Overview of the NTCIR-5 WEB Query Term Expansion Subtask

Masaharu Yoshioka
Graduate School of Information Science and Technology, Hokkaido University
N14 W9, Kita-ku, Sapporo-shi, Hokkaido, JAPAN
National Institute of Informatics
yoshioka@ist.hokudai.ac.jp

Abstract

The query term expansion subtask was conducted to establish an evaluation framework for information retrieval (IR) systems that focus on the effectiveness of query term expansion techniques. However, the quality of query term expansions are affected by several factors (e.g., IR system using expanded query, quality of initial query, etc.), so it is difficult to evaluate this technique.

In this subtask, I assume the topic difficulty for the query term expansion technique is caused by the mismatch between different information-need expressions (query terms and relevant documents). To take into account this topic difficulty, I propose feature quantities to characterize this difficulty and propose a new framework to evaluate each query expansion system.

1 Introduction

It is very difficult for many users of Information Retrieval (IR) utilities to select appropriate query terms to represent the information need. Because query terms are often imprecise and inappropriate, the documents selected may contain only some of the query words and be irrelevant to the user's needs.

To reduce the mismatch between query terms and information need, many IR systems use query term expansion techniques to find better query terms. However, the effectiveness of this technique depends on the quality of query terms in the initial query and documents used for query term expansion. Cronen-Townsend et al. [3] used a query clarity score based on a language model to decide if the query terms contain relevant information for the query term expansion; this approach was shown to be effective.

The Reliable Information Access (RIA) Workshop [5] conducted a failure analysis [1] for a set of topics using seven different popular IR systems and proposed a topic categorization based on the types of failures they encountered. They also conducted a relevance feedback experiment using a different IR systems [7]. This study, however, did not examine the relationship between topic difficulty based on the mismatch and the effect of relevance feedback.

In this subtask, I aim to establish an evaluation framework for IR systems that focuses on query term expansion. However the quality of query term expansion is affected by several factors, for example, when I evaluate a query term expansion technique using results of information retrieval, these results may be affected by the characteristics and quality of the IR system. In other cases, when a user carefully selects the query terms, the query term expansion is not performed well.

To evaluate the appropriateness of this evaluation framework and several IR systems with query term expansion techniques, each participant conducted retrieval experiments that used the survey type topics of the NTCIR-4 web test collection [4].

The remainder of this paper is divided into five sections. Section 2 proposes various statistical features for defining the topic difficulty and the effectiveness of the query expansion term. I also explain the collection of such information. In Section 3, I briefly review the NTCIR-4 web test collection and analyze the characteristics of topics in the test collection using the information defined in Section 2. Section 4 presents guideline for the retrieval experiment and for submission of the result. Section 5 analyzes the experimental results and Section 6 gives the conclusions of the paper.

2 Statistical Features for Evaluation of the Query Term Expansion Technique

Buckley et al. [2] hypothesized a possible reason why query expansion improves the query performance as follows.

1. one or two good alternative words to original query terms (synonyms)

2. one or two good related words

3. a large number of related words that establish that some aspect of the topic is present (context)

4. specific examples of general query terms

5. better weighting to original query terms

The first four reasons relate to the query term expansion. I can evaluate reasons 1, 2, and 4 using a thesaurus. However, since Voorhees [6] confirmed simple automatic query term expansion based on a thesaurus did not improve query performance, it may be inappropriate to use a thesaurus for this evaluation.

Therefore, I propose to use mismatch between different information-need expressions (query terms and relevant documents) for this evaluation.

2.1 Feature Quantities for Characterizing Mismatch between Initial Query and Relevant Documents

When a user carefully selects good query terms, query term expansion is not well performed. Therefore, it is crucial for this subtask to evaluate the quality of the initial query based on the mismatch between the initial query and the relevant documents.

When the query is represented with a Boolean operator this mismatch is characterized as a mismatch between the documents that satisfy this query and the relevant documents. When the initial query is good, documents that satisfy the Boolean query (Boolean satisfied documents) and relevant documents are equivalent ((1) and (3) in Figure 1 are an empty set).

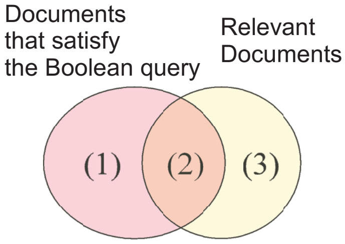

Figure 1. Mismatch between Initial Query and Relevant Documents

However, because it is difficult to construct good queries, (1) and (3) are not an empty set in almost every query. The size of (1) and (3) characterizes the quality of the query from the viewpoint of query expansion. For example, when there are many documents in (1), the initial query is too general and requires new query terms that define the context of the query. Conversely, when (3) has many documents the initial query is too strict and it is necessary to determine alternative words to relax the query.

Since the number of documents in (1) and (3) are affected by the number of relevant documents, I use following two feature quantities for this evaluation.

$R\&B/R$ The ratio between the size of relevant documents that satisfy the Boolean query ((2)) and the size of relevant documents ((2)+(3)).

$R\&B/B$ The ratio between the size of relevant documents that satisfy the Boolean query ((2)) and the size of the Boolean satisfied documents ((1)+(2)).

2.2 Feature Quantities for Evaluating Effectiveness of the Query Term

Feature quantities proposed in 2.1 can be also used to evaluate a query term when the Boolean query is constructed using only this term.

In addition to these features, I propose the following three criteria for selecting feature quantities.

1. Appropriateness of the alternative term for each initial query term.

2. Appropriateness of the context definition term for the query.

3. Appropriateness of the term that characterizes the relevant documents.

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.

The following feature quantities are defined for each query expansion term.

$total : Rel$ The number of relevant documents that have a query expansion term.

$total : Bool$ The number of Boolean satisfied documents that have a query expansion term.

$total : R\&B$ The number of Boolean satisfied relevant documents that have a query expansion term.

$total : All$ The number of documents that have a query expansion term in the document database.

The following feature quantities are defined for each set of an initial query term and query expansion term.

$comp : Rel$ The number of relevant documents that have a query expansion term and do not have an initial query term.

$cooc : Rel$ The number of relevant documents that have a query expansion term and an initial query term.

$cooc : Bool$ The number of Boolean satisfied documents that have a query expansion term and an initial query term.

$cooc : All$ The number of documents that have a query expansion term and an initial query term in the document database.

I use feature quantities that are based on mutual information content for evaluating the distinctiveness of each term [9]. These quantities are the mutual information content between relevant documents r and the term w. $p(w)$ is the probability of the term w being in the document database and $p(w|r)$ is the probability for the relevant documents.

$$MI(w) \quad = \quad p(w|r)log_2\frac{p(w|r)}{p(w)}$$

When term w exists explicitly in the relevant documents, $MI(w)$ increases.

3 Analysis of The NTCIR-4 Survey Type Topics

3.1 The NTCIR-4 Web Test Collection

The NTCIR-4 Web test collection [4] is a set of 100 gigabytes of html document data and 80 topics for retrieval experiments. 35 topics out of 80 are for the survey retrieval topics and the other 45 are for the target retrieval topics. The survey retrieval topics are designed for finding most relevant documents and the target retrieval topics are for finding just one, or only a few, relevant documents of the highly ranked documents. Since the target retrieval topics may miss relevant candidate documents, only the survey retrieval topics were used for this query term expansion subtask.

Figure 2 shows a sample topic in this test collection. <TITLE> includes 1–3 terms with Boolean expressions. The attribute “CASE” in <TITLE>, <ALT0>, <ALT1>, <ALT2>, <ALT3> means:

(a) All the terms are related to one another by the OR operator.

(b) All the terms are related to one another by the AND operator.

(c) Only two terms can be related using the OR operator; the rest are specified by the attribute “RELAT.”

For the sample topic described in Figure 2, I can formulate the Boolean query (オフサイド (offside) and (サッカー (soccer) or ルール (rule))) from TITLE and (オフサイド (offside) and サッカー (soccer) and ルール (rule)) from ALT3.

<TOPIC> <NUM>0001</NUM>
<TITLE CASE=“c” RELAT=“2-3”> オフサイド, サッカー, ルール </TITLE>
<DESC> サッカーのオフサイドというルールについて説明されている文書を探したい </DESC>
<NARR><BACK> サッカーでオフサイドとはどういうルールなのかを知りたい。</BACK><TERM> オフサイドはオフェンス側の反則である。オフサイドが適用される状況にはいくつかのパターンがむ爐蝓Ëサッカーのルールの中で最もわかりにくいものである。</TERM><RELE> 適合文書はオフサイドが適用される状況を説明しているもの</RELE></NARR>
<ALT0 CASE=“b”> オフサイド </ALT0>
<ALT1 CASE=“b”> オフサイド, 選手, 位置 </ALT1>
<ALT2 CASE=“b”> オフサイド, サッカー </ALT2>
<ALT3 CASE=“b”> サッカー, オフサイド, ルール </ALT3>
<USER> 大学 2 年, 男性, 検索歴 4 年, 熟練度 3, 精通度 5</USER>
</TOPIC>

(a) An original sample topic

<TOPIC> <NUM>0001</NUM>
<TITLE CASE=“c” RELAT=“2-3”> offside, soccer, rule </TITLE>
<DESC> I want to find documents that explain the offside rule in soccer. </DESC>
<NARR> <BACK> I want to know about the offside rule in soccer. </BACK> <TERM> Offside is a foul committed by a member of the offense side. There are several patterns for situations in which the offside rule can be applied, and it is the most difficult soccer rule to understand. </TERM> <RELE> Relevant documents must explain situations where the offside rule applies</RELE> </NARR>
<ALT0 CASE=“b”> offside </ALT0>
<ALT1 CASE=“b”> offside, player, position </ALT1>
<ALT2 CASE=“b”> offside, soccer </ALT2>
<ALT3 CASE=“b”> soccer, offside, rule </ALT3>
<USER> 2nd year undergraduate student, male, 4 years of search experience, skill level 3, familiarity level 5 </USER>
</TOPIC>

(b) An English translation of the sample topic

Figure 2. A sample topic from the NTCIR-4 Web test collection [4]

3.2 Feature Quantities of Topics in NTCIR-4 Web Survey Retrieval Topics

A graph in Figure 3 shows a characteristic of the topics in the Survey Retrieval Topics. The X axis of the graph is $R\&B/R$ and the Y axis is $R\&B/B$. The size of each circle indicates the size of the relevant documents.

These statistical values were calculated using an organizer reference IR system named The Appropriate Boolean Query Reformulation for Information Retrieval (ABRIR) [8]. Because such values may differ according to the method of extracting index keywords from the documents [1]

From this graph, all initial queries were not sufficiently appropriate to distinguish all relevant documents from the other documents. For the topics that have higher $R\&B/R$ and lower $R\&B/B$, such as topics 1, 4, 6, 55, and 98, the term for context definition may be good query expansion terms. For the topics that have lower $R\&B/R$ and higher $R\&B/B$, such as 65, 76, and 82, alternative terms may be good query expansion terms. The topics that have lower $R\&B/R$ and lower $R\&B/B$, such as 45, 62, 63, 80, and 84, may require various types of query expansion terms.

4 Guideline for the Retrieval Experiments

We used the NTCIR-4 web test collection data for the formal run. We used the survey type topics only (topic numbers 0001, 0003, 0004, 0006, 0019, 0021, 0022, 0023, 0028, 0029, 0034, 0044, 0045, 0055, 0058, 0061, 0062, 0063, 0065, 0068, 0070, 0071, 0073, 0074, 0076, 0080, 0082, 0084, 0086, 0088, 0091, 0095, 0097, 0098, and 0099).

The quality of query term expansions are affected by several factors (e.g., The IR system that uses the expanded query, the quality of the initial query, etc.) so it is difficult to evaluate this technique by itself.

Therefore, we conducted several retrieval experiments to reduce the effect of different elements.

- The effect of the query term expansion and the number of terms used for the query term expansion.
  - No query term expansion vs. query term expansion.
  - Query term expansion with a limited number of terms (10) vs. query term expansion with no limitation.
- The effect of documents that are used for query term expansion
  - Pseudo-relevant documents vs. user selected relevant documents
- The effectiveness of expanded query terms
  - Statistical analysis of expanded query terms and relevant documents.
  - Correlation between statistical analysis and retrieval performance.
- Comparison between ideal query expansion terms
  - Generation of ideal query expansion term candidates and retrieval results using all relevant document information defined in the test collection

The following are the specifications of the retrieval experiment formally run by the participants.

Initial Query Terms We used the TITLE(tt) field for all experiments.

Type of relevance feedback We use two different methods to select the feedback documents. <feedback-type>

- Automatic selection (e.g., use the top-N ranked documents) (auto):
- Used all relevant documents as selected as relevant documents (relevant-A: use “S(H)” and “A” as relevant documents, relevant-B: use “S(H)”, “A” and “B” as relevant documents):
- Simulation of user selection using relevant document information and system output (e.g., initial retrieval results. (user)
  - You could use relevant document information, however, you had to explain the interface of your IR system and make assumptions about the users’ behavior as a scenario. The assumptions should be consistent for all topics; you must not have changed this assumption manually for each topic.
  - You could use grades (“S(H)”, “A”, “B” and “C”) for relevant document information. For example, you could use “B” relevant documents when you could not find “S(H)” or “A” in a document list. You could also use this grade as a user’s confidence of relevance for calculating the weight of the document’s importance.
  - You could use other related information on WWW documents, such as document length, URL pattern, language, etc.

[1] ABRIR in NTCIR-4 [8] uses a noun plus two adjacent nouns as index words. Verbs were also used as index words in this experiment.

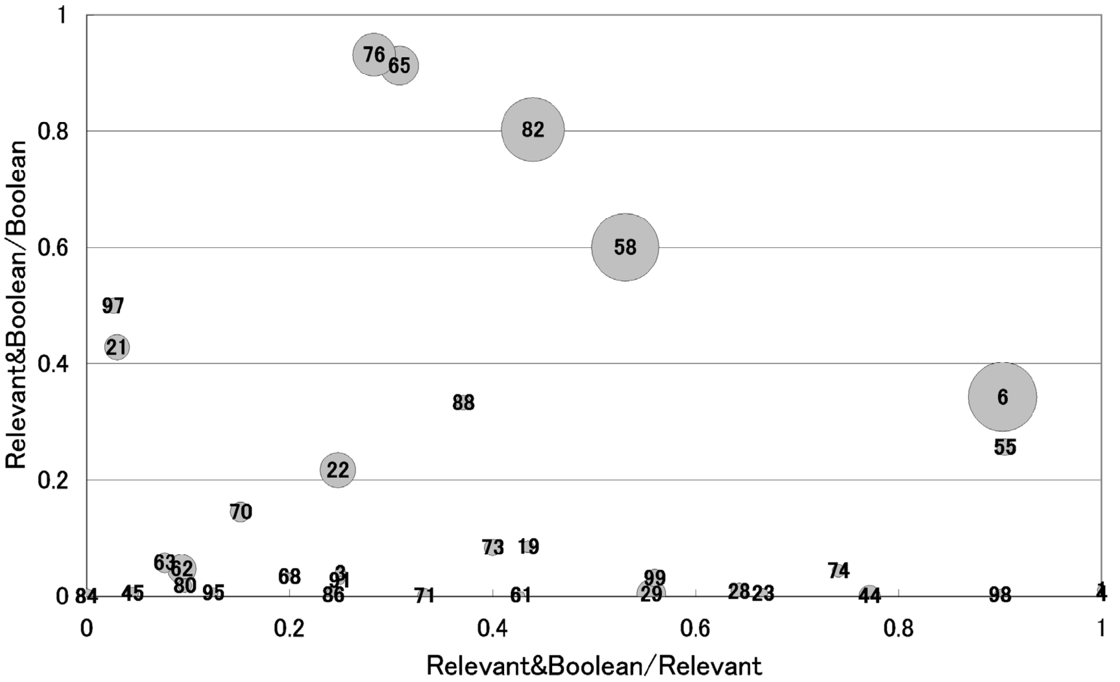

Figure 3. Characteristics of the Topics

- You could also use the top-N pseudo-relevant documents.
- To restrict amount of work required of a user, I introduced the following restriction for scenario definition.
  - ∗ A user could read 20 documents at most for relevant judgments and select at most five documents as relevant.
  - ∗ If your system required a user to check more than 20 documents, you could use more documents. However, the precise reason was noted in the scenario.
  - ∗ If your system required a user to select more than five relevant documents (e.g., selection of document cluster with more than five documents), you could use more documents. However, the precise was noted reason in the scenario.

Each participant submitted sets of retrieval results, expansion term candidates lists, and document lists that were used for query term expansion.

5 Analysis of the Retrieval Experiments Results

5.1 Summary of Participants

Four participants, listed below in alphabetical order of affiliation, and one organizer reference system submitted their completed run results.

- Justsystem Corporation
- National Institute of Informatics
- National Institute of Informatics, the University of Tokyo, and KYA group
- NTT Cyber Solutions Laboratories; NTT Corporation

Several query term expansion techniques and information retrieval models were used in these runs for which results were submitted.

5.1.1 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.

5.1.2 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.

5.1.3 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.

5.1.4 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.

5.1.5 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.

5.2 Additional Relevance Judgment

Since there were several submission results whose top-ranked documents are not included in the judged document list[2], it is unfair to evaluate the system results using the relevant document list in the test collection.

Due to time limitations the relevance judgment for this subtask is not the same as that of the NTCIR-4 Web[4][3].

5.2.1 Pooling

Because the quality of the document list used for query term expansion may affect the quality of the query term expansion, these document lists were used for pooling in addition to the submitted results. In addition to evaluating the Boolean query quality, experimental results using ABRIR with original Boolean query were included in the pool.

I took the top 50 ranked documents from each of the submitted results and the top 10 ranked documents from the document list used for the query term expansion.

The documents in the judged document list were removed from this pool. Pooled documents were ranked using the same method as in [4].

5.2.2 Relevance Assessment

In relevance assessment a page-unit document model[4] was used as the document model. However, the relevant document list of NTCIR-4 consists of a one-click distance document model[5].

The assessor judged the "Multi-Grade Relevance" of the individual documents as : highly relevant, fairly relevant, partially relevant or irrelevant. This is same as for NTCIR-4.

5.3 Summary of Evaluation Results

Table 1 shows the overall evaluation of the submitted runs. In most of the runs, the query term expansion technique improved the retrieval performance on average, but there was no run that improved the query performance for all topics.

Table 2, 3, 4, 5 show maximum, minimum, and average value for average precision, R-precision, and relevance retrieved for each topic. In this table, the highest performance run IDs for each team was selected and used to calculate these values.

Table 6 shows the number of Run IDs where query performance improved by using query term expansion techniques. There is no direct correlation between the effectiveness of the query term expansion technique and the mismatch between the initial query and the relevant documents showed in Figure 3.

It is interesting that there are several topics whose number of Run IDs for a user is lower than that for automatic (e.g., 0021, 0058, 0076, 0082). Those topics have higher $R\&B/B$ values compared with other topics and it means 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 nonrelevant documents may be useful for finding alternative terms for initial query terms.

6 Conclusion

In this paper, I propose a new framework to evaluate query term expansion techniques using mismatch between different information-need expressions (query terms and relevant documents). Although there is no direct correlation between the effectiveness of the query term expansion technique and the mismatch between the initial query and the relevant documents, I

[2] Documents that were checked by the assessor. A non-checked document may be relevant.

[3] I plan to rectify this mismatch before final data release.

[4] An assessor judged the relevance of a page only on the basis of the entire information given on it

[5] Assessors judged the relevance of a page by using out-linked page information in addition to the information given on it

Table 1. Evaluation Results for Average

Run ID	type of feedback	No. of topics where performace improve			10 query terms expansion (average)			No query term expansion (average)		
		AP	RP	RR	AP	RP	RR	AP	RP	RR
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	25	17	13	0.1523	0.2011	2139	0.139	0.1824	2137
ZKN-auto-02	automatic	23	18	16	0.1537	0.1968	2153	0.139	0.1824	2137
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

AP: Average Precision, RP: R-Precision, RR: Relevant Retrieved

confirmed that this mismatch is one of the factors that affects the performance of the technique.

For the future work further analysis is necessary to establish a framework to evaluate the query term expansion technique in isolation.

Acknowledgments

This research was partially supported by a Grant-in-Aid for Scientific Research on Priority (2), 15017202 from the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

References

[1] C. Buckley. Why current ir engines fail. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 584–585, New York, NY, USA, 2004. ACM Press.

[2] C. Buckley and D. Harman. Reliable information access final workshop report. Technical report, Northeast Regional Research Center, MITRE, 2004. http://nrrc.mitre.org/NRRC/Docs_Data/RIA_2003/ria_final.pdf.

[3] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, 2002.

[4] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa. Overview of the informational retrieval task at ntcir-4 web. In *Working Notes of the Fourth NTCIR Workshop Meeting*, 2004. http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/WEB/NTCIR4WN-OV-WEB-A-EguchiK.pdf.

[5] D. Harman and C. Buckley. Sigir 2004 workshop: Ria and "where can ir go from here?". *SIGIR Forum*, 38(2):45–49, 2004.

[6] E. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, 1994.

[7] R. H. Warren and T. Liu. A review of relevance feedback experiments at the 2003 reliable information access (ria) workshop. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 570–571, New York, NY, USA, 2004. ACM Press.

[8] M. Yoshioka and M. Haraguchi. Study on the combination of probabilistic and boolean ir models for www documents retrieval. In *Working Notes of the Fourth NTCIR Workshop Meeting, Supplement Volume*, pages 9–16, 2004. http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/WEB/NTCIR4WN-WEB-YoshiokaM.pdf.

[9] M. Yoshioka and M. H. W. Oniki. An appropriate boolean query reformulation interface for information retrieval based on adaptive generalization. In *International Workshop on Challenges in Web Information Retrieval and Integration*, pages 145–150, 2005.

Table 2. Evaluation Results for Each Topic (Automatic, 10 query terms expansion)

Topic	Average Precision			R-Precision			Relevant Retrieved		
	Max	Min	Average	Max	Min	Average	Max	Min	Average
0001	0.319	0.0038	0.17126	0.25	0	0.16668	12	8	10.4
0003	0.2887	0.1851	0.24148	0.35	0.2	0.28	20	14	17.8
0004	0.3559	0.082	0.23542	0.5	0.1667	0.26668	6	5	5.2
0006	0.4366	0.1583	0.364	0.5304	0.2686	0.47506	470	250	422.2
0019	0.127	0.0226	0.05954	0.1739	0	0.07826	18	14	16.8
0021	0.6828	0.0356	0.35538	0.69	0.06	0.4	97	6	72.4
0022	0.3524	0.1905	0.25428	0.4381	0.3093	0.35464	150	88	113.6
0023	0.027	0.0105	0.01808	0.0667	0	0.04002	12	7	9.8
0028	0.153	0.0461	0.08648	0.2381	0.0476	0.14284	33	11	23.4
0029	0.037	0.0117	0.02108	0.1278	0.0451	0.07968	38	30	34.2
0034	0.308	0	0.16002	0.3824	0	0.21766	27	0	19.6
0044	0.067	0.0016	0.03704	0.1408	0.0282	0.0986	27	3	19.8
0045	0.2545	0.0454	0.15978	0.3182	0	0.2091	19	14	16.8
0055	0.6531	0.0336	0.36368	0.619	0.119	0.40476	41	14	34.2
0058	0.5257	0.1165	0.40794	0.5662	0.2469	0.4887	474	204	400.4
0061	0.0345	0	0.0208	0.1429	0	0.05716	6	0	4
0062	0.3341	0.0001	0.22234	0.4062	0	0.28126	121	2	93.2
0063	0.0818	0.0005	0.04574	0.1667	0	0.11212	29	6	20.4
0065	0.4412	0.0004	0.1106	0.4599	0.0084	0.15864	165	9	66.6
0068	0.0736	0.011	0.04548	0.1	0	0.04	10	2	8.2
0070	0.3342	0.079	0.2275	0.3636	0.1212	0.29088	55	35	49.4
0071	0.0572	0.0017	0.02072	0.0606	0	0.0303	14	7	11.2
0073	0.231	0.0681	0.15914	0.1915	0.1489	0.17872	43	30	40.4
0074	0.2114	0.0862	0.14536	0.2593	0.0741	0.17778	26	14	21.6
0076	0.533	0.024	0.33062	0.5804	0.0769	0.36572	169	50	131.6
0080	0.0949	0.0006	0.0465	0.129	0	0.05162	27	4	17
0082	0.5247	0.0041	0.32178	0.5587	0.0508	0.39144	429	46	301.4
0084	0.0298	0	0.00786	0.0263	0	0.00526	21	0	8
0086	0.0173	0.0054	0.01396	0.0811	0.027	0.05408	20	5	12.6
0088	0.4652	0.2809	0.40524	0.4857	0.3143	0.42286	33	20	28.8
0091	0.2317	0.0585	0.10846	0.4167	0.0833	0.2	10	6	8.6
0095	0.1992	0.0083	0.11598	0.25	0	0.125	12	2	8.2
0097	0.302	0.0017	0.195	0.3421	0.0263	0.23684	27	3	18.8
0098	0.193	0.0256	0.10018	0.25	0.05	0.13	17	12	14.6
0099	0.2224	0.0214	0.12946	0.26	0	0.172	43	21	32.4

Table 3. Evaluation Results for Each Topic (Automatic, no expansion)

Topic	Average Precision			R-Precision			Relevant Retrieved		
	Max	Min	Average	Max	Min	Average	Max	Min	Average
0001	0.2716	0.068	0.18818	0.1667	0.0833	0.15002	12	10	11.6
0003	0.2543	0.2122	0.22402	0.35	0.2	0.27	20	15	17.8
0004	0.2165	0.057	0.12428	0.1667	0	0.13336	6	5	5.2
0006	0.4926	0.3903	0.42308	0.5601	0.5142	0.53336	506	454	470.2
0019	0.1317	0.0185	0.07324	0.1739	0	0.11302	18	12	15.8
0021	0.4097	0.0303	0.29346	0.47	0.05	0.34	96	5	71.2
0022	0.1836	0.1586	0.17044	0.3247	0.2784	0.2928	118	85	95.4
0023	0.0221	0.0089	0.0154	0.0667	0	0.01334	12	8	9.8
0028	0.0806	0.0468	0.0662	0.119	0.0714	0.0952	24	11	20.4
0029	0.0316	0.0061	0.02	0.1053	0.0376	0.0782	36	26	30.2
0034	0.2329	0	0.12838	0.3235	0	0.18824	24	0	18
0044	0.0793	0.02	0.05224	0.1549	0.0282	0.10706	30	4	22.2
0045	0.1397	0.0125	0.0804	0.1818	0.0455	0.12728	17	5	13.2
0055	0.5145	0.0131	0.32742	0.4762	0.0714	0.32858	41	11	33.2
0058	0.5185	0.1744	0.3953	0.5718	0.2706	0.47252	468	203	383.6
0061	0.0601	0.0002	0.03346	0.1429	0	0.08574	6	1	4
0062	0.2882	0.0323	0.2045	0.3438	0.0703	0.2578	119	52	100.6
0063	0.0908	0.0056	0.04444	0.1667	0.0303	0.09396	27	8	18.8
0065	0.1051	0.0076	0.04048	0.1899	0.0506	0.10634	103	24	54
0068	0.0845	0.0116	0.042	0	0	0	10	2	8
0070	0.258	0.0815	0.16518	0.3182	0.1212	0.20304	52	36	46.2
0071	0.0621	0.0002	0.03178	0.0909	0	0.0606	13	2	9.6
0073	0.197	0.0927	0.13876	0.1702	0.0851	0.14466	43	36	41.6
0074	0.1854	0.0507	0.12034	0.2222	0.1111	0.17776	26	14	22
0076	0.5197	0.2806	0.39016	0.5769	0.3147	0.43148	165	97	131.2
0080	0.0668	0.0011	0.03662	0.129	0	0.05808	27	5	16
0082	0.5132	0.0289	0.31354	0.5667	0.1302	0.40448	419	102	305.6
0084	0.0299	0	0.0084	0.0263	0	0.00526	21	0	8.4
0086	0.0183	0.0008	0.01256	0.0541	0	0.03786	15	4	10.4
0088	0.3919	0.1247	0.31798	0.4571	0.1714	0.35428	31	20	25.6
0091	0.0494	0.0191	0.0399	0.1667	0.0833	0.09998	9	6	7.8
0095	0.1973	0.0024	0.10554	0.25	0	0.125	12	3	7.6
0097	0.134	0.0269	0.08002	0.2105	0.0263	0.14736	20	3	16.2
0098	0.0841	0.0001	0.04762	0.1	0	0.07	14	1	11.2
0099	0.1699	0.037	0.11188	0.24	0.04	0.164	43	21	33.8

Table 4. Evaluation Results for Each Topic (User, 10 query terms expansion)

Topic	Average Precision			R-Precision			Relevant Retrieved		
	Max	Min	Average	Max	Min	Average	Max	Min	Average
0001	0.6276	0.0918	0.2384	0.5	0.1667	0.21668	12	11	7
0003	0.2717	0.2606	0.16034	0.35	0.2	0.18	20	15	10.8
0004	0.3375	0.2096	0.15288	0.3333	0.1667	0.13334	6	5	3.2
0006	0.485	0.3127	0.24788	0.5722	0.417	0.30338	494	363	265.2
0019	0.191	0.0116	0.07552	0.2609	0.0435	0.11306	21	11	10.4
0021	0.3876	0.1365	0.1524	0.41	0.17	0.18	90	48	43.2
0022	0.38	0.2791	0.20108	0.433	0.3866	0.24948	150	130	83.4
0023	0.0862	0.0097	0.02388	0.0667	0	0.02668	12	7	6
0028	0.1621	0.1293	0.0888	0.2619	0.1667	0.12858	35	33	20.2
0029	0.2685	0.1732	0.1324	0.3383	0.2556	0.17592	111	81	56.4
0034	0.2985	0	0.11296	0.4412	0	0.14706	32	0	10.8
0044	0.0722	0.0001	0.02736	0.1127	0	0.03662	26	1	10
0045	0.4089	0.0019	0.12644	0.4091	0	0.14546	20	7	9
0055	0.6531	0.1224	0.26578	0.619	0.2619	0.28094	41	28	21.6
0058	0.5257	0.1077	0.20132	0.5662	0.2273	0.25578	474	192	214.6
0061	0.0867	0	0.02424	0.1429	0	0.05716	5	0	1.8
0062	0.3418	0.0192	0.1306	0.3516	0.0781	0.15156	120	35	54.6
0063	0.1747	0.0005	0.04462	0.2121	0	0.07272	44	5	15.4
0065	0.4412	0.0404	0.1343	0.4599	0.1392	0.1713	165	75	69.8
0068	0.0782	0.0629	0.04294	0.1	0	0.02	10	9	5.8
0070	0.36	0.1325	0.15258	0.4242	0.1667	0.19394	55	45	30.2
0071	0.0745	0.0001	0.02288	0.1212	0	0.04242	18	1	6.4
0073	0.2213	0.094	0.0967	0.2128	0.1702	0.1149	43	38	24.8
0074	0.2669	0.1167	0.1247	0.2963	0.1481	0.14814	26	23	14.4
0076	0.4197	0.0632	0.15806	0.4336	0.1783	0.19232	160	88	74.8
0080	0.2205	0	0.0646	0.2903	0	0.07742	29	1	11.6
0082	0.5391	0.0033	0.2107	0.581	0.0429	0.2378	433	45	179.2
0084	0.0425	0	0.01376	0.0526	0	0.01578	21	0	7.2
0086	0.038	0.0112	0.01322	0.1081	0.0541	0.04866	20	5	8
0088	0.4905	0.3421	0.2495	0.4571	0.3714	0.25712	33	26	17.2
0091	0.2217	0.1101	0.0898	0.4167	0.0833	0.11666	10	8	5.4
0095	0.2368	0.0001	0.09394	0.25	0	0.1	12	1	4.4
0097	0.348	0.1998	0.16464	0.3684	0.2632	0.18948	33	26	17.8
0098	0.3767	0.0011	0.1143	0.3	0	0.12	20	5	8.4
0099	0.2462	0.0403	0.10556	0.3	0.1	0.136	44	33	23.8

Table 5. Evaluation Results for Each Topic (User, no expansion)

Topic	Average Precision			R-Precision			Relevant Retrieved		
	Max	Min	Average	Max	Min	Average	Max	Min	Average
0001	0.2691	0.068	0.1192	0.1667	0.0833	0.08334	12	12	7.2
0003	0.2231	0.2091	0.12854	0.25	0.2	0.14	19	16	10.8
0004	0.2165	0.0782	0.07508	0.1667	0.1667	0.10002	6	5	3.2
0006	0.4926	0.3897	0.2546	0.5601	0.5061	0.31472	506	451	282
0019	0.1381	0.0185	0.05894	0.2609	0	0.10436	21	12	10.8
0021	0.3554	0.2412	0.19	0.38	0.33	0.218	88	81	51.4
0022	0.1803	0.1619	0.101	0.3247	0.2938	0.1835	118	93	60.8
0023	0.0214	0.0132	0.01044	0.0667	0	0.02668	11	8	6
0028	0.0683	0.0577	0.0372	0.0952	0.0714	0.0476	24	22	13.6
0029	0.0163	0.0061	0.00768	0.0677	0.0376	0.0346	30	27	17
0034	0.2439	0	0.0962	0.3529	0	0.13528	24	0	9.4
0044	0.0681	0.02	0.02964	0.1127	0.0282	0.05072	31	4	12.8
0045	0.1148	0.0125	0.04744	0.1818	0.0455	0.08182	17	5	7.6
0055	0.5145	0.0131	0.20562	0.4762	0.0714	0.19524	40	11	18.2
0058	0.5185	0.1744	0.22012	0.5718	0.2706	0.26862	468	203	217.2
0061	0.055	0.0002	0.01826	0.1429	0	0.05716	4	1	1.6
0062	0.29	0.0323	0.1224	0.3516	0.0703	0.15314	119	52	58
0063	0.0502	0.0056	0.02072	0.1515	0.0303	0.0606	23	8	10.8
0065	0.1051	0.0083	0.03858	0.1899	0.0591	0.08018	103	26	42.2
0068	0.0599	0.0343	0.02924	0	0	0	10	9	5.8
0070	0.2743	0.0815	0.10142	0.2576	0.1364	0.1091	52	36	27.2
0071	0.0658	0.0002	0.02608	0.0909	0	0.03636	14	2	5.8
0073	0.135	0.0927	0.07204	0.1489	0.0851	0.07658	43	36	24.4
0074	0.1963	0.0507	0.08138	0.2222	0.1111	0.1037	26	20	13.8
0076	0.4557	0.2806	0.2204	0.5315	0.3147	0.24826	161	97	78.4
0080	0.0859	0.0011	0.03166	0.129	0	0.04516	28	5	12
0082	0.5296	0.0289	0.21374	0.573	0.1302	0.25366	430	102	190
0084	0.0333	0	0.0082	0.0526	0	0.01052	18	0	6.6
0086	0.0175	0.0008	0.00694	0.0541	0	0.02164	15	4	6.8
0088	0.3808	0.1247	0.17612	0.4571	0.1714	0.21142	29	20	15.2
0091	0.0419	0.0191	0.01958	0.0833	0.0833	0.04998	9	7	5
0095	0.2036	0.0024	0.08072	0.25	0	0.1	9	3	3.6
0097	0.0636	0.0269	0.03072	0.1579	0.0263	0.06842	20	3	8.6
0098	0.0444	0.0001	0.01708	0.1	0	0.04	14	1	5.8
0099	0.1616	0.037	0.0716	0.24	0.04	0.1	42	34	23

Table 6. Effectiveness of Query Term Expansion Technique for Each Topic

Topic	No. of Run IDs where performace improve					
	automatic			user		
	AP	RP	RR	AP	RP	RR
0001	6	3	2	4	3	0
0003	4	3	2	4	2	1
0004	9	7	0	3	1	0
0006	5	3	5	3	2	3
0019	2	1	1	2	1	0
0021	5	4	5	1	1	1
0022	8	7	6	5	4	5
0023	4	3	1	2	0	1
0028	5	4	4	5	5	5
0029	4	3	8	5	5	5
0034	7	5	6	3	1	1
0044	1	1	1	2	0	0
0045	7	6	8	3	3	3
0055	4	5	4	4	4	2
0058	7	5	7	1	0	1
0061	0	0	3	2	0	1
0062	6	6	6	2	1	1
0063	3	4	6	2	1	2
0065	5	5	4	4	4	4
0068	5	2	1	3	1	0
0070	5	6	7	4	3	4
0071	2	0	5	1	1	2
0073	6	7	0	4	4	1
0074	5	4	1	4	4	1
0076	2	4	7	0	0	0
0080	5	2	3	3	1	2
0082	6	4	6	2	1	1
0084	0	1	0	4	1	1
0086	3	2	8	2	2	2
0088	8	6	5	4	2	2
0091	10	2	0	5	1	2
0095	5	1	3	3	0	3
0097	7	6	4	5	5	5
0098	6	4	6	4	3	4
0099	5	3	0	3	3	2

AP: Average Precision, RP: R-Precision,
RR: Relevant Retrieved