Information Retrieval Boosted by Category for Troubleshooting Search System

08/13/2015

Bin Tong, Toshihiko Yanase, Hiroaki Ozaki, Makoto Iwayama

Central Research Laboratory
Hitachi, Ltd.

1. Outline

HITACHI
Inspire the Next

◆ Background:

✓ Information extraction for troubleshooting system in maintenance activities.

✓ Except for titles and documents, rich domain-specific category codes are available.

◆ Motivation:

✓ In maintenance logs, much information is redundant.

✓ In faceted search, information restricted to selected categories is displayed. The information related to the selected categories is important but not displayed.

◆ Idea:

✓ Extend a text-summarization method with a word-category graph and a category-category graph.

✓ The frequencies in the two graphs influence the words' scores.

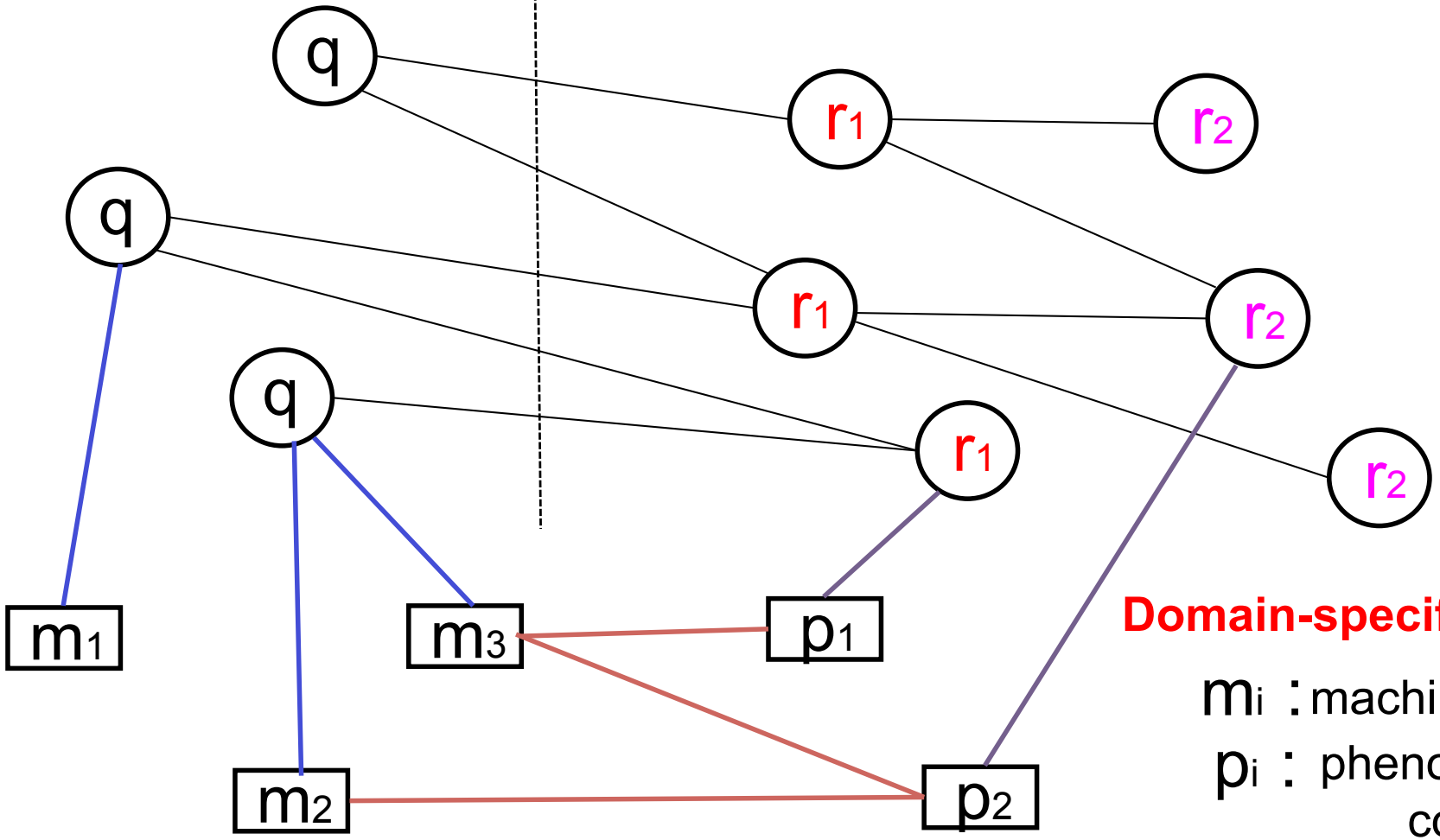
HITACHI
Inspire the Next

Titles

Hydraulic oil over heat

Documents

Hydraulic oil over heat cause fan pump damaged…
Check RMP fan motor.. Replace fan pump.



Domain-specific codes

m_i : machine mode

p_i : phenomenon code

3. Assumption:

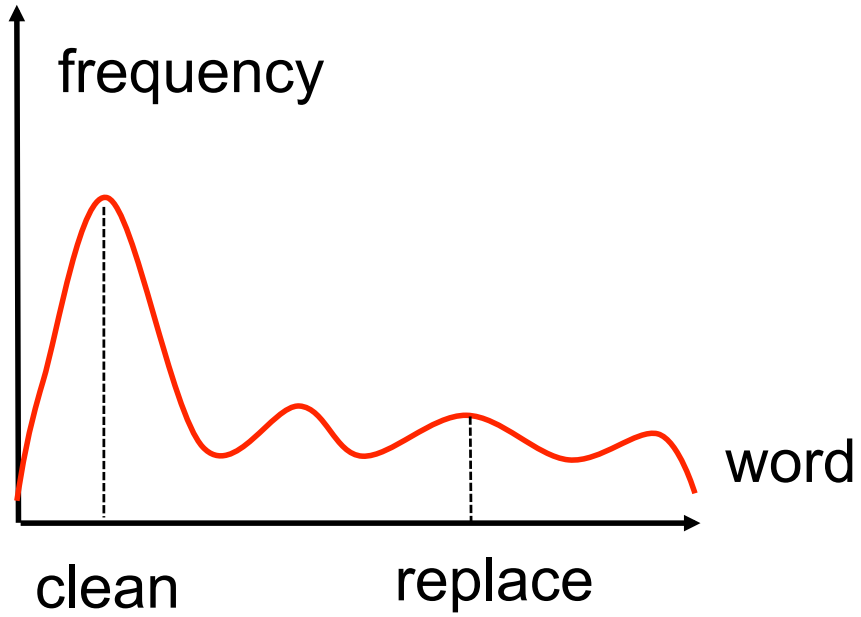
HITACHI
Inspire the Next

◆ The frequencies of words with respect to a category code are different.

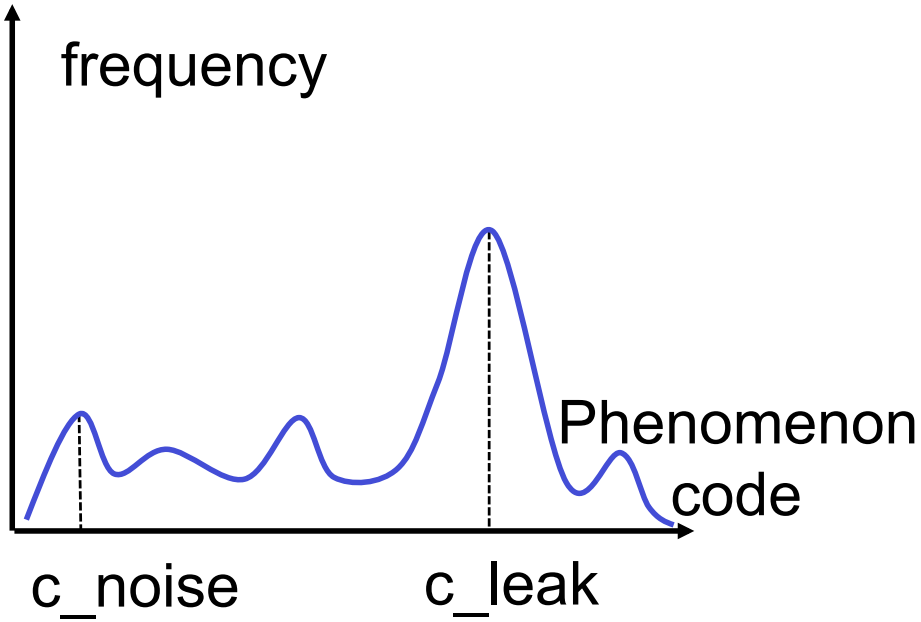
◆ The frequencies of categories with respect to a category code are different.

Mode Code: mc-1505

word-category graph



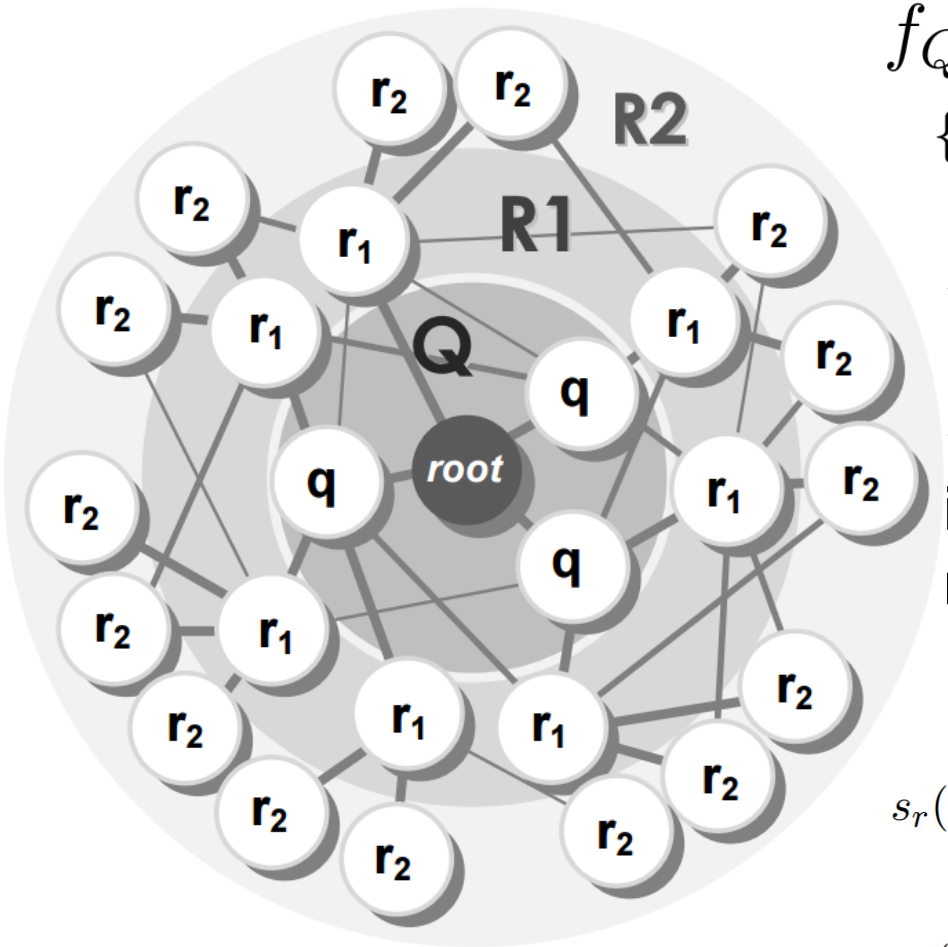
category-category graph



4. Baseline: Query Snowball (QSB)

HITACHI
Inspire the Next

A Text Summarization Method



$$f_{QSBP}(S) = \sum_{\{w_1,w_2|w_1 \neq w_2 \text{ and } w_1,w_2 \in u \text{ and } u \in S\}} s_p(w_1, w_2)$$

- ✓ The score of a sentence depends on scores of words.
- ✓ The score of a word is related to its frequency and closeness with respect to query words.

$$s_r(r1) = \sum_{q \in Q} s_b(r1) \left(\frac{s_b(q)}{sum_Q} \right) \left(\frac{freq(q, r1)}{distance(q, r1) + 1.0} \right)$$

$$s_r(r2) = \sum_{r1 \in R1} s_b(r2) \left(\frac{s_r(r1)}{sum_{R1}} \right) \left(\frac{freq(r1, r2)}{distance(r1, r2) + 1.0} \right)$$

5. Our approach: Category-based QSB

HITACHI
Inspire the Next

◆ The frequencies of query words with respect to a category code influence the score of a document word.
◆ The frequencies of categories with respect to a category code influence the score of a document word.

$$cqsb(w) = qsb(w)\exp(\lambda \cdot s_{ctg}(w))$$

category-based factor

Example:

$$s_{ctg}(r_1) = s_{wc}(r_1, Q^{r_1}_{C_M}) + s_{cc}(r_1, Q^{r_1}_{C_M})$$

$$s_{wc}(r_1, Q^{r_1}_{C_M}) = \sum_{q \in Q^{r_1}_{C_M}} \sum_{i=1}^{|C^q_{MN}|} \theta \frac{freq(c_i, q)}{freq(c_i)}$$

$$s_{cc}(r_1, Q^{r_1}_{C_M}) = \sum_{q \in Q^{r_1}_{C_M}} \sum_{\theta \in \Gamma} \sum_{c_i, c_j} \theta \frac{freq(c_i, c_j)}{freq(c_i)}$$

$$\Gamma = \{\beta, 1-\beta\}$$
$$c_i \in C^{r_1}_J$$
$$c_j \in C^q_M$$

6. Experiment

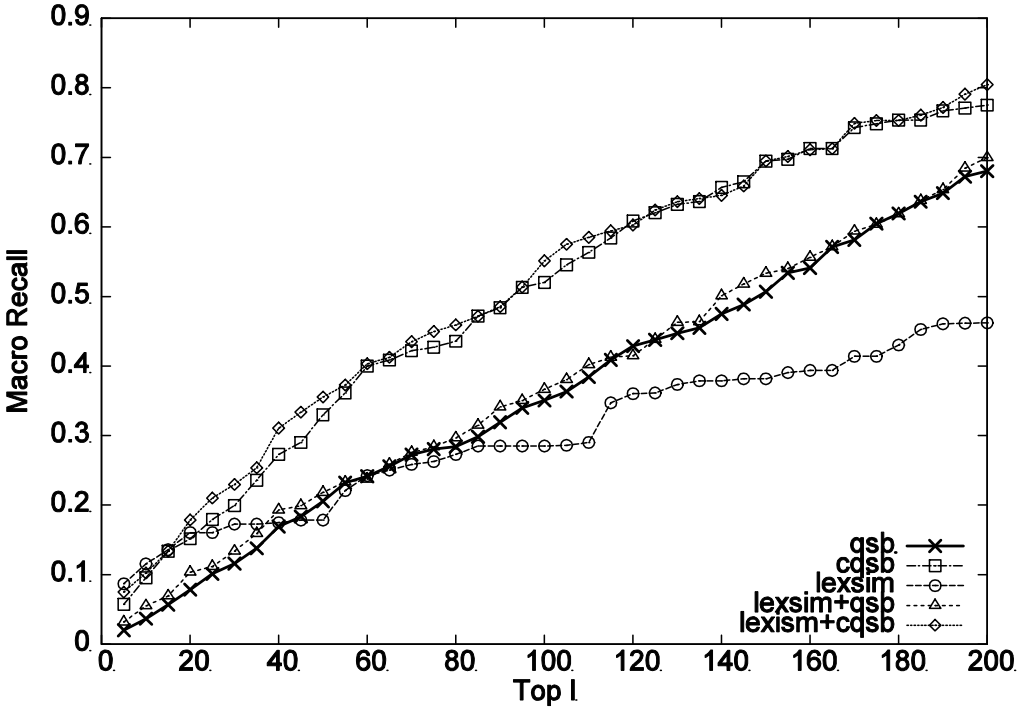
HITACHI
Inspire the Next

- **Data:**
  - ✓ Maintenance logs of construction machines
  - ✓ A title set and a document set are included.
  - ✓ A title is generally a problem statement. Each title corresponds to only one document, in which solutions of the problem are described.
  - ✓ **Domain-specific codes:** machine code, trouble code, phenomenon code, and countermeasure code.
  - ✓ Use the data of four dominated trouble codes.
- **Goal:**
  - ✓ Given a query and selected category codes, extract the most **informative** sentences from documents, which are useful for solving the problem a query states.
- **Evaluation:**
  - ✓ Macro Recall, Mean Average Precision (MAP), and F-score.
  - ✓ Recall is important in trouble shooting.

7. Experiment Results:

HITACHI
Inspire the Next

◆ The change of Macro Recall when increasing the number of the top l ranking sentences.



. Experiment Results

HITACHI
Inspire the Next

◆ The F_3 scores and MAPs when the top l ranking sentence is set to be 100.

Table 1: F_3 scores in trouble code 03 data set (best case)

Methods	Macro Recall	MAP	F_3 score
lexsim+cqsb	.5513	.0690	.3244
lexsim+qsb	.3665	.0407	.2036
lexsim	.2849	.0957	.2379
cqsb	.5200	.0609	.2964
qsb	.3503	.0312	.1731

Table 2: F_3 scores in trouble code 05 data set (worst case)

Methods	Macro Recall	MAP	F_3 score
lexsim+cqsb	.4645	.0557	.2679
lexsim+qsb	.3893	.0383	.2031
lexsim	.3431	.1346	.2971
cqsb	.4128	.0370	.2049
qsb	.3746	.0299	.1739

HITACHI

Inspire the Next