Overview of the NTCIR-5 WEB Query Term Expansion Subtask

Masaharu Yoshioka

Hokkaido University

National Institute of Informatics

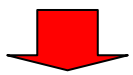
Background

- Query term expansion
  - Difficulties for selecting appropriate query terms to represent the information need
  - Estimation of useful query terms for filling the gap between the given query terms and the information need
- Evaluation of query term expansion technique
  - Usage of test collection
    - Most of query term expansion techniques improve retrieval results in general

However, the effectiveness of this technique depends on the quality of query terms in the initial query

Background (continue)

- Analysis on Topic difficulties
  - Analysis on NTCIR-1 test collection (Eguchi,2002)
    - Correlation between topic difficulties and average of each initial query term's IDF
  - Clarity measure based on a language model (Cronen-Townsend et al., 2002)
    - Identifying ambiguous or ill-formed queries
    - Usage of this measure to decide whether the initial query terms requires query term expansion or not



However, these approaches do not deal with the gap between initial query terms and information need directly

Objectives

- Proposal of an evaluation framework of query term expansion technique
  - Analysis on the several features that affect the quality of the techniques
    - Focusing on the variation of initial query term types.
      - Mismatch between the initial query and relevant documents
- New approach for evaluation of the techniques
  - From evaluation in average to evaluation of strong and weak topic types

Why Query Term Expansion Works (Buckley, 2004)

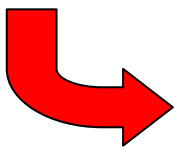
- (Buckley, 2004) hypothesized a possible reason why query expansion improves the query performance.
  - **one or two good alternative words to original query terms (synonyms)**
  - **one or two good related words**
  - **a large number of related words that establish that some aspect of the topic is present (context)**
  - **specific examples of general query terms**
  - **better weighting to original query terms**
- First 4 reasons are related to the query term expansion technique

Question: All topics requires each type of expansion terms in same ratio?

No

Mismatch between the Query Terms and Relevant Documents

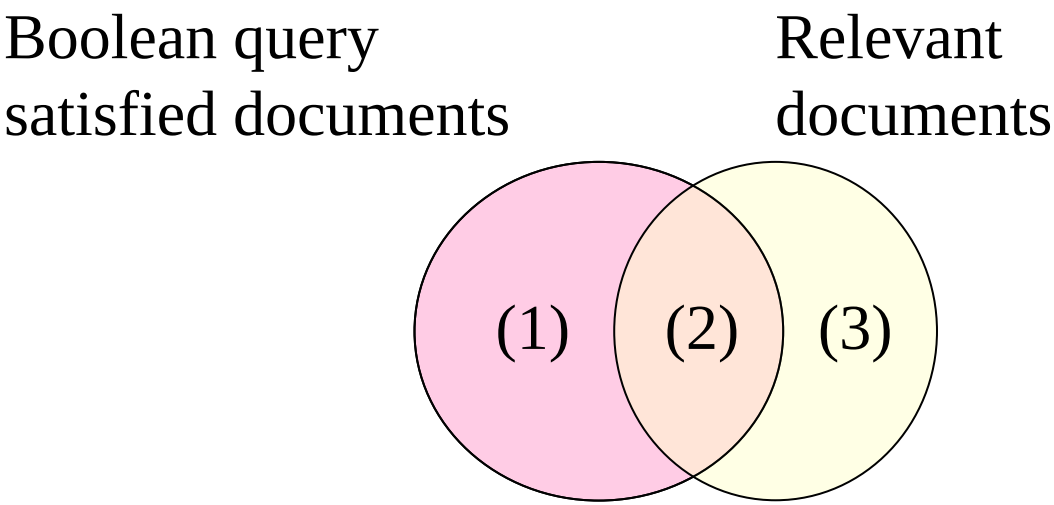
- Existence of relevant documents that do not contain a part of initial query terms
  - In NTCIR-4 Web test collection survey type
    - Almost 40% relevant documents do not satisfy title query (2-3 query terms with simple Boolean expression)
    - Many queries are not precise enough to narrow relevant document candidates
- Quality of initial query term sets affect on the quality of the query expansion technique



Evaluation of each query term sets based on this mismatch

Analysis on Mismatch between the Initial Query and Relevant Documents

- Analysis based on Boolean IR model
  - Comparison between documents that satisfy a initial Boolean query and relevant documents



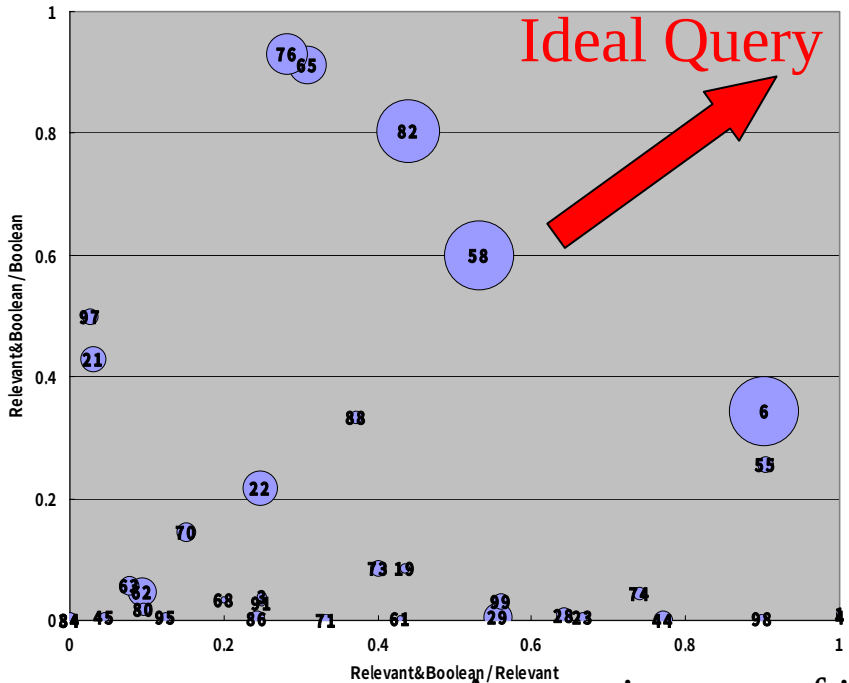
Ideal query: Both (1) and (3) are empty set
Large (1) = Ambiguous : needs context terms
Large (3) = Too strict : needs alternative terms

An Example of the Mismatch

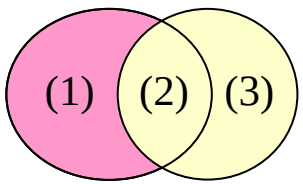
- NTCIR-4 Web test collection
  - Title field with Boolean expression for 35 Survey type topics
  - Index of ABRIR (organizer reference system) is used for this calculation

(2)/{(1)+(2)}

Appropriateness of Initial query quality from the viewpoint of precision



Boolean Relevant Documents



(2)/{(2)+(3)}

Appropriateness of initial query quality from the viewpoint of recall

Feature Quantities for Evaluating Effectiveness of the Query Term

- Feature quantities for both initial query terms and query expansion terms
  - Appropriateness of the term that characterizes the relevant documents
- Feature quantities for query expansion terms
  - Appropriateness of the alternative term for each initial query term
  - Appropriateness of the context definition term for the query

Appropriateness of the term that characterizes the relevant documents

- Comparison between the characteristic terms of relevant documents and query terms
  - Mutual information content between relevant documents *r* and the term *w*.

$$G'(w) = p(w \mid r) \log_2 \frac{p(w \mid r)}{p(w)}$$

  - The characteristic terms are terms with higher *G'(w)*.
    - A initial query term with higher *G'(w)* means good initial query term.
    - A query expansion term with higher *G'(w)* means good query term for defining context.

Feature quantities for query expansion terms

- A good alternative term
  - should exist for relevant documents that do not contain the initial query term.
  - Therefore, the number of documents that have a query expansion term and do not have an initial query term is useful for evaluation.
- A good term for context definition
  - is a distinct term that exists in relevant documents.
  - Therefore, the number of documents that have a query expansion term in the relevant documents, the Boolean satisfied documents, and total documents is useful for evaluation.

Retrieval Experiments

- Test collection
  - NTCIR-4 Web test collection survey type topics: 35 topics
  - Title field (2-3 query terms with simple Boolean expression)
  - Additional relevance judgment in almost same way in NTCIR-4
    - There were several submission results whose top-ranked documents are not included in the judged document list
- Type of retrieval experiments
  - Number of query expansion terms
    - No query term expansion
    - query term expansion
      - Limited number (10)
      - No limitation
  - Feedback Type
    - Pseudo-relevant documents
    - User selected relevant documents
      - Each participant uses relevant document for simulating document selection

List of Participants

- JSWEB
  - Experimented with relevant document vectors that were generated based on the existence of the keyword in the relevant documents. They also proposed combining relevant document vectors (one from the user selected relevant documents and the other from the pseudo-relevant documents). The retrieval method was based on a vector space IR model.
- NCSSI
  - Experimented with a clustering technique for the initial retrieval results and a named entity recognition technique for selecting query expansion terms from the appropriate cluster (user selected cluster or pseudo-relevant cluster). They used an organizer reference model, ABRIR, based on a probabilistic model as an IR system.
- R2D2
  - Experimented with Robertson's Selection Value (RSV) for selecting query expansion terms using pseudo relevant documents. The retrieval method was based on the modified Okapi. They also used link information for scoring the retrieved documents.
- ZKN
  - Experimented with Larvenko's relevance model for selecting query expansion terms using pseudo relevant documents. The retrieval method was based on the inference network and language model.
- ABRIR: Organizer Reference System
  - Experimented with mutual information between terms and relevant documents for selecting query expansion terms. The retrieval method was based on the Okapi.

Overall Evaluation Results of the Experiment

- Most of the cases, the retrieval results improve in average.
- However, there is no system that improves all topics.

	type of feedback	Number of topics where performace improve			10 query terms expansion (average)			No query term expansion (average)		
		Average Precision	R-Precision	Relevant Retrieved	Average Precision	R-Precision	Relevant Retrieved	Average Precision	R-Precision	Relevant Retrieved
JSWEB-auto-01	automatic	2	1	0	0.011	0.0236	212	0.0743	0.0992	1512
JSWEB-auto-02	automatic	3	2	0	0.0197	0.0344	516	0.0743	0.0992	1512
JSWEB-auto-03	automatic	17	15	11	0.0714	0.1094	1101	0.0743	0.0992	1512
NCSSI-auto-01	automatic	22	17	23	0.1708	0.2107	2432	0.1511	0.1991	2256
NCSSI-auto-02	automatic	21	12	17	0.1536	0.1962	2322	0.1511	0.1991	2256
R2D2-auto-01	automatic	19	15	21	0.1747	0.2239	2257	0.162	0.2066	2155
R2D2-auto-02	automatic	19	19	21	0.181	0.2236	2257	0.162	0.2066	2155
ZKN-auto-01	automatic	22	17	11	0.1523	0.2011	2139	0.1405	0.1839	2152
ZKN-auto-02	automatic	22	18	14	0.1537	0.1968	2153	0.1405	0.1839	2152
ABRIR-auto	automatic	28	20	25	0.2198	0.2506	2591	0.169	0.2085	2422
JSWEB-relevant-B-02	user	7	5	5	0.0235	0.049	755	0.0743	0.0992	1512
JSWEB-relevant-B-03	user	18	18	13	0.0976	0.1466	1453	0.0743	0.0992	1512
NCSSI-user-01	user	27	18	17	0.2434	0.2705	2508	0.173	0.2258	2353
NCSSI-user-02	user	28	15	16	0.2196	0.2487	2415	0.173	0.2258	2353
ABRIR-user	user	32	18	20	0.2569	0.2834	2689	0.1801	0.2268	2469

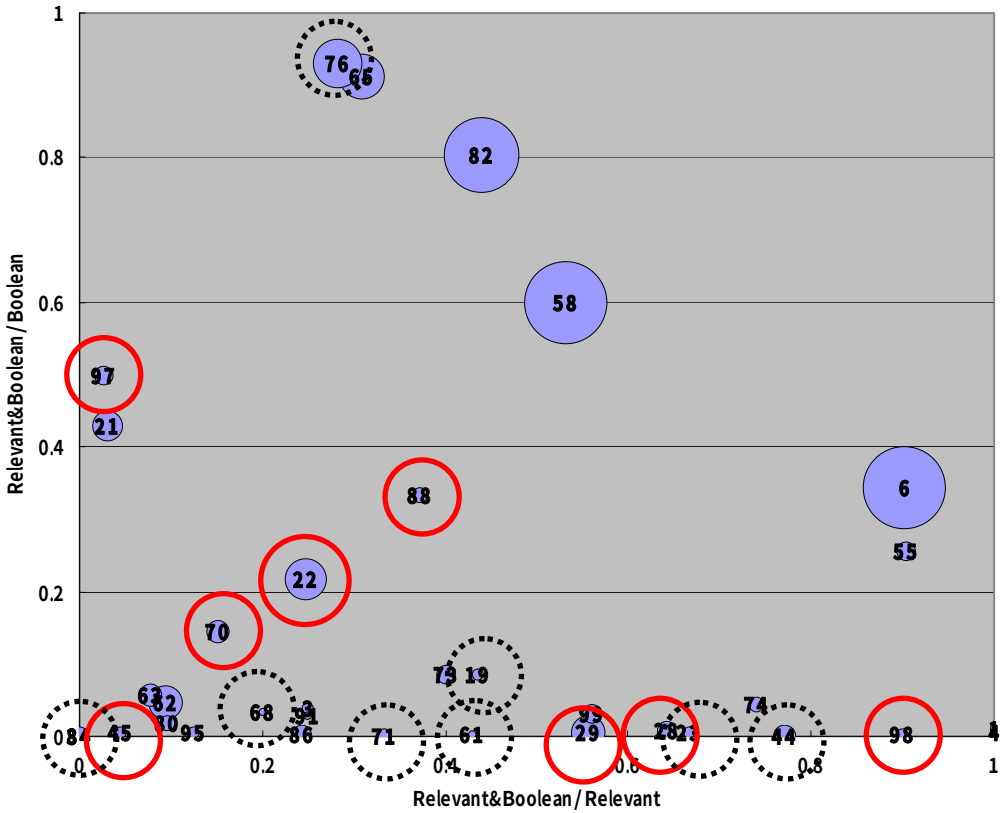
Topics where query term expansion works well

- Comparison between 10 query term expansion and no expansion
  - Count number of runs where expansion results is better than no expansion one
    - Average precision (AP)
    - R-Precision (RR)
    - Relevance Retrieved (RR)

	automatic			user		
	AP	RP	RR	AP	RP	RR
22	8	7	6	5	4	5
97	7	6	4	5	5	5
29	4	3	8	5	5	5
45	7	6	8	3	3	3
70	5	6	7	4	3	4
28	5	4	4	5	5	5
88	8	6	5	4	2	2
98	6	4	6	4	3	4
65	5	5	4	4	4	4
34	7	5	6	3	1	1
55	4	5	4	4	4	2
62	6	6	6	2	1	1
73	6	7	0	4	4	1
6	5	3	5	3	2	3
58	7	5	7	1	0	1
4	9	7	0	3	1	0
82	6	4	6	2	1	1
91	10	2	0	5	1	2
74	5	4	1	4	4	1
86	3	2	8	2	2	2
1	6	3	2	4	3	0
63	3	4	6	2	1	2
21	5	4	5	1	1	1
3	4	3	2	4	2	1
80	5	2	3	3	1	2
99	5	3	0	3	3	2
95	5	1	3	3	0	3
76	2	4	7	0	0	0
68	5	2	1	3	1	0
23	4	3	1	2	0	1
71	2	0	5	1	1	2
19	2	1	1	2	1	0
84	0	1	0	4	1	1
61	0	0	3	2	0	1
44	1	1	1	2	0	0

Topics where query term expansion works well

- No direct correlation between mismatch and effectiveness of the query term expansion

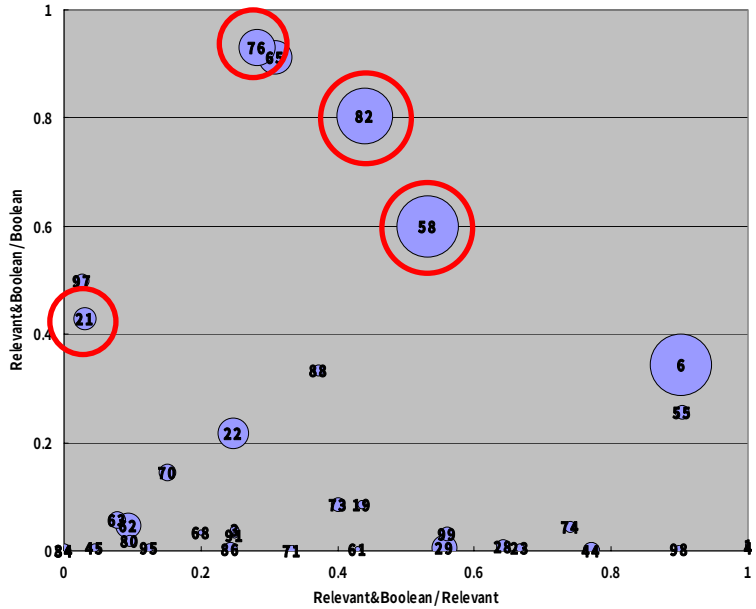


Topic where many number of performance measures improve

Topic where small number of performance measures improve

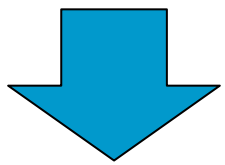
Effect of Feedback Documents Quality

- Comparison between automatic feedback and user feedback
  - Existence of topics whose retrieval results for user feedback is lower than automatic one
    - Run-Id: 0021, 0058, 0076, 0082
    - Those topics have higher *R&B/B* values compared with other topics
    - Good query expansion terms for these topics are terms that can be used as alternative terms for the initial query terms.
  - This result shows that non-relevant documents may be useful for finding alternative terms for initial query terms.



Discussion

- Topic types characterized by mismatch between the initial query and relevant documents may affect the performance of the query term expansion technique.
  - Parameter tuning for query term expansion technique may affect the mixed ratio of topic types.



For further analysis on the query term expansion, it is necessary to pay attention this issue.

Summary

- Proposal a new evaluation framework for query expansion technique
  - Topic type classification based on the mismatch between initial query term set and relevant documents
  - Enumeration of feature quantities that may affect the performance of query term expansion technique
- Further analysis is necessary for evaluating this framework

Errata

- Data for zkn with no query term expansion has a problem. Please replace following tables in the proceedings and “Summary of evaluation results” in a CD-ROM data.

Table 1in the proceedings
Summary of Evaluation Results (Page 5) in CD-ROM

	type of feedback	Number of topics where performace improve			10 query terms expansion (average)			No query term expansion (average)		
		Average Precision	R-Precision	Relevant Retrieved	Average Precision	R-Precision	Relevant Retrieved	Average Precision	R-Precision	Relevant Retrieved
JSWEB-auto-01	automatic	2	1	0	0.011	0.0236	212	0.0743	0.0992	1512
JSWEB-auto-02	automatic	3	2	0	0.0197	0.0344	516	0.0743	0.0992	1512
JSWEB-auto-03	automatic	17	15	11	0.0714	0.1094	1101	0.0743	0.0992	1512
NCSSI-auto-01	automatic	22	17	23	0.1708	0.2107	2432	0.1511	0.1991	2256
NCSSI-auto-02	automatic	21	12	17	0.1536	0.1962	2322	0.1511	0.1991	2256
R2D2-auto-01	automatic	19	15	21	0.1747	0.2239	2257	0.162	0.2066	2155
R2D2-auto-02	automatic	19	19	21	0.181	0.2236	2257	0.162	0.2066	2155
ZKN-auto-01	automatic	22	17	11	0.1523	0.2011	2139	0.1405	0.1839	2152
ZKN-auto-02	automatic	22	18	14	0.1537	0.1968	2153	0.1405	0.1839	2152
ABRIR-auto	automatic	28	20	25	0.2198	0.2506	2591	0.169	0.2085	2422
JSWEB-relevant-B-02	user	7	5	5	0.0235	0.049	755	0.0743	0.0992	1512
JSWEB-relevant-B-03	user	18	18	13	0.0976	0.1466	1453	0.0743	0.0992	1512
NCSSI-user-01	user	27	18	17	0.2434	0.2705	2508	0.173	0.2258	2353
NCSSI-user-02	user	28	15	16	0.2196	0.2487	2415	0.173	0.2258	2353
ABRIR-user	user	32	18	20	0.2569	0.2834	2689	0.1801	0.2268	2469

Table 3in the proceedings
Summary of Evaluation Results (Page 2) in CD-ROM

No expansion (automatic)									
	Average Precision			R-Precision			Relevant Retrieved		
tid	Max	Min	Average	Max	Min	Average	Max	Min	Average
1	0.2716	0.068	0.188	0.1667	0.0833	0.15002	12	10	11.6
3	0.2543	0.2122	0.22536	0.35	0.2	0.27	20	15	17.8
4	0.2165	0.0548	0.12384	0.1667	0	0.13336	6	5	5.2
6	0.4926	0.3903	0.42346	0.5601	0.5142	0.53416	506	454	471.4
19	0.1317	0.0185	0.07236	0.1739	0	0.11302	18	12	15.8
21	0.4097	0.0349	0.29438	0.47	0.05	0.34	96	5	71.2
22	0.1836	0.1586	0.17038	0.3247	0.2784	0.2928	118	85	95.4
23	0.0221	0.0089	0.0154	0.0667	0	0.01334	12	8	9.8
28	0.0818	0.0468	0.06644	0.119	0.0714	0.0952	24	12	20.6
29	0.0316	0.0061	0.02008	0.1053	0.0376	0.0782	36	26	30.4
34	0.2329	0	0.12758	0.3235	0	0.18824	24	0	18
44	0.0808	0.02	0.05254	0.1549	0.0282	0.10706	30	4	22.2
45	0.1397	0.0125	0.07896	0.1818	0.0455	0.12728	17	5	13.2
55	0.5145	0.0131	0.32546	0.4762	0.0714	0.32858	41	11	33.2
58	0.5185	0.1744	0.39512	0.5718	0.2706	0.47224	468	203	383.6
61	0.0601	0.0002	0.03362	0.1429	0	0.08574	6	1	4
62	0.2882	0.0323	0.2038	0.3438	0.0703	0.2578	119	52	100.6
63	0.0908	0.0056	0.04442	0.1667	0.0303	0.09396	27	8	18.8
65	0.1051	0.0076	0.03994	0.1899	0.0506	0.10466	103	24	54.2
68	0.0845	0.0121	0.0427	0	0	0	10	2	8
70	0.2521	0.0815	0.164	0.3182	0.1212	0.20304	50	36	45.8
71	0.0621	0.0002	0.03202	0.0909	0	0.05454	13	2	9.6
73	0.197	0.0927	0.1464	0.1702	0.0851	0.14466	43	36	41.6
74	0.1854	0.0507	0.1196	0.2222	0.1111	0.17776	26	14	22
76	0.5257	0.2806	0.39136	0.5839	0.3147	0.43288	168	97	131.8
80	0.0668	0.0011	0.03588	0.129	0	0.05808	27	5	16
82	0.5132	0.0289	0.31348	0.5667	0.1302	0.40448	419	102	306
84	0.0299	0	0.0084	0.0263	0	0.00526	21	0	8.4
86	0.0183	0.0008	0.01256	0.0541	0	0.03786	15	4	10.4
88	0.3919	0.1247	0.31782	0.4571	0.1714	0.35428	31	20	25.6
91	0.0798	0.0191	0.04618	0.1667	0.0833	0.11666	11	6	8.4
95	0.1973	0.0024	0.10524	0.25	0	0.125	12	3	7.6
97	0.134	0.0269	0.07996	0.2105	0.0263	0.14736	20	3	16.2
98	0.0847	0.0001	0.04774	0.1	0	0.07	14	1	11.2
99	0.1699	0.037	0.11338	0.24	0.04	0.164	43	21	33.8
All	0.169	0.0743	0.13938	0.2085	0.0992	0.17946	2422	1512	2099.4