情報検索における検索語拡張手法の有用性分析手法の提案

吉岡 真治
北海道大学大学院情報科学研究科
国立情報学研究所

概要: 検索語拡張手法は、ユーザの持っている検索要求に応じた検索語を初期検索語に追加することにより検索性能を向上させる技術であり、多くのシステムで用いられ、その効果が確認されている。しかし、この手法が検索性能に与える効果は、様々な要因と関係しているため、検索語拡張手法の持つ性質を分析する方法に関しては十分議論されているとはいえない。そこで、本研究では、検索語拡張手法を分析するために、その有用性に影響を与えると考えられる要素の整理を行なう。また、実際の検索実験の結果に基づき、検索語拡張手法の分析を行なう。

Proposal of an Analysis Method for Evaluating the Effectivess of Query Term Expansion in Information Retrievval

Masaharu Yoshioka
Graduate School of Information Science and Technology, Hokkaido University
National Institute of Informatics

Abstract: Query term expansion is a technique to modify initial query by adding some terms that represents users' information need for achieving higher retrieval performance. However, a good framework for evaluating this technique in isolation does not exist because the effect of query expansion depends on several issues. In this paper, I enumerate several features that may affect the quality of this technique and analyze retrieval experiment results by using these features.

1 緒言

情報検索システムのユーザにとって、自分の欲しいと考えている検索要求を、検索語の組み合わせとして適切に表現することは、必ずしも容易ではない。そのため、検索要求と検索語の間のミスマッチによるギャップを埋めてくれる技術が、検索性能の向上のための重要な課題として研究が行なわれている。

検索語拡張は、ユーザが入力した検索語と、情報検索システムが検索に役立つであろうと推定した検索語を組み合わせて、情報検索を行なう手法であり、情報検索システムの性能向上のために広く用いられている技術である [1]。この検索語拡張の代表的な方法としては、(疑似) 関連文書中に含まれる語を利用する方法 [2, 3]、シソーラスなどの類義語情報を利用する方法 [4] などがある。

この検索語拡張が検索性能に与える影響に関しては、古くから議論がされているが [5, 4]、様々な検索課題に共通して有効に働くための明確なガイドラインは、まだ、得られていない[1]。

一方、検索課題ごとの検索語拡張が与える影響を考慮して、検索語拡張を行なうべきかどうかを判断する手法 [7, 8] も提案されており、検

[1] 例えば、Billerbeck ら [6] は、様々な検索システムにおいて、検索語拡張のパラメータを変更することにより、どの様に検索性能が変わるかを分析しているが、明確なガイドラインの作成にはいたっていない

索実験により、その有用性が確認されている。

ただし、これらの分析に用いられている指標の多くは、検索課題と文書データベース全体とのマッチングを評価した指標であり、検索要求と検索語の間のミスマッチを直接的に量る指標ではない。

一方、本研究では、これまでに、検索要求と検索語の間のミスマッチが検索性能に与える影響の分析を行なっているが [9]、検索要求と検索語の間のミスマッチの統計的指標と検索性能の間に強い相関は認められるものの、検索語拡張の有用性の全てを議論できる指標ではない。

よって、本論文では、これらの研究成果を踏まえ、検索語拡張の有用性を分析するにあたり、考慮するべき要因の整理を行なうと共に、NTCIR-4 の Web テストコレクション [10] を用いて、具体的な検索語拡張手法の有用性分析を行なう。

2 検索語拡張に影響を与える要因の分析

まず、最初に、検索語拡張に影響を与える要因について考えるにあたり、検索語拡張が検索性能の向上に役立つ理由について議論する。次に、検索要求と検索語の間のミスマッチについての議論を行なうと共に、検索語拡張の必要性という立場から、検索課題の特徴を分析するために役立つ指標を用いた検索課題の分析方法を提案する。

さらに、情報検索システムにおける適合文書フィードバックの基本的なモデルに基づき、検索語拡張の性能に影響を与えると考えられる要因について議論を行なう。

2.1 検索語拡張の効果

Buckley らは、適合文書フィードバックにより検索性能が向上する理由として、次のような効果を挙げている [11]。

1. 初期検索語の代替となる言葉を追加する (類義語)。
2. 関連する語を追加する (関連語)。
3. 多くの語を追加することにより、検索課題が持つ観点を表す。
4. 初期検索語の具体例となる語を追加する。
5. 検索語の重要度を適切に設定する。

このうち、最後の項目は、検索語の重み付けの話になるので、検索語拡張に直接関連する項目は 1.〜4. である。よって、これらの 4 つの理由に対して分析を行なうための評価基盤が必要となると考えられる。

上記の項目の内、1.2.4. については、シソーラスなどの情報を用いることにより、分析をすることも可能ではある。しかし、単純なシソーラスによる検索語拡張では、検索性能が向上しないことが確認されている [4] 事を踏まえると、適切な方法であるとはいえない。そのため、検索に役立つ検索語を分析するためには、ユーザの情報要求に応じた評価が不可欠である。

2.2 ブーリアンモデルを基礎とした検索課題の特徴分析

検索課題として与えられた検索語が十分に適切なものであるならば、検索語拡張は不必要である。一方、検索語が適切に選ばれていない場合には、検索語拡張の重要性が増す。

この検索語の適切さを分析するために、ここでは、検索課題として与えられているブーリアン式と適合文書リストの一致度を分析する。図 1 を用いて分析の方針を説明する。理想的なブーリアン式と適合文書リストの関係は、ブーリアン式を満たす文書リストが適合文書リストと一致する ((1) と (3) のサイズが 0) 状況である。

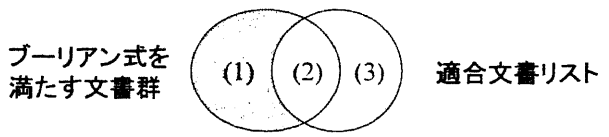

図 1: 検索課題と適合文書リストの一致度

本研究の以前の成果 [9] から、初期検索式として与えられたブーリアン式を精度の観点から評価した指標である $FocusAppropriateness = (2)/\{(1) + (2)\}$ の値と検索性能に強い正の相関があることが確認されている。

しかし、実際にこの様なブーリアン式を作成するのは困難であるため、(1) と (3) に相当する文書が存在する。一般に、(1) の割合が多い場合には、適合文書が十分に絞り込めていない状態を意味し、適合文書を絞り混むための検索語の拡張が必要な状況であると考えられる。これは、2.1 節の 2,3 に相当する検索語拡張が必要な課題だと考えられる。これに対し、(3) の文書が多い場合は、元のブーリアン式の制約が強すぎる状況であり、初期の検索語を補完する検索語の拡張が求められる。これは、2.1 節の 1,4 に相当する検索語拡張が必要な状況であると考えられる。

2.3 適合文書フィードバック

適合文書フィードバックとは、初期検索として与えられた検索語を適合文書の情報により修正する方法である。具体的には、適合文書中の語の出現確率などの情報に基づく、初期検索語の重要度を示す重みの修正や、適合文書中に存在する語を利用した検索語拡張がある。しかし、本研究の目的は、検索語拡張の分析にあるため、検索語の重みの修正に関しては、分析の対象としないこととする。

また、適合文書フィードバックの適合文書選択の手法としては、ユーザによる選択を行なう方法だけでなく、システムの初期検索の上位の文書を適合文書として取り扱う疑似関連文書フィードバックという方法が存在する。この疑似適合文書フィードバックは、多くの検索課題で有効に作用することが知られており、実際に与えている疑似適合文書に、本当の適合文書を含まないような場合においても、検索性能を向上させる場合がある。

この様な現象を分析するためには、2.2 節で述べた検索課題の特徴分析を組み合わせて分析を行なうのが有効であると考えられる。

例えば、絞り込みが不足しているようで、コンテキストを表すような検索語の追加が求められる検索課題については、初期検索の結果は、ユーザの情報要求にそったものである必要性が高く、疑似適合文書フィードバックにおける選択した疑似適合文書中の本当の適合文書の割合が高い事が望まれる。一方、初期の検索語を補完するような検索語が必要な検索課題においては、類義語や関連語などが見つけやすい文章であれば良く、必ずしも、初期検索の結果が、ユーザの情報要求にそったものでなくても構わない。逆に、適合文書であっても、類義語などを含まない文書であれば、適切ではない場合が考えられる。

2.4 適合文書群に特徴的な語

ある検索語がコンテキストを表す語として適切かどうかについては、適合文書群に特徴的に現れるか否かについて分析する方法が考えられる。本研究では、式 1 に示す、適合文書全体と文書集合全体における語の分布の異なりに注目した相互情報量に基づく指標 $G'(w)$[12] を用いて、この分析を行なう。

$$G'(w) = p(w|r)log_2\frac{p(w|r)}{p(w)} \quad (1)$$

この指標を適合文書中に含まれる全ての異なり語の値を計算し、その値の大きい方から順番に語を選択することにより、適合文書群に特徴的な語を選択することが可能になる。一方、理想的な検索課題においては、初期検索語が特徴的な語に含まれる事が望ましい。

よって、この特徴的な語と初期検索語の重なり具合いを見ることにより、初期検索語の適切性を量る指標になると考えられる。

3 検索語拡張の分析実験

3.1 検索実験の概要

前節の分析を踏まえ、実際の検索語拡張の結果を分析するための、要因に関するデータを収集し、そのデータと検索語拡張の有効性の関係について分析を行なった。

本研究では、NTCIR-4 の Web タスクで用いられている 35 のサーベイ検索の課題を利用してデータの分析を行なった [10][2]。本テストコレクションは、100GB の jp ドメインを中心に集められた Web ページの文書集合、検索課題、各々の検索課題についての多段階適合文書判定 (S:完全

[2]適合文書リストには、NTCIR-5 の検索語選択支援タスクのために追加判定した文書データを含む。

に適合、A：適合、B：部分的に適合、C：非適合)結果から構成されている。本実験では、SとAの二種類の判定の文書のみを適合文書として扱い、システムの評価を行なった。また、検索語などの統計量をとるためのベースラインとするシステムとして、[9] と同様に、確率型情報検索モデルOkapiとブーリアン情報検索モデルを組み合わせた情報検索システムであるABRIR(Appropriate Boolean query Reformulation for Information Retrieval) [13] を改良したシステムを用いた[3]。本システムはNTCIR-4参加者中で最も良い検索性能を実現しており、検索課題の難易度をはかるベースラインシステムとして有用であると考えている [10]。

本システムでは、初期検索結果に基づく、疑似適合文書を利用した適合文書フィードバックを行なっている。このフィードバックでは、検索語拡張の候補を幅広く考慮するために、一定の異なり語数 (本実験では、4語) を含まない文書を疑似適合文書の候補から削除している。また、同じテンプレートを用いて生成されたために、内容とは関係なく類似する単語を共通に含むようなページを除外するために、既に選択された文書と比較して類似している文書 (コサイン尺度を用いて評価) を候補から除外している。本システムでは、初期検索の上位の文書について、上記の条件を満たすかどうかを判定し、総数が5件になるまで、順番に文書を追加している。

また、検索語拡張の効果を調べるため、このABRIRにおいて、検索語拡張を全く行なわない検索実験と、相互情報量に基づく指標により疑似適合文書に特徴的に現れる10語,300語を検索拡張語として追加する実験を行なった。

全課題を総合した検索結果に関する基本データを表1に、課題毎に、検索語拡張を行なった場合と、検索語拡張を行なわなかった場合の性能の変化を分析した表を表2に示す (性能向上については、課題数が多いので、課題名の記述を省略)。

疑似適合文書として用いた文書の適合性判定は、表3の通りである。これと表2を比較すると、疑似適合文書が適合文書でなくても性能が向上する場合 (課題21「作家の値打ち and 福田和也」[4]) や、疑似適合文書が全て適合文書であった場合でも性能が悪化する場合 (課題76「ヴィトゲンシュタイン and 思想」,82「社会主義市場経済 and 中国」) などが存在することが確認できる。

表 3: 疑似適合文書の適合性判定結果

1	S(1),C(4)	65	S(1),A(4)
3	S(1),A(1),B(3)	68	C(5)
4	A(1),C(4)	70	S(2),A(1),B(2)
6	S(2),A(2),C(1)	71	A(2),C(3)
19	A(1),C(4)	73	A(1),B(2),C(2)
21	C(5)	74	S(1),B(1),C(3)
22	A(2),B(1),C(2)	76	S(1),A(4)
23	C(5)	80	A(1),B(1),C(3)
28	A(1),C(4)	82	S(1),A(4)
29	C(5)	84	B(1),C(4)
34	S(1),C(4)	86	A(1),B(2),C(2)
44	A(2),B(3)	88	S(1),A(3),C(1)
45	A(1),B(1),C(3)	91	C(5)
55	S(2),A(3)	95	S(1),A(2),C(2)
58	S(1),A(4)	97	A(2),C(3)
61	C(5)	98	S(1),B(1),C(3)
62	A(2),C(3)	99	A(2),C(3)
63	A(1),C(4)		

3.2 検索課題の特徴

次に、各検索課題を2.2節で議論した指標により、分析を行なった結果を示す。図2は、縦軸に、(2)/{(1)+(2)} を、横軸に、(2)/{(2)+(3)} をとり、丸の大きさにより、適合文書の大きさを示した散布図である。

この散布図から、検索課題ごとの検索式と検索要求の一つの表現である適合文書リストのギャップに様々なバリエーションがあることが見て取れる。

その結果として、役に立つ検索拡張語のタイプについても異なることが想像される。例えば、課題6「競馬 and 血統」は、ほとんどの適合文書がブーリアン式を満たすため、その中から適

[3]名詞を中心としたインデックスに加え、名詞性接尾辞と組み合わされた動詞・形容詞、ならびに、動詞をインデックスに追加した。

[4]利用した検索課題のタイトルから作成したブーリアン式を「」でしめす。

表 1: 検索結果に関する基本データ (全課題)

	検索語拡張 (300 語)	検索語拡張 (10 語)	検索語拡張なし
Retrieved	2720 / 3986	2591 / 3986	2422 / 3986
MAP	0.250	0.220	0.169
Prec@5	0.457	0.429	0.349

表 2: 検索性能の変化 (課題毎)

	性能の変化	検索語拡張 (300 語)	検索語拡張 (10 語)
Retrieved	向上	28 課題	24 課題
	同等	3 課題 (1,23,73)	9 課題 (1,4,19,61, 68,73,80,84,99)
	悪化	4 課題 (19,68,82,99)	2 課題 (71,74)
MAP	向上	25 課題	27 課題
	同等	0 課題	1 課題 (84)
	悪化	10 課題 (1,19,23,29,34,61, 68,71,76,82)	7 課題 (3,29,61,68, 71,74,76)
Prec@5	向上	16 課題	13 課題
	同等	14 課題 (1,4,23,34,44,55,58, 63,68,73,74,76,82,84)	18 課題 (1,3,4,23,34,44,55,58, 63,68,71,74,76,82,84,86,91,95)
	悪化	5 課題 (6,19,29,61,71)	4 課題 (6,19,29,61)

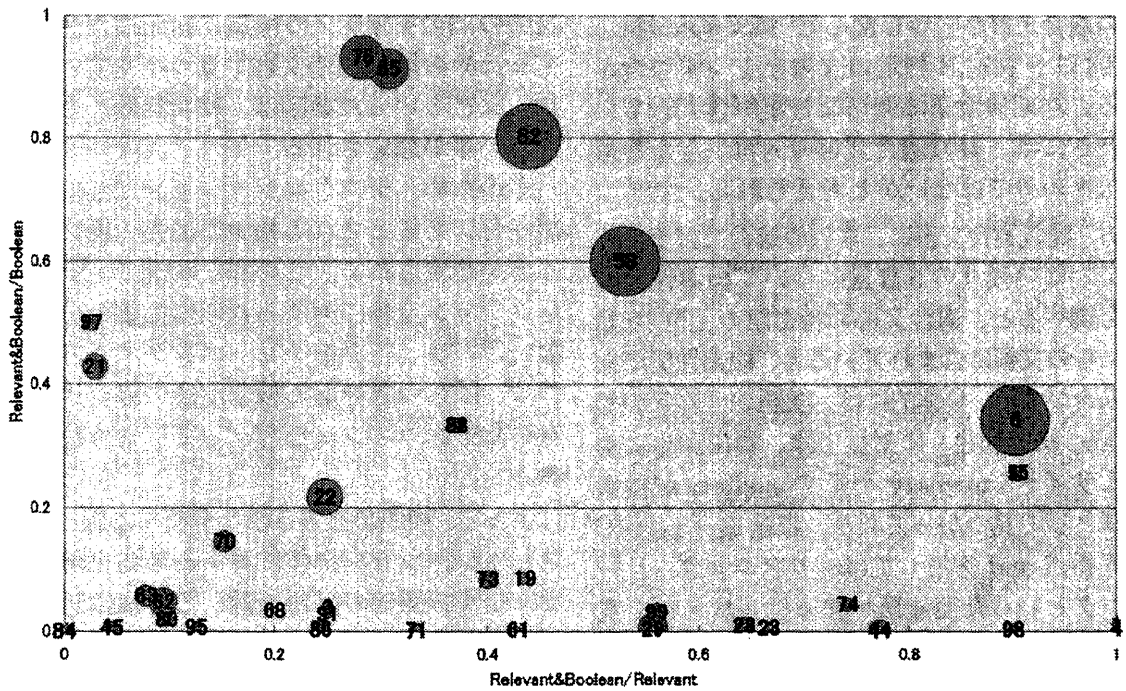

図 2: 検索課題の特徴

切なものを選びだすために、絞り込みのための検索語が役に立つと考えられる。逆に、ブーリアン式を満たしている文書のほとんどが適合文書であるが、その網羅性が不十分である課題76「ヴィトゲンシュタイン and 思想」などでは、初期検索語を補完する検索語の拡張が役に立つと考えられる。

この散布図と表2を比較しても、検索語拡張の手法が検索性能を向上させる課題と悪化させる課題に対する特徴的な関係を見いだすことは難しく、他の要因を考慮した分析が必要である。

次に、各々の検索課題に関して、2.4節で述べた適合文書の特徴語と初期検索語の比較を行なった。その結果、多くの検索課題では、初期検索語に適合文書の特徴語の上位五位までの語を含んでいたが、二つの課題(課題29「(バイオ or ソニー)and パソコン」 課題98「世界遺産 and 日本」)において、適合文書の特徴語と初期検索語の解離が見られた。これらの課題については、検索語拡張による性能改善を比較すると、課題29は、多くの指標で検索性能を悪化させているが、課題98では全ての指標で検索性能が向上している。ここで、表2を考慮すると、課題29は本当の適合文書を一つも含んでいないのに対し、課題98はS判定,B判定の文書を1つずつ含んでいる。初期検索語に適切な言葉が含まれていないときには、疑似適合文書の内容が与える影響は大きいのではないかと考えられる。

一方、疑似適合文書に、本当の適合文書を全く含んでいない残りの5課題について検索性能の変化を調べると、適合文書の特徴語の上位10語がこれらの文書に含まれていると、検索性能が向する場合が多い。具体的には、課題21「作家の値打ち and 福田和也」に対しては、福田和也に関係するサイト i-mediatv を示す mediatv が検索拡張語として見つかっており、課題91「TOEIC and 高得点 and 方法」では、リスニング、リーディングといった関連語が検索拡張語として見つかっており、各々、検索性能の向上に役立ったと考えられる。

また、疑似適合文書が全て適合文書であった場合においても、多くの検索語を追加すると、性能が悪化する場合(課題76「ヴィトゲンシュタイン and 思想」, 82「社会主義市場経済 and 中国」)があった。これらの2つの課題は、図2に示すように、初期検索語を補完する検索語の拡張が求められる課題であるため、適合文書であるかどうかよりも、補完する語を含むか否かが重要になったためであると考えられる。

3.3 考察

図2に示したように、本実験で用いた実際の検索課題の全てにおいて検索語と適合文書の間のギャップが存在する。そのため、本実験で分析した検索語拡張の手法は、多くの課題において有効に作用していることが確認された。

しかし、個別の検索課題について、分析を行なうと、検索性能が悪化している課題も見うけられた。これらの課題について、性能悪化の要因を分析したところ、その原因は、様々な要因に関連していることを確認した。

これは、少なくとも、実験した検索語拡張の手法においては、その有効性に影響を与える要因の中に、支配的な要因がないことを示しており、本研究で提案したような多角的な分析が必要であることを示していると考えている。また、本課題の分析結果を踏まえると、検索語と適合文書の間のギャップの種類、初期検索語を補完するための疑似関連文書が見つけやすいかどうか、などといった要因は、検索課題ごとの差が大きい事が想定される。

そのため、テストコレクションごとに、検索課題毎に必要とされる検索語拡張の種類が異なる事が想定される。この様な状況で、テストコレクション全体に対する検索語拡張の分析を行ない、パラメータをチューニングするという操作は、ばらつきのある検索課題の混合比率に依存したチューニングを行なってしまう危険性がある。

よって、検索語拡張手法を分析する際には、テストコレクション全体に対する平均を用いるのではなく、個別の課題に対して、検索語拡張がうまく行く場合、うまく行かない場合の要因を多角的に行なう必要があると考えている。

4 結言

本論文では、検索語拡張に影響を与える要因について考察を行ない、その要因として、検索

語と適合文書の間のギャップや、初期検索結果などをリストアップした。さらに、NTCIR-4のWebテストコレクション[10]を用いた具体的な検索語拡張手法の有用性分析により、これらの要因に基づく多角的な分析が不可欠であるという知見を得た。

ただし、今回の分析は、あくまでも、一つの検索語拡張手法のみを分析したものであり、この手法による影響が存在することは否定できない。よって、様々な検索語拡張手法について、同様の検討を行なうことにより、検索語拡張分析手法の一般性を検討する必要がある。そのため、現在、実行中のNTCIR-5のWebタスクにおけるサブタスクとして検索語選択支援タスクを行なっている[5]。本論文で提案した指標と様々なシステムにおける検索語選択の結果を比較することにより、特定の検索システムに依存しない形の分析を行ない、検索語拡張手法の有用性分析のための手法の確立を目指したいと考えている。

謝辞

NTCIRコレクションは国立情報学研究所の許諾を得て使用した。本研究の一部は、文部科学省科学研究費補助金(特定領域研究 課題番号16016201)によって実施された。

参考文献

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*, chapter 5 Query Operations, pp. 19–71. Addison-Wesley, 1999.

[2] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 206–214, New York, NY, USA, 1998. ACM Press.

[3] S. E. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In *Proceedings of TREC-8*, pp. 151–162, 2000.

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

[5] Peter Willett Helen J. Peat. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, Vol. 42, No. 5, pp. 378–383, 1991.

[6] Bodo Billerbeck and Justin Zobel. When query expansion fails. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 387–388, New York, NY, USA, 2003. ACM Press.

[7] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. A framework for selective query expansion. In *CIKM '04: Proceedings of the thirteenth ACM conference on Information and knowledge management*, pp. 236–237, New York, NY, USA, 2004. ACM Press.

[8] Tetsuya Sakai and Stephen E. Robertson. Relative and absolute term selection criteria: a comparative study for english and japanese ir. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 411–412, New York, NY, USA, 2002. ACM Press.

[9] 吉岡真治. 情報検索のテストコレクションにおける検索語の有用性に関する検討. 情報処理学会情報学基礎研究会, 2005-FI-79, pp. 25–30, 2005.

[5] http://research.nii.ac.jp/ntcweb/cfp-ntcir5web-ja.html

[10] Koji Eguchi, Keizo Oyama, Akiko Aizawa, and Haruko 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.

[11] Chris Buckley and Donna 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.

[12] 吉岡真治, 原口誠. 検索語の網羅性に注目した汎化概念により検索語選択支援を行う情報検索システムの研究. 人工知能学会論文誌, Vol. 20, No. 4, pp. 270–280, 2005.

[13] Masaharu Yoshioka and Makoto 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*, pp. 9–16, 2004. http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/WEB/NTCIR4WN-WEB-YoshiokaM.pdf.