Exploiting Anchor Text for the Navigational Web Retrieval at NTCIR-5

Atsushi Fujii (Univ of Tsukuba)
Katunobu Itou (Nagoya Univ)
Tomoyosi Akiba (TUT)
Tetsuya Ishikawa (Univ of Tsukuba)

WEB-7

Introduction

Taxonomy of queries on the Web

- Navigational — Focus of today's talk
  - A user has a Web site in mind and requires to reach that site
- Informational
  - A user searches for Web sites that provide knowledge for his/her information need

In NTCIR-5 Navi-2 Subtask, a hypothetical user requires to find the representative pages of an item (e.g., person and product)

Contribution of our research

- Exploiting anchor text in Navigational Web Retrieval
  - Modeling anchor text
  - Extracting synonyms from anchor text for query expansion
- Combining different information types
  - anchor text, page content, link structure

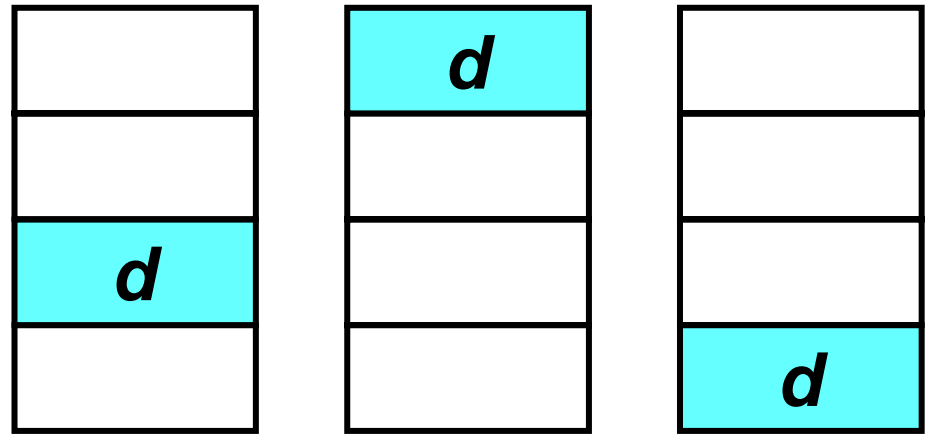
System overview

- We use different information types to compute the score (RSV) of document *d* with respect to query *q*

- Anchor text: $P(d \mid q)$
- Page content: Okapi BM25
  - Indexing with word and bi-word
- Link structure: PageRank $P(d)$
  - Probability that a user surfing on the Web reaches document *d*

Combining different scores

- Scores computed by different information types have different meanings
- The final score of *d* is determined by a weighted harmonic mean of the ranks



Final score of *d*

$$\frac{1}{\alpha \times \frac{1}{3} + \beta \times \frac{1}{1} + \gamma \times \frac{1}{4}}$$

Reality is …

- The best performance was obtained with

$$\alpha = 0.8 \quad \beta = 0.2 \quad \gamma = 0$$

Anchor Content Structure

- Anchor text was definitely effective for Navigational Web Retrieval
- In the remaining of this talk, we focus only on exploiting anchor text

Exploiting anchor text

- Modeling anchor text
- Extracting synonyms from anchor text for query expansion

Modeling anchor text: Basis

- $P(d \mid q)$: probability that document d is the representative page for the item expressed by query q

$$\arg\max_d P(d \mid q) = \arg\max_d P(q \mid d) \times P(d)$$

Computation of $P(q \mid d)$ is crucial

$$\frac{\#\text{inlinks of } d}{\#\text{inlinks in collection}}$$

Computation of $P(q \mid d)$

- We assume independence of terms in q

$$P(q \mid d) = \prod_{t \in q} P(t \mid d)$$

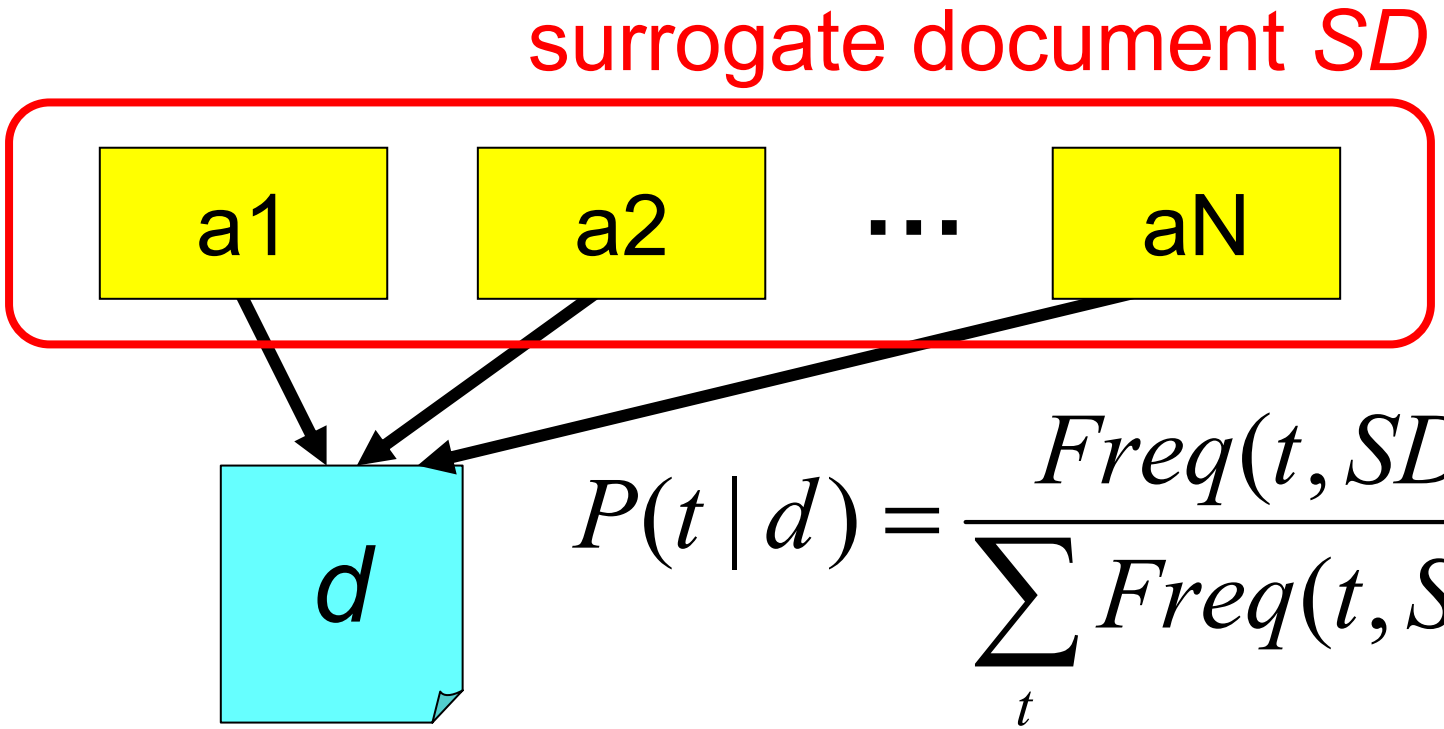
- ChaSen is used to extract term t
- If $P(t \mid d)$ is not modeled, we use $P(t)$ for smoothing

- We compare two alternative methods to model $P(t \mid d)$

First method: Document model

- Use all anchor texts linking to *d* as a single surrogate document for *d*

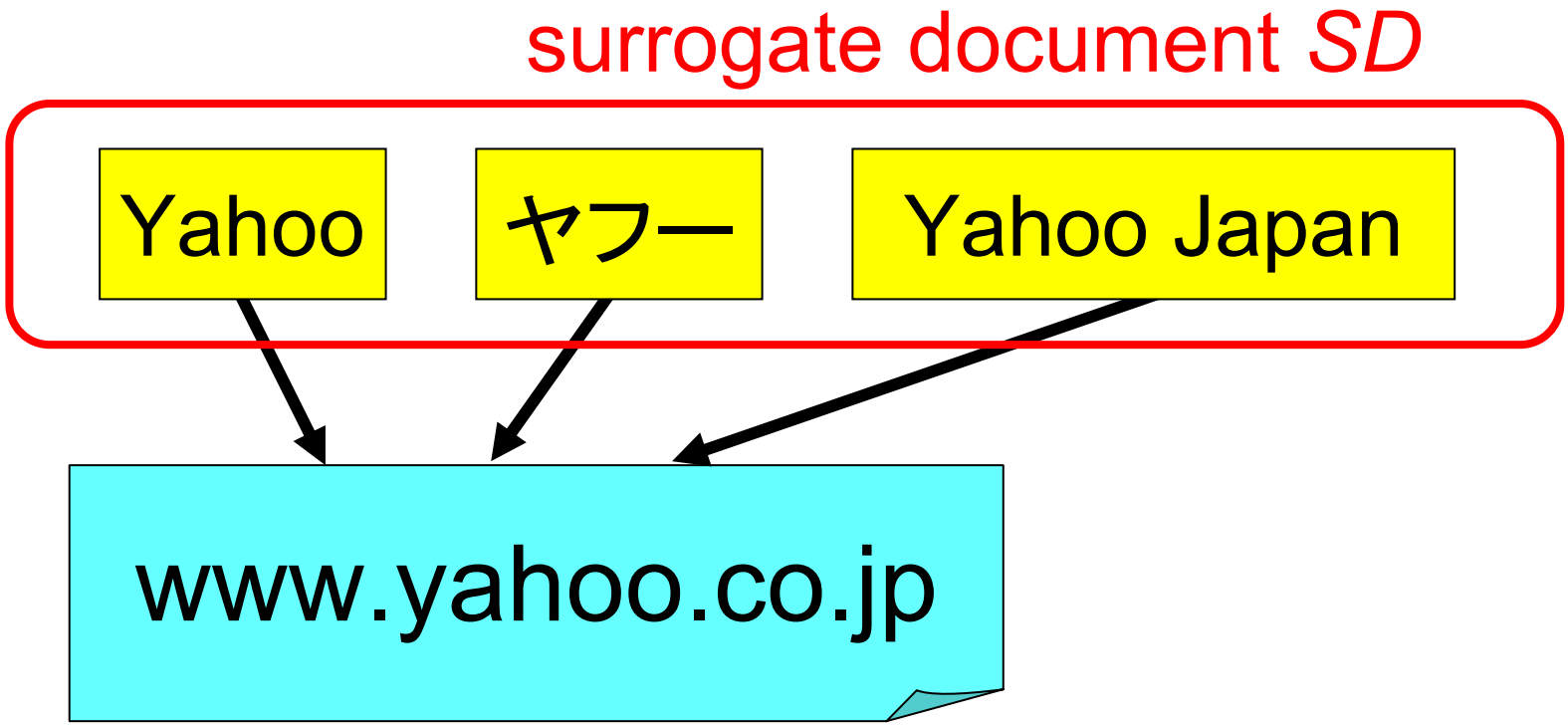
  [Westerveld et al., TREC 2001]



$$P(t \mid d) = \frac{Freq(t, SD)}{\sum_{t} Freq(t, SD)}$$

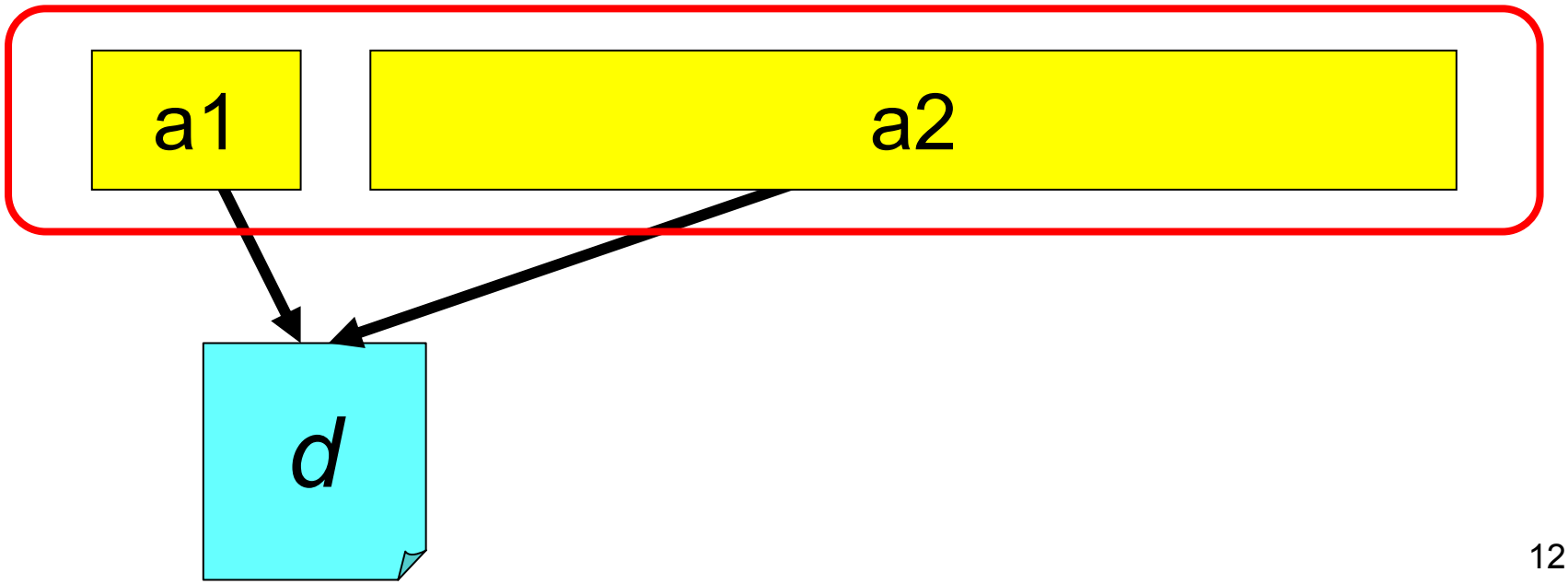
Problem of DM

- *P*(*t* | *d*) is same for ヤフー（*yafuu*） and “Japan”
- But, “Japan” is useless without “Yahoo”



Problem of DM (cont.)

- Document model is spammable
  - A user can change $P(t \mid d)$ purposefully
- Computation of $P(t \mid d)$ is affected by a2



Second method: Anchor model

- Use anchor texts linking to *d* independently to compute *P*(*t* | *d*)

$$P(t \mid d) = \sum_{a} P(t \mid a) \times P(a \mid d)$$

Probability of term *t* is normalized on an anchor-by-anchor basis

Exploiting anchor text

- Modeling anchor text
- Extracting synonyms from anchor text for query expansion

Query expansion in $P(t \mid d)$

- If $P(t \mid d)$ is not modeled, synonym term s is used

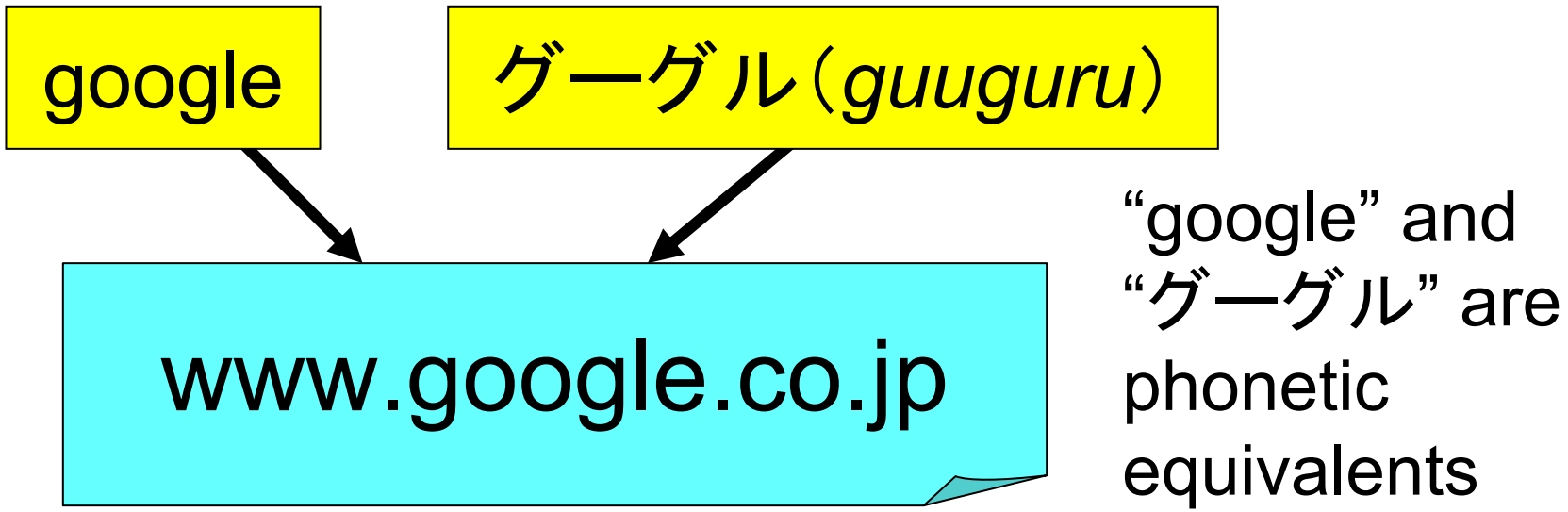
$$P(t \mid d) = P(t \mid s, d) \times P(s \mid d)$$
$$\approx P(t \mid s) \times P(s \mid d)$$

Probability that s is replaced with t

- We need synonym pairs to compute $P(t \mid s)$

Extracting synonym pairs

- Multiple anchor texts linking to the same document have same/similar meaning
- We extract phonetic equivalents (transliterations) from those anchor texts



Identifying phonetic equivalents

- Our transliteration method is used to determine whether two words are phonetic equivalents

  [Fujii and Ishikawa, CHUM 2001]

  google / グーグル（*guuguru*） → Yes

  google / エンジン（engine） → No

Example of query expansion

Topic ID	Source term	Expanded term
1041	UNESCO	ユネスコ
1097	エキサイト	excite
1131	ダンス	dance
	ディライト	delight
1138	トヨタ	Toyota
1172	ディレクトリ	directory

Evaluation result (TYPE=A)

	DCG-3-0	WRR-1-0
AM+Syn+C	2.522	0.605
AM+Syn	2.499	0.600
AM	2.464	0.596
DM+Syn	2.460	0.593
DM	2.431	0.590
C	0.381	0.080

Analysis

- We analyzed the results of AM+Syn+C by topic subcategories
  - Type
  - Category
  - Specialty

Topic Type (TYPE=A)

	#Topics	DCG-3-0	WRR-1-0
keyword	145	3.101	0.767
keywords	96	2.033	0.446
incomplete	28	1.383	0.356

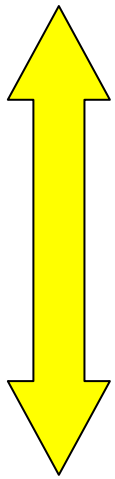
If a query can be expressed by a single keyword precisely, the performance was better than those for other cases

Topic category (TYPE=A)

	#Topics	DCG-3-0	WRR-1-0
product	49	2.256	0.540
company	60	3.071	0.717
person	29	2.376	0.517
facility	29	2.502	0.637
sight	16	2.206	0.649
resource	47	2.403	0.555
online shop	29	2.329	0.598
event	19	3.117	0.768

Specialty of a user (TYPE=A)

Specialty	#Topics	DCG-3-0	WRR-1-0
High	62	2.577	0.592
	106	2.720	0.632
	73	2.654	0.669
Low	28	1.594	0.435



Queries produced by specialists did not match with anchor texts produced by "general" people

Example of mismatch b/w specialist query and anchor text

- Topic 1063
  - query: Yahoo housing information
  - anchor text: Yahoo real estate

Conclusion

- Following methods were effective for Navigational Web Retrieval
  - Anchor model
  - Synonym-based query expansion
  - Combination of anchor and content retrieval

Please visit **WEB-7**