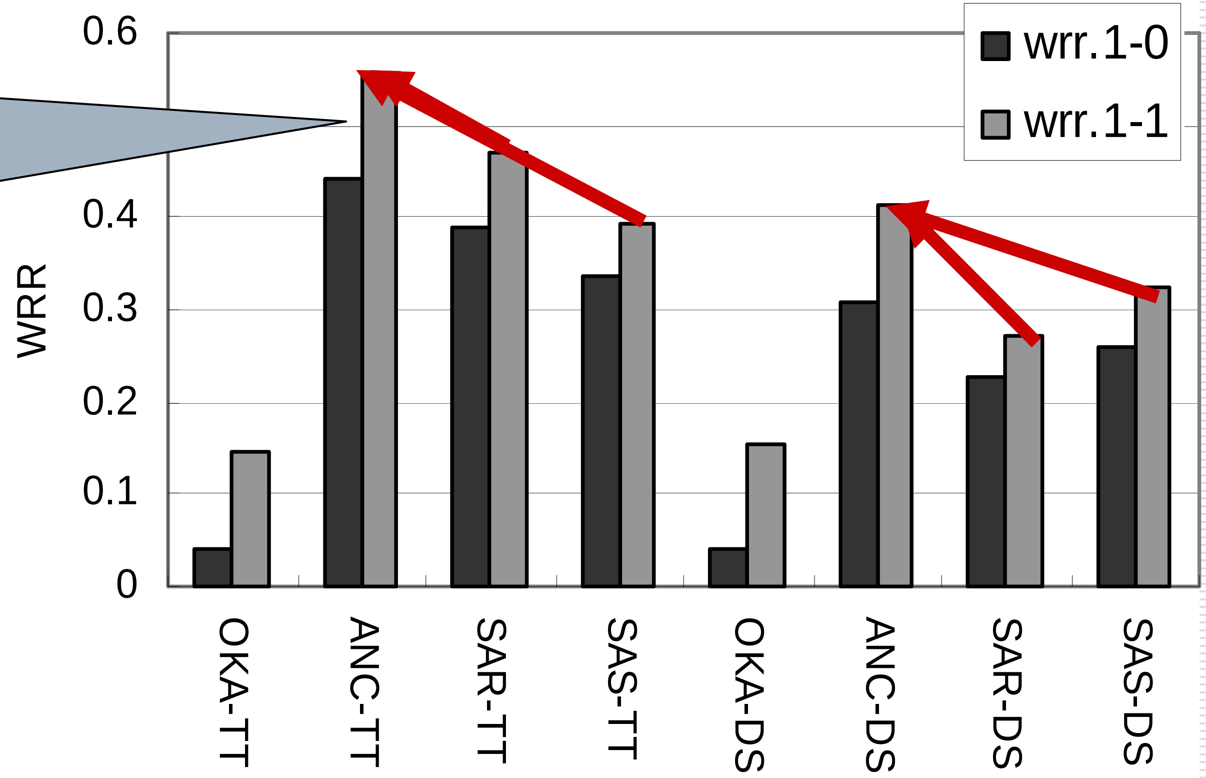
Site anchor text retrieval (SAR and SAS) outperformed the simple full text retrieval (OKA)

※ TT : <TITLE> / DS : <DESC> for TopicPart

Result and discussion (2/4)

- Some important information in anchor text can be lost when site anchor text was extracted.
  - e.g. http://abc.jp/~usr1/ and http://abc.jp/~usr2/ are dealt with as the same site.

Anchor weighted full text retrieval (ANC) was better than site anchor text retrieval (SAR and SAS)



※ TT : <TITLE> / DS : <DESC> for TopicPart

Result and discussion (3/4)

- Despite a very small index, SAR and SAS were comparable with ANC (up to 88% on WRR)
- Especially accuracy ratio tends to be higher in data series that give a score only for the "relevant" pages.
- Site anchor text can pinpoint highly relevant documents.

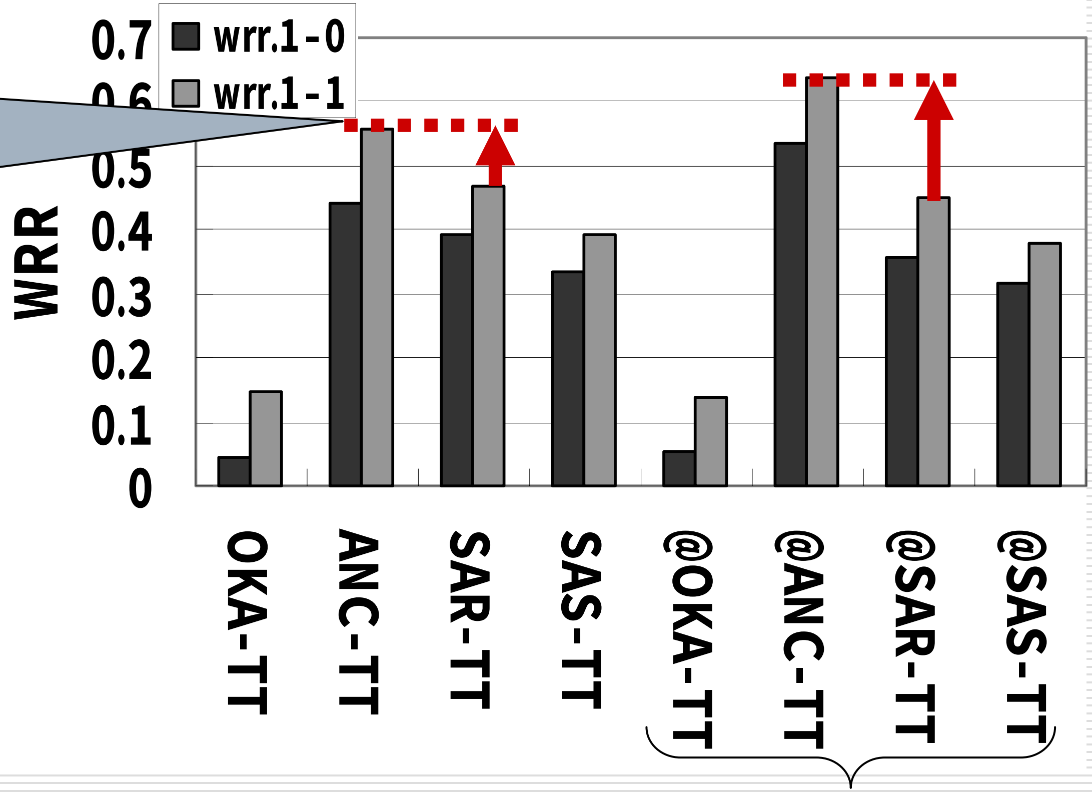
	SAR/ANC	SAC/ANC
dcg.3-0	0.84	0.81
dcg.3-2	0.75	0.71
dcg.3-3	0.72	0.68
wrr.1-0	0.88	0.76
wrr.1-1	0.84	0.71

※TopicPart is <TITLE>

Result and discussion (4/4)

- Some uncrawled pages are "relevant" and relevancy for the uncrawled pages can be determined based on reference information.

The gap of WRR value increased between SAR and ANC (or SAS and ANC) crawled documents



WRR for crawled documents only

Conclusion and Future work

- Site anchor text retrieval system ...
  - Has very small index size (one-thousands of original document set)
  - Outperforms simple full-text retrieval.
  - Is comparable with anchor text weighted full-text retrieval (up to 88% accuracy).
  - Tends to pinpoint highly relevant pages.
  - Can retrieve uncrawled pages as well as crawled pages based on only referential information.
- In future work ...
  - Integrate site anchor text retrieval and traditional retrieval system
  - Address the problem of Web site boundaries