On a Combination of Probabilistic and Boolean IR Models for WWW Document Retrieval

MASAHARU YOSHIOKA and MAKOTO HARAGUCHI
Hokkaido University

Even though a Boolean query can express the information need precisely enough to select relevant documents, it is not easy to construct an appropriate Boolean query that covers all relevant documents. To utilize a Boolean query effectively, a mechanism to retrieve as many as possible relevant documents is therefore required. In accordance with this requirement, we propose a method for modifying a given Boolean query by using information from a relevant document set. The retrieval results, however, may deteriorate if some important query terms are removed by this reformulation. A further mechanism is thus required in order to use other query terms that are useful for finding more relevant documents, but are not strictly required in relevant documents. To meet this requirement, we propose a new method that combines the probabilistic IR and the Boolean IR models. We also introduce a new IR system—called appropriate Boolean query reformulation for information retrieval (ABRIR)—based on these two methods and the Okapi system. ABRIR uses both a word index and a phrase index formed from combinations of two adjacent noun words. The effectiveness of these two methods was confirmed according to the NTCIR-4 Web test collection.

Categories and Subject Descriptors: H.3.3 [**Information Systems**]: Information Search and Retrieval—*Query formulation*; *Retrieval models*; *Relevance feedback*

General Terms: Information Retrieval, Algorithms, Experimentation

Additional Key Words and Phrases: Boolean IR model, probabilistic IR model

1. INTRODUCTION

Large quantities of textual data can now be accessed through the Internet and many search engines have been implemented for commercial use. Most such systems mainly use simple Boolean query operators such as "+" with terms that should be included in retrieved documents and "−" with terms that should be excluded. However, most users have great difficulty specifying appropriate queries in Boolean format [Hearst 1999; Young and Shneiderman

Authors' address: M. Yoshioka and M. Haraguchi, Graduate School of Information Science and Technology, Hokkaido University, North 14 West 9, Kita-ku, Sapporo-shi, Hokkaido, 060-0814 Japan; email: yoshioka@ist.hokudai.ac.jp.

1993]. As a result, according to an analysis of real queries used in the AltaVista search engine, most users only use implicit AND, and do not use such Boolean features [Spink et al. 2001]. In addition, Eastman and Jansen [2003] analyzed the impact of these query operators for different search engines (e.g., Google, AOL, and MSN); they found that most user-defined operators do not improve search engine results, except for some PHRASE operator cases.

Because of the difficulties in constructing effective Boolean queries, these search engines use Boolean queries only as filters that select document sets for ranking and use information other than content information, such as link structures, for ranking to find highly relevant pages as highly scored ones. PageRank [Brin and Page 1998] is a famous algorithm for that purpose and it works well for the user who wants to find one or a few appropriate pages.

In contrast, there are professional users who would like to retrieve exhaustive (high-recall) documents for particular tasks (e.g., compiling surveys). An IR system that uses a Boolean query as a filter for selecting a document set for ranking may miss many relevant documents when the given Boolean query is not appropriate enough. To utilize a Boolean IR model, it is, therefore, desirable to have a support mechanism for constructing an appropriate Boolean query that represents his/her information need. Several studies have considered support of Boolean query formulation [Anick et al. 1990; Young and Shneiderman 1993; Jones 1998]. In those studies, Boolean query reformulation process was supported by showing interactively how retrieved results would change with a revised query, but they did not focus on how to reformulate a query according to relevant documents.

In the current study, we propose a method for modifying a given Boolean query by using information from a relevant document set. This method is based on the assumption that the (pseudo-) relevant document set should satisfy the newly constructed Boolean query. However, some important keywords may be excluded as a result of this query reformulation process, thereby causing difficulties when searching for relevant documents that contain the excluded keywords. To compensate for this difficulty, we also propose a new method that combines the probabilistic IR model and the Boolean IR model.

Based on these two methods, a new IR system, "appropriate Boolean query reformulation for information retrieval" (ABRIR) is introduced. This system uses a modified version of the Okapi system as a probabilistic IR engine. It uses both a word index and an index of phrases comprising combinations of two adjacent words.

The rest of this paper is divided into five sections. Section 2 introduces our baseline probabilistic IR model based on the Okapi system and discusses its characteristics. In Section 3, two methods that combine the probabilistic and Boolean IR models are proposed. In Section 4, our new IR system, ABRIR is described and evaluated by using the NTCIR-4 Web test collection [Eguchi et al. 2004]. Section 5 compares our approach with related work; Section 6 concludes the paper.

2. AN IR SYSTEM BASED ON THE PROBABILISTIC IR MODEL

Our baseline system is mostly similar to Okapi BM25 [Robertson and Walker 2000] with pseudo-relevance feedback and query expansion and to the IR system proposed by Toyoda et al. [2002], which was the highest-performing system in the NTCIR-3 workshop. However, we have modified several aspects of these existing systems, so these modifications in our system are briefly explained in this section.

Our system is designed to handle mainly Japanese documents. It uses BM25 [Robertson and Walker 2000] as a basic probabilistic IR model and ChaSen [Matsumoto et al. 2000] as a morphological analyzer to extract index terms. It uses a word index and a phrase index comprising combinations of adjacent words [Toyoda et al. 2002]. Like Uchiyama and Isahara's system [Uchiyama and Isahara 2001], it employs pseudo-relevance feedback and query expansion by using the five top-ranked documents retrieved initially. The generic engine for transposable association (GETA) tool is used[1] as a database engine. Indexes for documents in other languages totally depend on the results of ChaSen.

2.1 Indexing Each Document

Japanese text can have different coding systems such as Shift-JIS, EUC, and UTF-8, so before applying the indexer we converted all texts into EUC code. We also removed HTML tags from documents that contained them.

After these preprocessing steps, we applied the following procedure to extract the word and phrase indexes from the text.

1. Morphological analysis—ASCII text characters are converted into 2-byte EUC codes by using KAKASI[2] as a code converter and ChaSen as a morphological analyzer.
2. Extraction of index terms—Noun words (nouns, unknowns, and symbols) are extracted as index terms. We excluded numbers, prefixes, postfixes, and pronouns from the index terms. We removed "—" from the end of a term when the length of the term was longer than two katakana characters. All alphabetical letters were then normalized to 1-byte ASCII codes and stored in lower case.
3. Extraction of phrasal terms—The aim was to use compound nouns as phrasal terms extracted from pairs of adjacent nouns. In addition, prefixes, postfixes, and numbers were used for extracting phrasal terms.

2.2 Pseudo-Relevance Feedback and Query Expansion

The five top-ranked documents were used for pseudo-relevance feedback. However, when the score was normalized by the number of terms existing in the document, some texts with fewer terms tended to score highly. For example, when the single query term was "TOEIC," a document containing only the term

[1]http://geta.ex.nii.ac.jp/
[2]http://kakasi.namazu.org/

"TOEIC" scored highly. This may occur, for example, when the title of an HTML page is "TOEIC" and the contents are Macromedia Flash or image objects. Because these documents are not useful for query expansion, we excluded documents containing fewer than four terms from the relevant documents list.

Pseudo-relevant documents are also used as a source of query expansion. The computational cost increases when large numbers of expanded terms are added, so expansion was restricted to 300 terms. If there were more than 300 different terms in a pseudo-relevant document set, the 300 different terms with the highest mutual information content between a relevant document set and a term would be selected [Yoshioka and Haraguchi 2003].

2.3 Term Weighting

The BM25 weighting formula was used to calculate the score for each document:

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

Here, $w^{(1)}$ is the weight of a (phrasal) term T which is a term or a phrasal term in query Q, and is calculated using Robertson-Sparck Jones weights:

$$w^{(1)} = log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \tag{2}$$

where N is the count of all documents in the database, n is the count of all documents containing T, R is the given number of relevant documents, and r is the count of all relevant documents containing T. In addition, tf and qtf are the number of occurrences of T in a document and in a query, respectively, and k_1, k_3, and K are control parameters.

The results of term extraction in our system may vary because of the results of the morphological analyzer. The effect of this difference must, therefore, be minimized.

For example, suppose a phrasal term "情報科学 (information science)" ("情報 (information)" + "科学 (science)") exists. When "情報科学 (information science)" is registered in the dictionary of the morphological analyzer, the term "情報科学 (information science)" is extracted. When "情報 (information)" and "科学(science)" are registered separately, but "情報科学 (information science)" is not registered, the terms "情報 (information)" and "科学 (science)" and a phrasal term "情報科 学 (information science)" are extracted. In the latter case, in addition to "情報科 学 (information science)," the terms "情報 (information)" and "科学 (science)" are also used to calculate the score.

Phrasal terms should, therefore, have lower weights than regular terms. For this purpose, we introduced a parameter $c(0 \leq c \leq 1)$ for counting the phrasal terms in a query, where qtf is incremented by c, rather than one when a phrasal term is found.

For the query expansion, Rocchio-type feedback was used [Uchiyama and Isahara 2001]:

$$qtf = \alpha qtf_0 + (1-\alpha) \frac{\sum_{i=1}^{R} qtf_i}{R} \tag{3}$$

where qtf_0 and qtf_i are the number of times T appears in the query and in relevant document i, respectively.

To estimate parameters, we conducted retrieval experiments using the NTCIR-3 Web test collection, and we set $k_1 = 1, K = \frac{dl}{avdl}, c = 0.3, \alpha = 0.7$. Here, dl is the length of a document (the number of terms and phrasal terms) and $avdl$ is the average length of all documents. We set $k_3 = 1000$ for initial retrieval and $k_3 = 7$ for final retrieval.

2.4 Retrieval Procedure

The retrieval procedure used in our IR system is as follows.

1. Morphological analysis—An identical morphological analysis process was applied to generate an index of each document and to extract terms and phrases for the query.
2. Initial retrieval—The query was applied to obtain the top-ranked documents. We set $R = r = 0$ to calculate the score of each document.
3. Pseudo-relevance feedback and query expansion—The five top-ranked documents were selected as the relevant documents. When this set included documents that had fewer than four terms, it removed them from the relevant documents list and included the next higher-ranked documents.

   We did not use phrasal terms for the query expansion because they may be too specific for use with pseudo-relevance feedback [Toyoda et al. 2002]. When there were many terms in the relevant documents, the 300 terms that shared the highest mutual information were selected [Yoshioka and Haraguchi 2003].
4. Final retrieval—The expanded query was applied to obtain the final results.

2.5 Implementation

We implemented the baseline IR system using the generic engine for transposable association (GETA) tool.

This system has almost equivalent retrieval performance in terms of mean average precision to the highest-performance IR system in NTCIR-3 [Toyoda et al. 2002], which is based on an Okapi BM25.

Because GETA cannot handle all documents as a single database, the documents were divided into eight subsets. To obtain an equivalent score from all databases N, n, and $avdl$ were shared. A given query was applied to all eight databases and the results were merged.

2.6 Evaluation

We used the NTCIR-4 Web test collection [Eguchi et al. 2004] to evaluate the system. This collection contains 100 gigabytes of document data, 35 queries for survey-type retrievals, and 45 queries for target-type retrievals. Figure 1 shows a sample topic of this test collection. <TITLE> includes 1–3 terms with Boolean expressions. Attribute "CASE" in <TITLE>, <ALT0>, <ALT1>, <ALT2>, and <ALT3> has the following meanings.

<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あり、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 a sample topic

Fig. 1. A sample topic of NTCIR-4 web test collection. [Eguchi et al. 2004].

(a) All of the terms have relationships with one another that can be used as an OR operator.
(b) All of the terms have relationships with one another that can be used as an AND operator.
(c) Only two of the terms have a relationship that can be used as an OR operator and they are specified by the attribute of "RELAT."

For the sample topic described in Figure 1, we can construct the Boolean query (オフサイド(offside) and (サッカー(soccer) or ルール(rule))) from TITLE and (オフサイドフサイド (offside) and サッカー (soccer) and ルール (rule)) from ALT3.

In this test collection, assessors judged the "multigrade relevance" of documents, i.e., highly relevant (S), fairly relevant (A), partially relevant (B) or irrelevant (C). Rigid relevance levels, where "S" or "A" documents were classified as relevant, were used for overall evaluation in this paper.

Table I lists the evaluated results from our IR system. Survey-type experiments were conducted with 35 topics selected by the organizers and target

Table I. Evaluation of Results from Our System[a]

	AvePrec	RPrec	Prec@10	Prec@20
tt (s)	0.223	0.254	0.411	0.361
ds (s)	0.200	0.234	0.383	0.341
tt (t)	0.215	0.232	0.344	0.306
ds (t)	0.235	0.242	0.378	0.333

[a]tt, title only; "ds," description only; s, survey type; "t," target type; AvePrec, average precision; "RPrec," R precision; Prec@10; Prec@20, Precision at 10, 20 documents.

types with another 45 topics selected by the organizers. A thousand documents were retrieved for every topic.

In most cases, our system DBLAB-tt-02, DBLAB-ds 02 in [Eguchi et al. 2004] has one of the highest retrieval performances. However, in several cases, it has poorer performance than average.

We assume that the quality of phrasal terms used in a query may affect the retrieval performance. For example, topic 0058 uses the terms "存在論 (ontology)" = "存在 (onto-)"+ "論 (-logy)" in the title.

In contrast, it uses "「存在とはなにか」ついて哲学的観点から… (from the philosophical aspect, (find documents that explain) "What is (onto-, existence)" that includes "哲学的観点 (philosophical aspect)" = "哲学 (philosoph-)" + "的 (-cal)" + "観点 (aspect)" in the description.

Because "存在論 (ontology)" is a technical term in philosophy and artificial intelligence, "存在論 (ontology)" is a more appropriate word than "存在 (onto-, existence)." On the other hand, because "哲学的観点 (philosophical aspect)" is more important than "存在 (onto-, existence)," which is a common word, our system tends to neglect "存在 (onto-, existence)."

The difference between these terms causes the quality of the initially retrieved results to vary, so the final results for retrieving the description are worse than average, but the final results for retrieving the title are better than average.

Another problem arises from pseudo-relevance feedback with irrelevant and similar document sets. In topic 0006, the system retrieves quite similar documents (NW002999258, NW002999245, NW002999257, NW002999256, and NW002999253) that contain formatted record data. Because these five documents have a similar term list, our query expansion method generates a bad query. To reduce the effect of irrelevant documents, we believe that it is better to check for similarity among the top-ranked documents and to remove similar documents from the query expansion. A further problem arises from our indexing method. Topic 0034 uses the following three terms "料理 (cooking)," "切り方 (cutting method)," and "名称 (name)" in the title.

Because we do not use verbs for indexing, we do not identify "切り方 (cut + -[t]ing method)" as an index term in our system, so the retrieved results for topic 0034 are poor. There are two possibilities for including "切り方 (cutting method)" as an index term. The first is to include verbs as index terms; the second is to include phrasal terms made with noun postfixes. Because "方 (method)" is a noun postfix, "切り方 (cutting method)" can be included as the phrasal term "切り (cut)" +"方 (-[t]ing method)".

3. COMBINATION OF A PROBABILISTIC IR MODEL AND A BOOLEAN IR MODEL FOR QUERY REFORMULATION

There are three major IR model types: a probabilistic model such as the one on which our proposed IR system is based, a vector-space model, and a Boolean model [Baeza-Yates and Ribeiro-Neto 1999]. The most distinctive differences between the Boolean model and the other models are the assumptions about the appropriateness of selected IR query terms.

For example, a probabilistic model and a vector-space model may retrieve documents that do not contain the user-specified query terms. In contrast, a Boolean model assumes that the user will select appropriate terms and it retrieves only documents that contain the user-specified required query terms, but do not contain the user-specified query terms with a NOT operator.

However, it is not easy to formulate an appropriate Boolean query. For example, some user-formulated Boolean queries defined in this test collection are not precise enough for retrieving all relevant documents, as we have shown in the retrieval results for the Boolean query (see Section 4 for detail).

In this research, we, therefore, propose a new IR system ABRIR based on the following two new proposed methods.

- A method for formulating a Boolean query that includes more relevant documents, by using information about relevant documents.
- A method for combining a probabilistic IR model and a Boolean IR model.

Each approach is discussed in details in the following sections.

3.1 Reformulation of a Boolean Query Based on Relevant Documents

Because we assume that all relevant documents contain words that the user intends to retrieve, words that exist in all relevant documents were selected. To remove common words, only words that exist in the original query were used. The original Boolean query was reformulated by using these words.

The following procedure is used to reformulate a Boolean query. Figure 2 shows an example of this process.

1. Selection of Boolean candidate words. All terms used in the original query that also exist in all relevant documents are selected. A Boolean query is reformulated by using the selected words with the AND operator. In this example, “A” and “C” exist in all relevant documents, so “A and C” is selected as a candidate query.
2. Reformulation of the Boolean query based on the initial query. When an original Boolean query has been created, it is relaxed. When there are one or more words in the initial query that are used within an OR operator, the generated query is expanded by using this OR operator information. In this example, because “C or D” exists in the original query, the generated query is modified to “A and (C or D).”

Because the description query does not have an original Boolean query, the first step only is applied to generate a new Boolean query.

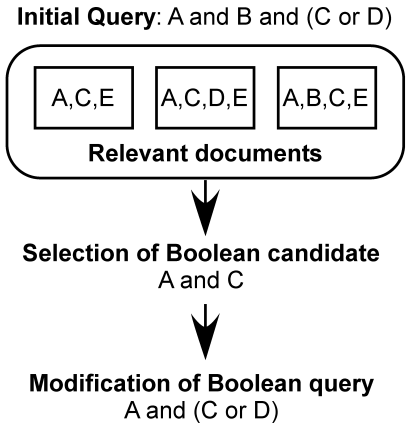

Fig. 2. Boolean query construction.

We think that the methodology proposed here is applicable not only for finding the N top-ranked pseudo-relevant documents, but also for user-selected relevant documents. However, the meaning of this reformulation procedure varies with the nature of the relevant documents.

When we use user-selected relevant documents, the meaning is simple. Because user-selected documents should be included in the retrieved set, it is necessary to reformulate a Boolean query so that it can be satisfied, at least, by all the selected documents. However, because this Boolean query is not compared with all relevant documents, it may not be sufficient for all.

In contrast, when the N top-ranked pseudo-relevant documents are used, the meaning is different. In this case, the method deals with the co-occurrence patterns of the given query terms. When term co-occurrences described by an initial Boolean query are frequent, the N top-ranked pseudo-relevant documents may satisfy the Boolean query. However, if such term co-occurrences are rare, we must modify the Boolean query. Because we use BM25 term weighting for initial retrieval, n in Eq. (2) (the count of all documents containing the term "T") affects the score of each term. Therefore, when we assume that query terms are independent of each other, this algorithm tends to exclude high-occurrence terms (common terms) from the new Boolean query.

3.2 Modification of the Score Based on the Boolean Query

A query expansion technique based on relevance feedback improves retrieval performance, because the expansion may include terms that can specify a desired document domain and/or terms that may help to find similar terms based on co-occurrences. These terms are not necessarily included in relevant documents but are useful for finding new relevant documents that are difficult to find with the original query.

However, when a query's terms are expanded by using relevant documents in the probabilistic IR model, there is a chance that documents without all the required query terms will receive a higher score than documents with these terms, although the original query includes terms that are necessarily included in

relevant documents. In such cases, conventional query expansion may degrade retrieval performance.

Because we assume that documents that do not satisfy the Boolean query may be less appropriate than documents that do satisfy it, a penalty score is subtracted from documents that do not satisfy the Boolean query.

The penalty is applied according to the importance of the word. For a probabilistic IR model, the BM25 weighting formula was used to calculate the score of each document (Eq. 1). In this equation, $w^{(1)}\frac{(k_3+1)qtf}{k_3+qtf}$ shows the importance of the word in the query. A control parameter β is used to calculate the penalty score.

$$Penalty(T) = \beta * w^{(1)}\frac{(k_3+1)qtf}{k_3+qtf} \tag{4}$$

For the OR operator, we use the highest penalty from all the OR terms as the overall penalty.

We describe how to calculate the penalty, using, as an example, the Boolean query ("A" and ("C" or "D")) given in Figure 2. First, we calculate the penalty score for all words ("A," "C," and "D"). In this case, we assume $Penalty(C) \geq Penalty(D)$. Documents not possessing terms "A," "C," or "D" receive the penalty $Penalty(A) + Penalty(C)$. Documents possessing only the "C" term receive $Penalty(A)$.

4. ABRIR (APPROPRIATE BOOLEAN QUERY REFORMULATION FOR INFORMATION RETRIEVAL)

4.1 Implementation

We implement ABRIR based on our baseline IR system discussed in Section 2 (BM25 + pseudo-relevant feedback + query expansion by using terms in pseudo-relevant documents (max 300).

The GETA tool has a mechanism for applying the Boolean AND operator, but not for applying the Boolean OR operator by itself. In previous experiments, when the system retrieved the number of top-ranked documents for each database, we could find the desired number of top-ranked documents for the total database. However, if we apply the Boolean OR operator to the retrieved results and reject documents from them, the resulting number of top-ranked documents for each database may not be large enough to retrieve the desired number of documents from the entire database. To reduce the effect of this problem, we, therefore, add a margin for retrieved document numbers.

4.2 Evaluation

We also applied ABRIR to the NTCIR-4 Web test collection. That is, we constructed initial Boolean queries from the topic descriptions for the title-retrieval task. When given terms were split into two or more index words by ChaSen, the last phrase was used for an initial Boolean query in order to avoid constructing complicated Boolean queries. For example ("利用者 (利用 (use)-者 (-er))" or 新人

Table II. Evaluation Results for Our System with Boolean Reformulation (Survey)[a]

	AvePrec	RPrec	Prec@10	Retrieved	Total
tt-b	0.200	0.236	0.431	1843	23488
tt-o	0.153	0.184	0.374	1685	17927
ds-b	0.155	0.196	0.370	1327	22534

[a] tt-b, title only with Boolean reformulation; tt-o, title only by using original Boolean query; ds-b, description only with Boolean reformulation; Retrieved, number of relevant retrieved documents.

Table III. Evaluation Results for Our System with Boolean Reformulation (Target)[a]

	RPrec	Prec@5	Prec@10	Retrieved	Total
tt-b	0.255	0.382	0.371	1451	23613
tt-o	0.247	0.400	0.378	1390	17470
ds-b	0.246	0.422	0.387	1166	19683

[a] tt-b, title only with Boolean reformulation; tt-o, title only by using original Boolean query; ds-b, description only with Boolean reformulation; Retrieved, number of relevant retrieved documents.

研究者 (新人 (new)-研究 (research)-者 (-er))") is a query described in the topic, an initial Boolean query is ("!c 利用者 (user)" or "!c 研究者 (researcher)").[3]

Tables II and III list the results of this experiment. Eighteen Boolean queries from 35 survey-type topics and 19 Boolean queries from 45 target-type topics were modified. We used the following three types of reformulation:

- *Removal of terms*. Remove term(s) (13 from survey and 13 from target);
- *Break phrase into word*. Remove phrase(s) and add terms that are parts of the phrase(s) (four from survey and six from target)
- *Add terms*. Add new phrases and/or terms that are excluded from the initial Boolean construction process (five from survey and five from target)

The first two reformulations relax the initial Boolean query, the last one strengthens the initial Boolean query.

Since the Boolean query obtained by the proposed method retrieves more documents that the original user-constructed Boolean query does, it is confirmed that the original Boolean query is stricter than the constructed query. There were 158 more "Retrieved" documents from the survey task (1843 − 1685 from 35 topics: 3893 relevant documents in all), and 61 more from the target task (1451 − 1390 from 45 topics: 2891 relevant documents in all).

In the baseline system, 1000 documents are retrieved for each topic. However, because we restrict the retrieved results by using a Boolean query, there is a chance that the system cannot retrieve 1000 documents. The "Total" column in Tables II and III shows the number of retrieved documents for all topics. Comparing the increased number of total retrieved documents and that of retrieved relevant documents, shows that this Boolean reformulation works well for survey topics (158 from 5561), but not for target topics (61 from 6143).

[3] "!c" is a prefix for the phrase index.

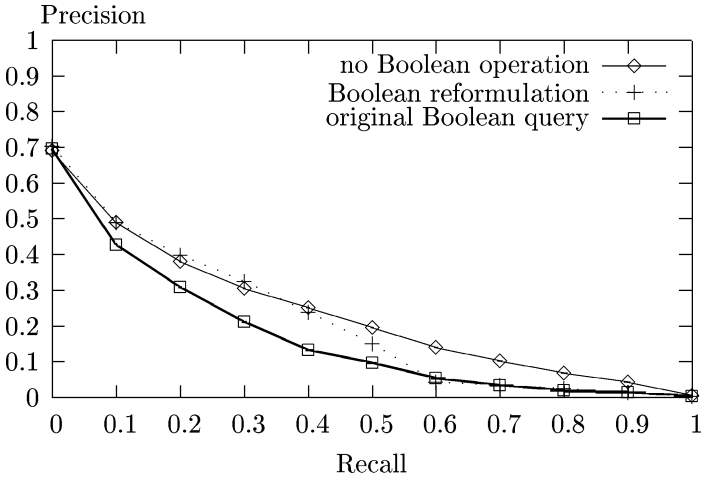

Fig. 3. Recall-precision graph for different Boolean queries (title only, survey).

We think that this result comes from the method for constructing relevant document data. Relevant document data of the target-type retrieval task are generated by using pools that are constructed by the top 20-ranked documents from each run results and that of survey-type one uses top 100-ranked documents [Eguchi et al. 2004]. There is, therefore, less chance of including relevant documents with complementary terms (e.g., synonyms) in target-type than in survey-type retrieval.

In addition, the precision of higher-ranked documents (Prec@5 and Prec@10) for target-type retrieval is no better than that of the original query. In target-type retrieval, the task is to find a limited number of appropriate pages from the query, and the precision of higher-ranked documents is important. Because our Boolean query reformulation method mostly relaxes the original Boolean query to achieve higher recall, it may reduce the precision of the original Boolean query, thus reducing the precision of higher-ranked documents. We, therefore, think that this strategy is useful mainly for survey-type retrievals and not for target-type ones.

When we compare the above results with those from the probabilistic IR model only (Table I), it becomes clear the developed system performs worse for "Average Precision" and "RPrec" values. This problem arises because of the difference in the number of relevant retrieved documents [for our baseline system: tt (s) 2166, ds (s) 2177, tt (t) 1843, and ds (t) 1616], and implies that the given Boolean query is not precise enough to represent the user's information need.

Figures 3 and 4 show the recall-precision graph of the retrieved results when using different Boolean queries for survey-type retrieval. This Boolean query reformulation method improves the performance precision, especially for smaller recall values.

We assume that this improvement is a result of removing documents that have expanded query terms with higher Robertson-Sparck Jones weights and

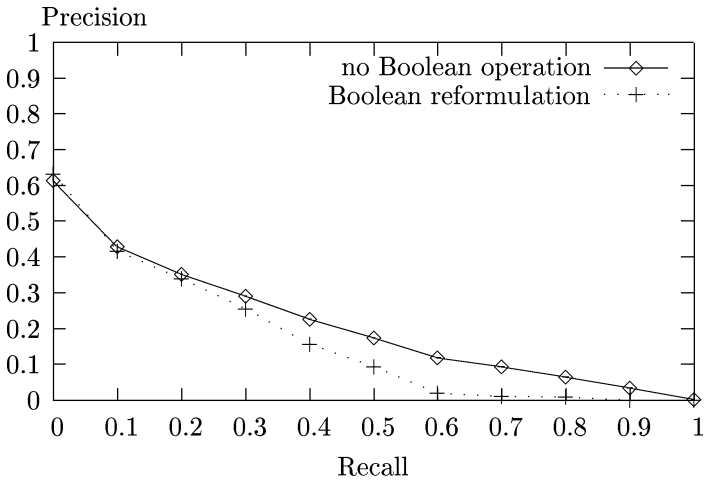

Fig. 4. Recall-precision graph for different Boolean queries (description only, survey).

Table IV. Evaluation Results for Our System with Penalties (Survey)[a]

	AvePrec	RPrec	Prec@10	Prec@20
tt-0.2	0.229	0.255	0.420	0.364
tt-1.0	0.241	0.263	0.431	0.376
tt-2.0	0.241	0.265	0.429	0.380
ds-0.2	0.207	0.235	0.391	0.349
ds-1.0	0.218	0.242	0.389	0.346
ds-2.0	0.211	0.237	0.394	0.346

[a]tt-β, title only $\beta = 0.2, 1.0, 2.0$; ds-β, description only $\beta = 0.2, 1.0, 2.0$.

Table V. Evaluation Results for Our System with Penalties (Target)[a]

	RPrec	Prec@5	Prec@10	Prec@20
tt-0.2	0.245	0.373	0.342	0.311
tt-1.0	0.255	0.373	0.364	0.327
tt-2.0	0.256	0.382	0.358	0.322
ds-0.2	0.251	0.427	0.373	0.338
ds-1.0	0.258	0.418	0.382	0.337
ds-2.0	0.263	0.422	0.382	0.337

[a]tt-β, title only $\beta = 0.2, 1.0, 2.0$; ds-β, description only $\beta = 0.2, 1.0, 2.0$.

do not have initial query terms with smaller Robertson-Sparck Jones weights. However, in collecting all relevant documents, this restriction is too strict and this is one reason why we have poor performance for higher recall values.

We also conducted retrieval experiments using the score-modification method. Because the constructed Boolean queries perform better than the original queries, we use them for calculating the penalties. Tables IV and V show the results for this method with different β values.

These results confirm that the penalty calculation improves the retrieval results.

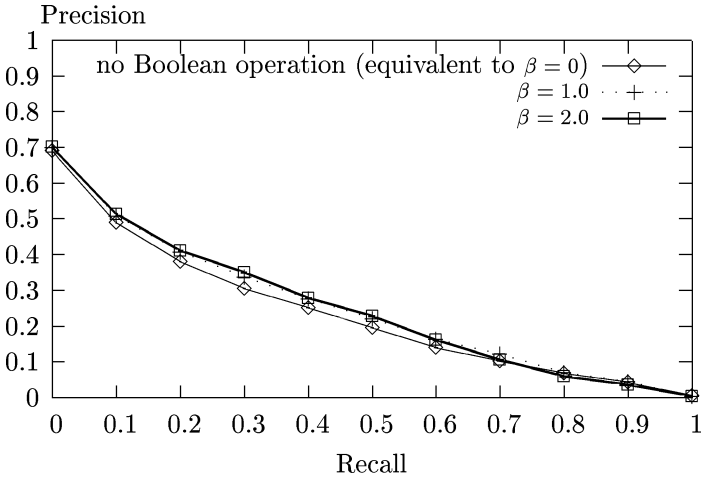

Fig. 5. Recall-precision graph for different β (title only, survey).

Because we aim to make a recall-oriented system, we mainly use "AvePrec" and "RPrec" for the survey-type task. For the target-type task, we mainly use "Prec@5" and "Prec@10," because this task mainly focuses on the precision of the lower recall value.

In the title-only experiment, the best performance was obtained when $\beta = 2.0$. In contrast, the results for $\beta = 1.0$ in the description-only experiment show better performance, in most cases, than that when $\beta = 2.0$. We assume that this difference comes from the quality of the given Boolean query, because our constructed Boolean query used for the description gives worse performance than the titles query in terms of the relevant retrieved document sizes. We, therefore, think that the estimation of an appropriate value for β should be based on a user model that has information on how a user may describe a Boolean query correctly.

Figures 5 and 6 show recall-precision graphs of the retrieved results for different values of β for survey tasks. This method improves performance, especially for recall values of 0.1 to 0.6.

Unlike the strict Boolean model, because this model does not remove documents that do not have the required initial terms, there is a chance that the system will find relevant documents that have complementary terms (e.g., synonyms) of the initial required query terms that do not exist in the document. On the other hand, documents that do not satisfy a Boolean query have less chance of being considered as relevant documents, so the penalty works well in reducing the importance of such documents.

5. RELATED WORKS

There have been several studies on Boolean query formulation [Anick et al. 1990; Young and Shneiderman 1993; Jones 1998] with graphical user interfaces. Their main support methods for Boolean query formulation consisted of showing how retrieved results changed as the user interactively modified

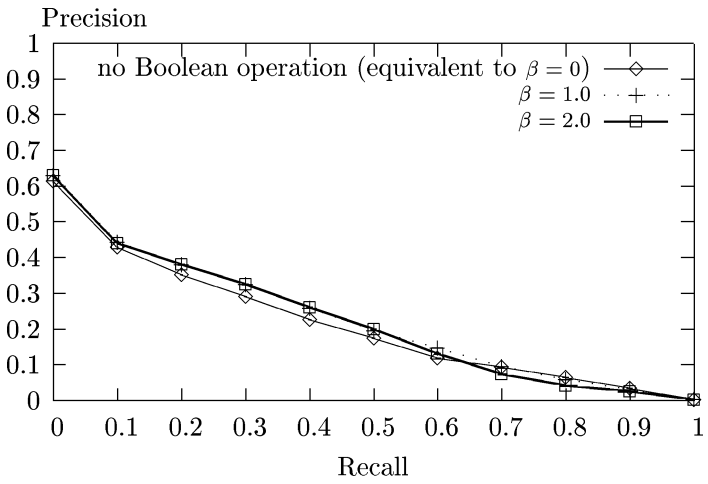

Fig. 6. Recall-precision graph for different β (description only, survey).

the Boolean query. They did not focus on how to reformulate a Boolean query according to relevant documents.

The extended Boolean information retrieval model [Salton et al. 1983] is another approach to finding documents that do not satisfy a given Boolean query with a ranking algorithm. This method uses a metric space defined by the given query terms. However, this method itself did not distinguish query terms that are useful for Boolean-type operations from those that are useful for finding related documents. Shaw and Fox [1994] proposed a method that combines this extended Boolean IR model and a vector-space IR model. This approach is similar to our approach. However, they did not discuss how to reformulate a Boolean query into a more appropriate one.

Kekalainen and Jarvelin [1998] analyzed how query structure, which means the use of operators to express the relations between search keys and query expansion, affect retrieval performance. They found that query expansion was not very effective for a Boolean-structured query, while strong structures (synonym operator: SYN of INQUERY) with many query expansions, based on a thesaurus, achieved the highest performance. This research is similar to ours in that it combines Boolean operators with a probabilistic IR model. Their findings were similar to ours. For example, both studies confirmed that the Boolean IR model itself cannot achieve higher performance and query expansion also improves retrieval performance as a combination of a Boolean IR and a probabilistic IR model. However, there are two points of difference between this research and ours. One is that our system divides query terms into two groups: terms that are used to formulate a Boolean query and ones that are used only in the probabilistic IR model. The other point is that our methodology includes a Boolean query-reformulation process that is based on relevant documents and is different from thesaurus expansion.

Several studies have tried to find useful terms from retrieved results and/or information from relevant documents. Xu and Croft [1996] proposed

an automatic query expansion method that used relevant document information. RU-INQUERY [Koenemann and Belkin 1996] and DualNavi [Takano et al. 2001] are interfaces for selection of keywords that are useful for characterizing relevant documents. Scatter/Gather's approach [Cutting et al. 1992] is to select useful terms by using text clustering information.

These systems are good for adding new query terms that are useful in finding new relevant documents. However, because they do not focus on Boolean query reformulation, they do not consider whether those terms should be strictly contained in relevant documents or not. As a result, their query expansion corresponds to the simple query expansion method used in our baseline system. This means that we could employ this method to improve the quality of query expansion in our system.

6. SUMMARY

We developed an information retrieval (IR) System—called appropriate Boolean query reformulation for information retrieval (ABRIR)—that modifies a given Boolean query by using relevant documents and combines the probabilistic IR model and the Boolean IR model. We confirmed that ABRIR improves the retrieval performance of our baseline system, which was one of the highest-performing retrieval systems among the NTCIR-4 Web task participants. We also confirmed that a user-constructed Boolean query is not precise enough to represent the information need and we proposed a method for Boolean query reformulation by using relevant documents to improve the retrieval performance. We confirmed that calculating a penalty based on the Boolean query improves the retrieval performance. In future work, we plan to use a thesaurus for constructing more expressive Boolean queries.

ACKNOWLEDGMENTS

We thank the organizers of the NTCIR Web Task for their efforts in constructing this test data. 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

ANICK, P. G., BRENNAN, J. D., FLYNN, R. A., HANSSEN, D. R., ALVEY, B., AND ROBBINS, J. M. 1990. A direct manipulation interface for Boolean information retrieval via natural language query. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Brussels, Belgium, 5–7 September 1990, Proceedings*, J.-L. Vidick, Ed. ACM, New York. 135–150.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*. Addison-Wesley, Reading, MA.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems 30*, 1–7, 107–117.

CUTTING, D. R., PEDERSEN, J. O., KARGER, D., AND TUKEY, J. W. 1992. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 318–329.

EASTMAN, C. M. AND JANSEN, B. J. 2003. Coverage, relevance, and ranking: The impact of query operators on web search engine results. *ACM Transactions on Information Systems 21*, 4, 383–411.

EGUCHI, K., OYAMA, K., AIZAWA, A., AND ISHIKAWA, H. 2004. Overview of the informational retrieval task at ntcir-4 web. In *Working Notes of the Fourth NTCIR Workshop Meeting*. http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/WEB/NTCIR4WN-OV-WEB-A-EguchiK.pdf.

HEARST, M. A. 1999. *Modern Information Retrieval*. Addison-Wesley, Chapter 10 User Interfaces and Visualization, 257–323.

JONES, S. 1998. Graphical query specification and dynamic result previews for a digital library. In *ACM Symposium on User Interface Software and Technology*. 143–151.

KEKALAINEN, J. AND JARVELIN, K. 1998. The impact of query structure and query expansion on retrieval performance. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 130–137.

KOENEMANN, J. AND BELKIN, N. J. 1996. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 205–212.

MATSUMOTO, Y., KITAUCHI, A., YAMASHITA, T., HIRANO, Y., MATSUDA, H., TAKAOKA, K., AND ASAHARA, M. 2000. *Morphological Analysis System ChaSen version 2.2.1 Manual*. Nara Institute of Science and Technology.

ROBERTSON, S. E. AND WALKER, S. 2000. Okapi/Keenbow at TREC-8. In *Proceedings of TREC-8*. 151–162.

SALTON, G., FOX, E. A., AND WU, H. 1983. Extended Boolean information retrieval. *Communications of the ACM 26*, 11, 1022–1036.

SHAW, J. A. AND FOX, E. A. 1994. Combination of multiple searches. In *Text REtrieval Conference*. 105–108.

SPINK, A., WOLFRAM, D., JANSEN, M. B. J., AND SARACEVIC, T. 2001. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology 52*, 3, 226–234.

TAKANO, A., NIWA, Y., NISHIOKA, S., HISAMITSU, T., IWAYAMA, M., AND IMAICHI, O. 2001. Associative information access using dualnavi. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*. 771–772.

TOYODA, M., KITSUREGAWA, M., MANO, H., ITOH, H., AND OGAWA, Y. 2002. University of tokyo/ricoh at ntcir-3 web retrieval task. In *Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering*. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-WEB-ToyodaM.pdf.

UCHIYAMA, M. AND ISAHARA, H. 2001. Implementation of an IR package. In *IPSJ SIGNotes, 2001-FI-63*. 57–64 (in Japan).

XU, J. AND CROFT, W. B. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 4–11.

YOSHIOKA, M. AND HARAGUCHI, M. 2003. Construction of personalized and purpose-oriented thesaurus. In *Proceedings of Asian Association for Lexicography '03 (ASIALEX)*. 461–466.

YOUNG, D. AND SHNEIDERMAN, B. 1993. A graphical filter/flow representation of Boolean queries: A prototype implementation and evaluation. *Journal of the American Society of Information Science 44*, 6, 327–339.

Received July 2004; revised May 2005; accepted May 2005