

On a Combination of Probabilistic and Boolean IR Models for GeoTime Task

Masaharu Yoshioka e-mail: yoshioka@ist.hokudai.ac.jp
Graduate School of Information Science and Technology, Hokkaido University

Motivation

Information retrieval for Question Answering about a particular named entity

Documents that do not contain information about the entity are irrelevant.

Query: When Paul Nitze die?

Paul
McCartney
..... Paul
.....
die...Paul...
death...

..... Paul Nitze
.....
die.....
.....

..... Nitze
.....
.....

Score of IR System

With Partial match

IR for QA

O

O

x

x

O

x

ABRIR (Combination of Boolean IR model and Probabilistic IR model)

IR System (Probabilistic IR Model)

- Modified version of OKAPI

- Use BM25 formula to calculate each document score

$$w^{(1)} = \log \frac{\sum_{T \in Q} \frac{(k_1 + 1)tf}{K + tf} \frac{(k_3 + 1)qtf}{k_3 + qtf}}{(n - r + 0.5)/(N - n - R + r + 0.5)}$$

- 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 index

$$qtf = c * qtf_c$$

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

Experimental Results

Runs	Boolean for NE	Boolean for others	
JA-JA-01-D	Filter	Filter	
JA-JA-02-DN	Filter	Filter	
JA-JA-03-D	Penalty	Penalty	
JA-JA-04-D	Penalty	No	
JA-JA-05-D	No	No	Baseline:Okapi

run	01-D	02-DN	03-D	04-D	05-D
AP	0.3697	0.3867	<u>0.3719</u>	0.3627	0.2881
nDCG	0.4117	0.4268	<u>0.4162</u>	0.4072	0.3282
Q	0.5710	0.5685	0.5881	0.5717	0.4993

Statistical Significance Test between 03-D and 05-D (Baseline: Okapi)

Statistically significant:

t-Test for a two-sided tests :nDCG(0.018) and Q(0.040)

Wilcoxon Signed Rank test: AP(0.0015), nDCG(0.0006) and Q(0.0024)

Modification of ABRIR for QA

Usage of verb as index terms

- Verb synonym set construction by using EDR

Handling named entities

- Identification of named entity is important for constructing a Boolean query.

Varieties of representation of foreign people in Japanese Katakana

Number of relevant documents

- There may be only a few relevant documents for a query.

Number of query expansion terms

- Large number of query expansion terms may cause concept drift for QA

Flow chart of Query Construction

0. Initial Query



女優のオードリ・ヘップバーンが
亡くなったのはいつですか？
(When did the actress Audrey Hepburn died?)

1. Remove question part of the query



女優のオードリ・ヘップバーンが亡くなった
(the actress Audrey Hepburn died)

2. Morphological analysis and NE tagging



NE: オードリ・ヘップバーン
Keywords and types
女優(actress)
オードリ(Audrey)
ヘップバーン(Hepburn)
亡くなる(die)
NE
NE
verb

NE Tagger:
Cabocha
Morphological
Analysis:
Mecab

3. Generation of synonym and variation list



オードリ:オドリ
ヘップバーン:ヘブバーン, ヘップバン,
ヘップバーン
亡くなる:死ぬ, 死亡, ...

EDR for
synonym list
construction

4. Initial retrieval



Query
女優, オードリ, ヘップバーン, 亡くなる

5. Comparison between query and pseudo relevant documents



女優: All documents
オードリ:オードリ
ヘップバーン:ヘップバーン, ヘブバーン
亡くなる:亡くなる, 死ぬ

Probabilistic IR
model
Usage of 3
pseudo relevant
documents

6. Construction of Boolean Query



女優 and オードリ and (ヘップバーン or
ヘブバーン) and (亡くなる or 死ぬ)

Usage of 5 terms
with high MI from
pseudo relevant
documents

7. Query expansion



女優, オードリ, ヘップバーン, ヘブバーン,
亡くなる, 死ぬ, ローマ(Rome), 休日
(Holiday), ...

Combination of
Boolean and
Probabilistic IR

8. Final Retrieval



Conclusion

- Proposal of using ABRIR as an IR system for question and answering for particular named entities.
- From the evaluation experiment, we confirm that ABRIR can make appropriate Boolean query and penalty based system outperform the baseline system (probabilistic IR model: Okapi BM25).