Study on the Combination of Probabilistic and Boolean IR Models for WWW Documents Retrieval

Masaharu YOSHIOKA　Hokkaido Univ.
Makoto HARAGUCHI

Background and Objectives

- Background
  - Requirement for IR system with large scale text data
  - Different IR models
    - A probabilistic model
      - The user may not select query term appropriately.
    - A Boolean model
      - The user must select query term appropriately.
      - A Boolean query formula is expressive but is very difficult to construct appropriate one.
- Objectives
  - Evaluate following IR systems.
    - our IR system, which is based on the probabilistic IR model.
    - our method for combining probabilistic and Boolean IR models for clarifying queries.

Index for our IR system

- Word and phrasal index
  - Use ChaSen as morphological analyzer and select noun (noun, unknown, symbol) for word index
  - Phrasal index: a pair of adjacent noun terms
    - We use prefixes, postfixes, and numbers in addition to words that are used for word index

Input: 道路交通法について

ChaSen

道路	noun
交通	noun
法	noun-prefix
について	postpositional particle

Word index 道路, 交通

Phrasal index
!c道路交通, !c交通法

Our IR System (a Probabilistic IR Model)

- Modified version of OKAPI
  - Use BM25 formula to calculate each document score

$$\sum_{T \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf}$$

$$K = \frac{document\ length}{average\ document\ length}$$

$$w^{(1)} = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)}$$

tf : frequency of T in a document
qtf : frequency of T in a query
k_1,k_3: parameter（k_1=1,k_3=1000 (initial) or 7 (final))
N: :the count of all documents in the database,
n: the count of all documents containing T
R: the given number of relevant documents
r : the count of all relevant documents containing T

  - Term weighting for phrasal terms
    - Document score may differ according to the dictionary entry

情報処理→ Word 情報処理
情報科学→ Word 情報, 科学 Phrase !c情報科学

    - Discount score for phrasal terms

$$qtf = c * qtf_c$$

qtf_c : frequency of phrase T in a query
c : parameter (c≤1;c=0.3)

Relevance Feedback

- Relevance feedback
  - Pseudo-relevance feedback
    - Use top 5 ranked documents of initial retrieval are used as relevant documents.
      - Reject documents with small number of terms in it.
  - Query expansion
    - Use terms in relevant documents as query terms
      - Max: 300 terms
    - Rocchio-type feedback

$$qtf = \alpha * qtf_0 + (1-\alpha) * \frac{\sum_{i=1}^{R} qtf_i}{R}$$

qtf_0 : frequency of T in a initial query
qtf_i : frequency of T in a i-th relevant documents
R: the given number of relevant documents
α : parameter (α=0.7)

Implementation of Our IR System

- Text normalization
  - Use cooked data
  - Remove tag such as <NWD:img>
  - All alphabet and number are converted with ASCII character
  - Remove "ー" at the end of katakana words
- Database Engine
  - Generic Engine for Transposable Association (GETA)
  - Divide texts into 8 database of GETA
  - Merge results after retrieving documents from all databases

Evaluation of Our IR System (a Probabilistic IR Model)

- Retrieval Performance (Debugged-System)
  - Use “S” and “A” documents as relevant one
- Better performance in a submitted system
- Problem of pseudo-relevance feedback
  - Similar template generated page may take similar score → Too much biased with relevant page

	AvePrec	RPrec	Prec@10	Prec@20
tt (survey)	0.223	0.254	0.411	0.361
tt (target)	0.218	0.242	0.374	0.330
ds (survey)	0.200	0.234	0.383	0.341
ds (target)	0.220	0.239	0.380	0.337

Characteristics of IR models

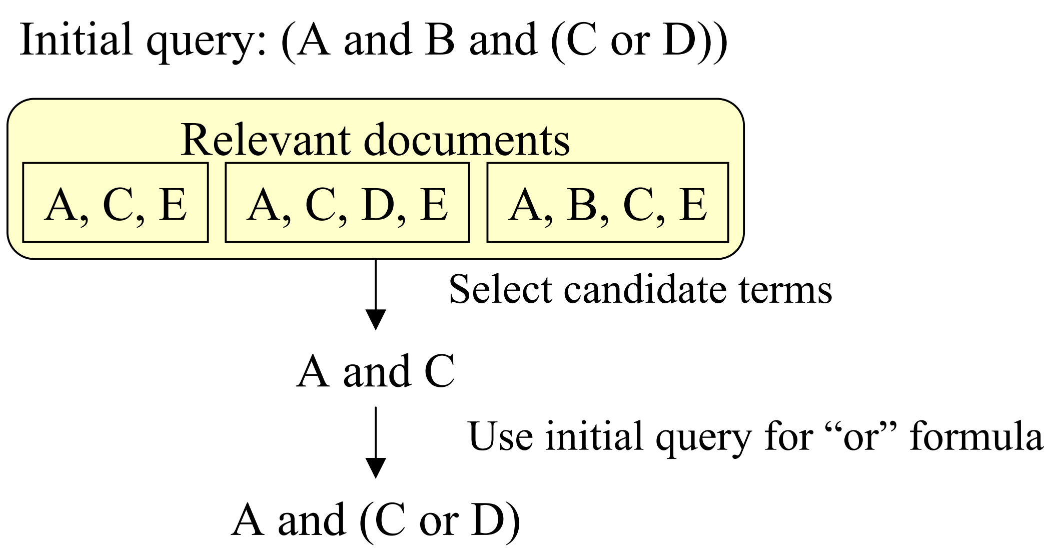
- A probabilistic model
  - The user may have difficulties to select appropriate query terms.
    - Documents that do not contain a part of query terms may select as higher relevant ones.
  - The system can represent users' retrieval intention by using a large number of query terms that includes words with higher cooccurrence
    - Difficulties to understand appropriateness of query
- A Boolean model
  - The user can select appropriate query terms.
    - Documents that do not satisfy a Boolean formula is not selected
  - Limited number of required query terms are used
    - Higher readability
    - The user can easily understand why the IR system selects the documents

Problem on a Boolean IR model

- Retrieval performance of a Boolean IR model is worse than a probabilistic one
  - A Boolean query formula is expressive but is very difficult to construct appropriate one.
- Requirement for a Boolean query construction support
  - Use relevant documents for clarifying a Boolean query formula
    - Initial document retrieval without using a Boolean IR model
    - Relax a Boolean query formula by using relevant documents

Reconstruction of a Boolean Query Formula

- Relax an initial Boolean query formula to include given relevant documents as relevant one
  - Use terms that exists in all relevant documents and also exists in an initial query as a candidate to construct a relaxed Boolean query formula
  - Use an initial query for “or” formula



Combination of Probabilistic and Boolean IR Models

- Two approach
  - Use a Boolean IR model first and calculate score of each retrieved document by using a probabilistic model
  - Use a probabilistic IR model first and apply penalty for documents that do not satisfy a Boolean query formula
    - Penalty is calculated by using term importance in BM25

$$\beta \times w^{(1)} \times \frac{(k_3+1)qtf}{k_3+qtf} \qquad \beta : \text{parameter}$$

    - Penalty is calculated for each "and" element
    - For "or" formula, use penalty of a term that has highest one among them.

Evaluation of the Boolean Reconstruction

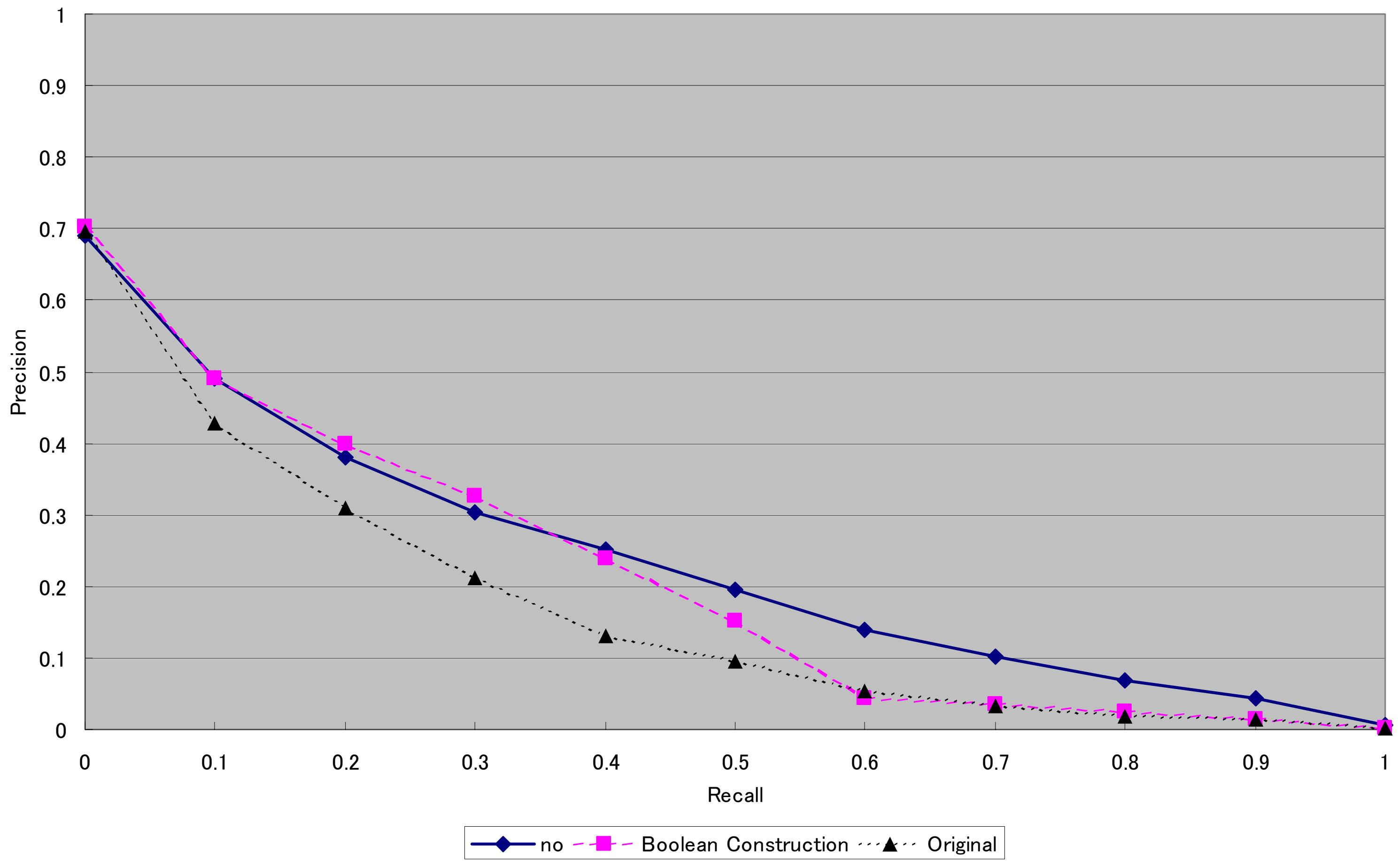
- Retrieval Performance
  - Use "S" and "A" documents as relevant one
  - Original boolean query formula is not appropriate one
  - Poorer performance than a probabilistic IR model

	AvePrec	RPrec	Prec@10	Retrieved
tt-b (survey)	0.200	0.236	0.431	1843
tt-b (target)	0.210	0.247	0.398	3294
tt-o (survey)	0.153	0.184	0.374	1685
tt-o (target)	0.183	0.216	0.381	3075
ds-b (survey)	0.155	0.196	0.370	1327
ds-b (target)	0.192	0.224	0.388	2493

"-b": reconstructed Boolean query formula
"-o": original Boolean query formula

Evaluation of the Boolean Reconstruction

- tt (survey)
  - Improve performance in lower recall



Evaluation of the Boolean Penalty

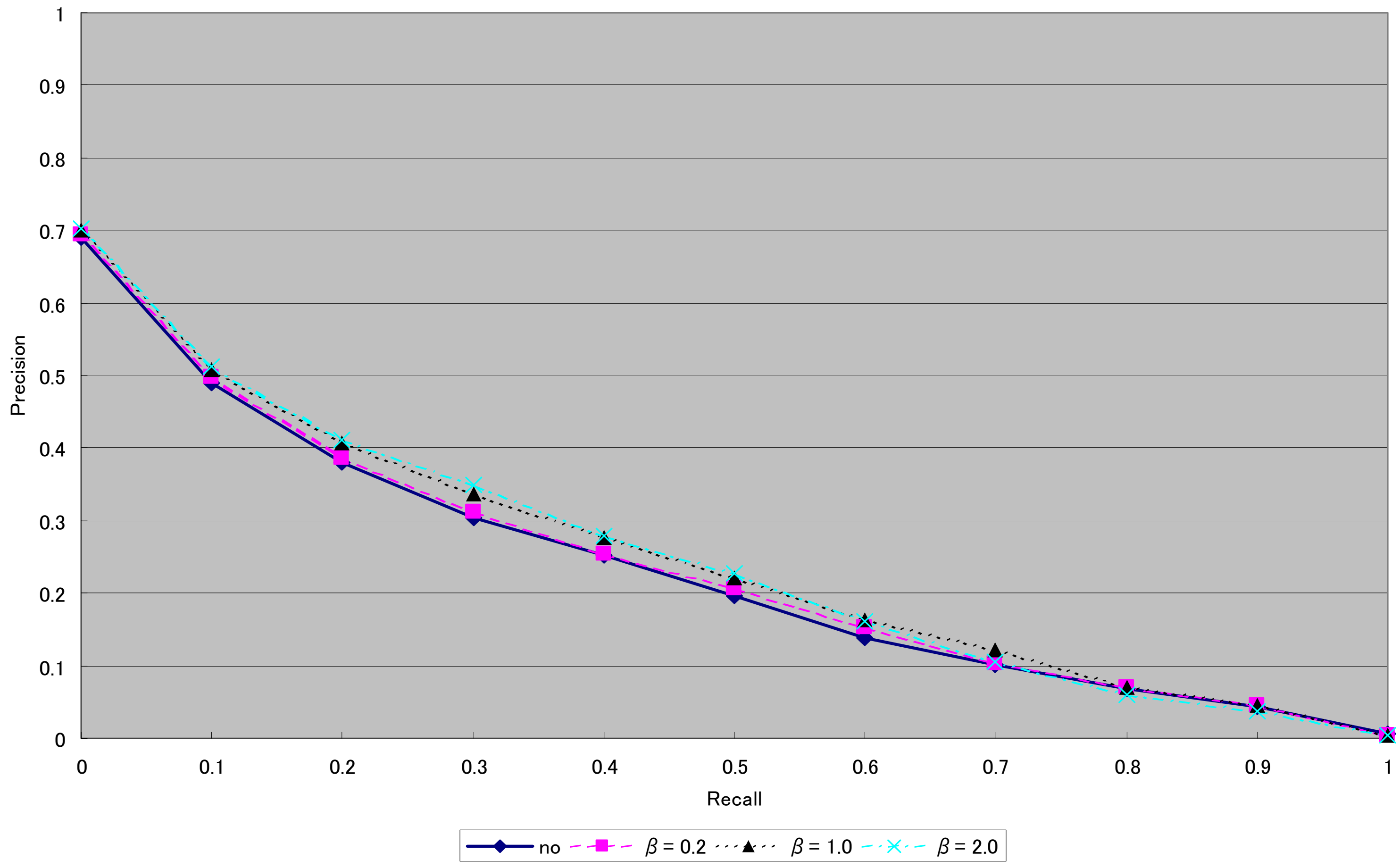
- Use reconstructed Boolean Query formula
- Retrieval Performance
  - Use “S” and “A” documents as relevant one

	AvePrec	RPrec	Prec@10	Prec@20
tt-1.0 (survey)	0.241	0.263	0.431	0.376
tt-1.0 (target)	0.239	0.259	0.394	0.348
tt-2.0 (survey)	0.241	0.265	0.429	0.380
tt-2.0 (target)	0.241	0.260	0.389	0.348
ds-1.0 (survey)	0.218	0.242	0.389	0.346
ds-1.0 (target)	0.238	0.251	0.385	0.341
ds-2.0 (survey)	0.211	0.237	0.394	0.346
ds-2.0 (target)	0.234	0.251	0.388	0.341

“-1.0”: $\beta = 1.0$
“-2.0”: $\beta = 2.0$

Evaluation of the Boolean Penalty

- tt (survey)
  - Improve performance almost all recall value



Conclusion

- A proposal of our IR system based on a probabilistic IR model
  - We confirm the system has better performance in NTCIR-4 submission.
  - This system may be good enough to use as a benchmark system.
- A proposal of a combination of two IR models
  - User defined Boolean query is not precise enough to retrieve all relevant documents
  - Relaxing an initial Boolean query formula by using relevant documents improve quality of a Boolean query formula
  - Penalty calculation by using a Boolean query formula improves retrieval performance