2004－ＦＩ－74　(12)
2004－ＤＤ－43　(12)
2004／3／26

# 適合的汎化に基づく情報検索システムの研究(第2報)
## – 検索語の網羅性に注目した検索インタフェースの作成 –

○吉岡 真治　原口 誠
北海道大学大学院工学研究科

**概要:**　情報検索システムを利用する検索者にとって、適切なキーワードを選択することは必ずしも容易なことではない。本研究では、検索者にも理解しやすい概念階層の汎化という考え方を利用して、ユーザの検索意図を明示化すると共に、精度落ちを抑えた情報検索システムを提案している。本報では、汎化概念の関連文書における網羅性に注目し、検索に役に立つと考えられる汎化概念を明示化するインタフェースの作成について報告する。また、この網羅性に注目することにより、検索意図をより明確に表現するBoolean検索式の構築の支援とWeb検索への応用についても述べる。

# Research on Information Retrieval System based on Adaptive Generalization (2nd Report)
## - Construction of IR User Interface that focuses on Generalized Concept with High Coverage -

Masaharu Yoshioka　Makoto Haraguchi
Graduate School of Engineering, Hokkaido University

**Abstract:**　It is not easy for a user of Information Retrieval (IR) system to select appropriate keywords. In this research, we proposed a new IR system that uses adaptive generalization of keywords. When the system can selects appropriate generalization by estimating user's intent, the system can generate good keywords that have high readability and good retrieval performance. In this report, we proposed new IR user interface that displaies generalized concepts by using relevant documents information. In addition, we propose a method to construct Boolean retrieval formula by using these generalized concepts and apply this method for Web information retrieval.

## 1　緒言

現在の情報検索システムにおいては、検索語の入力による検索が主流であるが、一般の検索者にとって自分が思っている検索意図に基づいて適切な検索語を選択する事は必ずしも容易ではない。これに対し、ユーザモデルの利用や、関連文書中の語を検索キーワードに加える事により、検索者の検索意図の推定を行っているシステムなどが提案されている。しかし、これらのシステムは、検索性能の向上という成果をあげているが、推定された検索意図の表現が検索者にとって理解困難なものが多く、本当に検索者の検索意図とマッチしているのかを検索者が確認するのが困難であるという問題がある。

検索者にとって理解しやすい検索語の選択支援の手法として、シソーラスを用いた支援がある。しかし、用途に応じた適切なシソーラスを構築することは手間がかかるという問題がある。また、

一般的な目的で構築されたシソーラスを単純に用いた検索拡張では、検索精度が向上しないことが WordNet[1] を使った実験により確認されている [2]。よって、本研究では、入力された検索語と数個の関連文書を用いて、検索語や関連文書中に存在する語の汎化レベルを推定することにより、検索者に理解しやすい検索拡張を行う適合的汎化に基づく情報検索システムを提案している。

本報では、汎化概念の関連文書における網羅性に注目し、検索に役に立つと考えられる汎化概念を明示化するインタフェースの作成について報告する。また、この網羅性に注目することにより、検索意図をより明確に表現する Boolean 検索式の構築の支援と Web 検索への応用についても述べる。

## 2　適合的汎化に基づく情報検索システム

### 2.1　概念階層に基づく検索語の汎化

一般的な検索者は、検索語が持つ適合文書の分別能力などについて深く気にせずに、検索語の選定を行っている場合がある。例えば、「ビデオ」という概念に関連して図 1 の様な概念階層を考えたときの、次の 3 つの事例における「ビデオ」というキーワードが持つ意味について考える。

1. 映画を見たいと思って、「レンタル」「ビデオ」という検索語を利用する人にとって、「ビデオ」というのは代表的な手段であって、「DVD」などを含む「映像機器」でも良いと考えている。

2. ビデオの構造を知りたいと思って、「ビデオ」「構造」という検索語を利用する人にとっては、ビデオ一般（VHS ビデオ、8mm ビデオなど）ならどれでも良いと考えている。

3. VHS ビデオのデッキを買いたいと思って、「ビデオ」「デッキ」という検索語を利用する人（ビデオといえば VHS だと思っている）にとって、「VHS ビデオ」が良い検索語である。

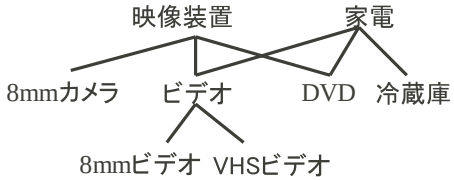


図 1: ビデオに関する概念階層

これらの事例からわかるように、検索者は、検索意図を表現するのに適切な抽象度の概念を必ずしも用いない場合がある。そのため、検索意図に応じた適切な抽象度の概念を選択し、検索語に用いると、検索者にも理解しやすく効率的な検索語になると考えられる。本研究では、この様な電子化辞書やシソーラスに記述されている概念階層構造を利用し、検索意図に応じて検索語を汎化する方法を「適合的汎化」と呼ぶ。

一方、既存の電子化辞書やシソーラスに記述される概念階層は、必ずしも、ユーザの特定の検索意図を表現するために十分な概念の詳細度を持っていない場合がある。例えば、先のビデオで映画を見たいと考え、ビデオを検索語に利用する事例を考えると、汎化概念である「映像装置」に含まれる概念の内、DVD は検索意図にあうが、8mm カメラは検索意図と必ずしも一致しないと考えられる。

よって、ユーザが持つ細かな検索意図を適切に表現するために、検索目的に応じた適切な抽象概念を設定し、概念階層を再構築することが必要である。この様な中間階層の概念カテゴリーは、

特定の個人の特定の目的に応じたものであり、本研究では、この様な概念階層の構築を目的指向の概念階層の構築と呼ぶ。本研究では、関連文書の情報を用いて、この目的指向の概念階層の構築を行う。

## 2.2 検索語が持つ適合性判定への寄与度

本研究では、検索語あるいは概念の存在が適合文書の判別において、どの程度貢献するかによって、その語あるいは概念が役に立つか立たないかを判断する。そのため、本研究では、語の存在と適合文書の間の相互情報量に基づいた以下の指標 $G'(w)$ により、語あるいは概念の有効性を判断する。

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

ただし、$w$ は各単語を表す変数 $r$ は適合文書群を表す確率変数

この指標は、主に、$p(w|r)$ と $p(w)$ の比に注目しているため、次のような性質を持つ。以下では説明のため、検索語として $a$、汎化語の概念として $A$ を考える。

1. 汎化を行う事は、対応する語の数が増えるため、$p(w|A) \geq p(w|a)$ と $p(A) \geq p(a)$ の関係が成り立つ。
2. 汎化を行うことにより、より多くの関連文書をカバーする文書が増加する場合には、$p(w|A)$ の増加分が大きいことになり、$G(A)$ が大きくなる可能性が高くなる。
3. 汎化を行っても、関連する文書が増えない場合には、$p(A)$ の増加分が大きくなり、$G(A)$ が減少する。

この 3 番目の性質により、非適合文書を明示的に与えなくても過剰汎化を防ぐことができる指標になっている。

## 2.3 適合的汎化に基づく情報検索システム

上記の考え方に基づいた適合的汎化に基づく情報検索システムのプロトタイプを作成した。このシステムは、通信総研で作成されている BM25[3] を利用した情報検索のパッケージ [4] (以降では、ベースラインシステムと呼ぶ) をベースとして作成した。また、概念階層を与える電子辞書としては、EDR[5] を利用する。

本システムを、情報検索システム評価用のテストコレクションである NTCIR-1 テストコレクション [6] に適用した。その結果、本システムでは、初期検索の結果に検索性能が影響されやすく、初期検索のランク上位文書を関連文書として利用するオートマチックフィードバックを利用する場合には、性能の向上が見られなかったが、テストコレクションに記載されている正解文書を利用する場合には、少ない検索語の拡張で、関連文書全てに含まれる語を検索拡張に利用するベースラインシステムと同等の検索性能を得ることが確認できた。

# 3　検索語の網羅性に注目したBoolean型検索への応用

本研究で提案している適合的汎化の手法では、単純に、検索語や文書中に存在する語を用いるのではなく、関連文書に広く特徴的に現れる抽象度の高い概念がある場合には、その抽象度の高い概念で検索をすることにより、より適切な検索式が作れるという考えに基づいたものである。

この抽象化と目的指向の概念の汎化は、検索語として役に立つ特徴的な語に対し、その検索語と同じ概念を表しながら補完的な役割を果たす検索語を見つけることにより、検索式を拡張する方法と考えることができる。

この性質を用いることにより、検索語や関連文書の情報に基づいた、より適切なBooleanの検索式を作ることが可能であると考えられる。また、この適合的汎化による検索拡張の結果をBoolean式として提示することにより、単なる検索拡張による検索語の提示ではなく、検索語の選択理由についての理解が深まり、よりユーザにとって理解しやすい表現形式になると考えられる。

以下では、Boolean式に利用するという観点からの検索語の汎化の方法論と、そのBoolean式の表現手法について述べる。

## 3.1　検索語の汎化

従来の適合的汎化に基づく情報検索システムでは、検索語がどれだけ特徴的に正解文書に含まれるかという相互情報量に注目していたため、抽象化した語はもとの語に比べ、関連文書に含まれる可能性が高くはなっているが、必ずしも、全ての関連文書を網羅するものではない。そのため、Boolean式として利用するためには、関連文書群に対する網羅性という観点から検索語の指標をとらえ直す必要がある。

よって、現在の汎化操作に関する基準を相互情報量に基づく指標である$G'(w)$に加え、関連文書に対する網羅性を考慮した次の2つの基準を導入する。

- 一定の割合以上の関連文書に含まれない検索語や抽象概念は汎化の対象とする。
- ユーザの視認性を考慮して、一定数以上の概念の抽象化は行わない。また、$G'(w)$が一定の値より小さい場合には、抽象化を行わない。

この汎化操作により、関連文書に対する網羅性の高い概念への汎化が行われることになる。その結果得られた抽象概念が全ての関連文書に含まれる場合には、その抽象概念を、Boolean式のandとして設定することにより、初期にユーザにより与えられたキーワードを補完したBoolean式が作成可能になる。ただし、ここで、作成する抽象概念は、検索インデックスとして存在しないので、実際の、検索式では、抽象概念に対応する検索語全体をorで結合したものとして表現される。

また、このような汎化を行わない検索語についても、全ての関連文書に含まれる語が存在する。このような語を全てBoolean式のandを構成する要素として利用する事も可能であるが、今回のシステムでは、初期検索式に含まれていた語のみをBoolean式に利用することとした。

この汎化によるBoolean式の作成について、「ローマの休日を見たい」という検索要求に基づいた例を用いて考える。この時の、ビデオという言葉に関する概念階層ならびに、各々の概念に対応する$G'(w)$の値、関連文書については、図2に示すようなものであったとする。また、抽象概念は、全体の70%以上の文書をカバーする必要があり、最小の$G'(w)$の値は1.0と設定した。

この時、ビデオという言葉は、一番、$G'(w)$が高い言葉であるが、2つの関連文書の内の1つに存在するのみであり、網羅性の基準を満たしていない。よって、更なる汎化の対象となり、ビデオとDVDの汎化概念である映像機器がこの検索要求に適切な汎化概念として選択される。また、検索語にある「ローマ」「休日」という語は、全ての文書に存在するため、Boolean式のandとして利用する。

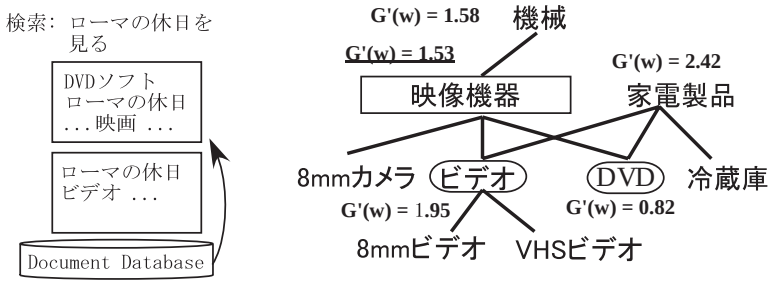


図 2: 適合的汎化による Boolean 式の作成

よって、この検索の場合に作成される Boolean 式は (ローマ and 休日 and (ビデオ or DVD)) となる。

## 3.2 検索インタフェースの作成

ユーザが作った検索式に対し、ビジュアルな表現を通して検索式の修正を行っていくシステムとしては、検索文書に特徴的に現れる語のみを表現する DualNavi[7] やシソーラスの概念階層をインタラクティブに表示させながら検索式を作っていく研究 [8] などがある。しかし、DualNavi において表示される検索語間の関係は主に共起関係に基づくものであり、本研究で取り扱おうとしているような補完関係にあるような語を見つけ出すという目的には、あまり適していないと考えられる。また、シソーラスの概念階層を見せる方法では、どの程度の抽象度で検索式を表現するのが適切なのかといったことや、本研究で提案しているような目的指向の概念階層の修正と言った操作が含まれておらず、ユーザへの負荷が高いと考えられる。

よって、本研究では、検索者が関連文書を選んでいくと、その関連文書に応じた適合的汎化を行い、その結果できあがる検索式をユーザに提示するシステムを提案する。つまり、このシステムでは、ユーザは、関連文書を選んでいくことにより、一般的なシソーラスに基づき生成された抽象概念を見ることができ、その概念の修正などを通して、検索式の修正を行う。

作成した検索インタフェースのハードコピーを図 3 に示す。この検索インタフェースは左側が検索語の入力や修正を行う部分であり、右側が文書のリストを見るための部分となっている。

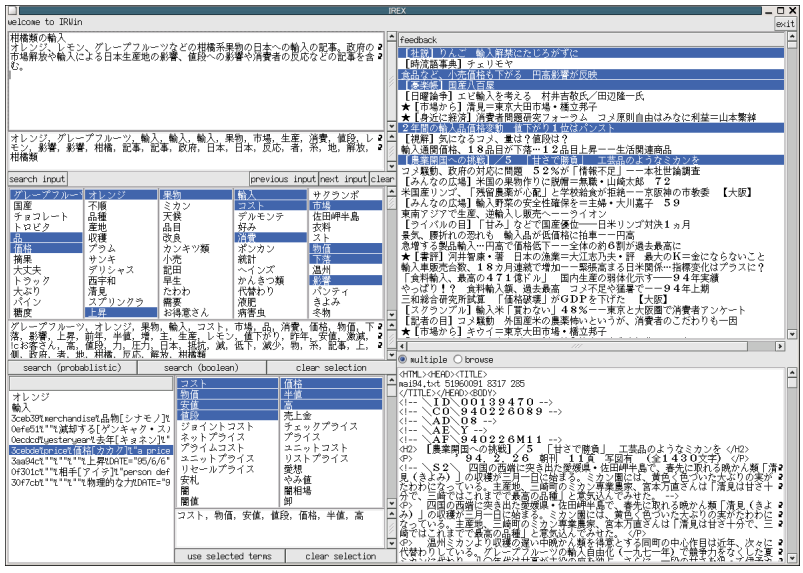

図 3: 検索インターフェース

左の真ん中にある語のリストは、上位 10 件もしくは選択した関連文書に特徴的に現れる語のリストであり、左下のリストが、汎化概念、並びに、Boolean として利用可能な検索語を表示している。

ユーザは右側のリストから関連文書を選択することにより、左側のリストに表示される語のリストならびに、生成された抽象概念が変化する。ユーザは、実際に、抽象概念に対応する語のリストを見ることにより、適切な語の追加や削除が可能となっている。また、ユーザは検索意図が明確になったかどうかを考慮しながら、Boolean による検索と、確率モデルによる Boolean による制約をかけない検索を切り替えて利用することができる。

さらに、作成した検索式の Boolean 式で表される部分の検索式を用いて、AltaVista などの検索エンジンによる Web 検索を利用することができる。このようにして作成された Boolean 式をユーザが最初から作成するのは困難であると考えられるが、本システムで提案している適合的汎化の考え方に基づいて、ローカルデータベースとやり取りを行うことにより、容易にこの様な検索式が作成できる。

## 4 結言

本報では、汎化概念の関連文書における網羅性に注目し、検索に役に立つと考えられる汎化概念を明示化するインタフェースを提案した。このインタフェースでは、網羅性に注目することにより、検索意図をより明確に表現する Boolean 検索式の構築を支援することができ、Web 検索への応用が実現できた。

今後の展望としては、本システムの有効性を検討するためのユーザ実験などが必要であると考えている。さらに、Web 検索の結果、得られた文書などの情報とローカルのデータベースに蓄積されている情報をうまく組み合わせて、更なる検索式の洗練化を支援する方法や、Boolean 式を作成するための基準についての検討を行っていきたいと考えている。

## 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させて頂きました。また、毎日新聞 1994 年版、1995 年版 CD-ROM と IREX 実行委員会の作成したデータを利用させて頂きました。本研究の一部は、文部科学省科学研究費補助金 (特定領域 (2) 課題番号 15017202) によって実施された。

## 参考文献

[1] George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[2] 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.

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

[4] 内山将夫, 井佐原均. 情報検索パッケージの実装. 情報処理学会情報学基礎研究会 ,2001-FI-63, pp. 57–64, 2001.

[5] 日本電子化辞書研究所. EDR 電子化辞書 (第 2 版) 仕様説明書, TR2-006(改), 2001.

[6] 神門典子. 情報検索システムの評価プロジェクト：NTCIR ワークショップ. 情報処理, Vol. 41, No. 6, pp. 689–697, 2000.

[7] Akihiko Takano, Yoshiki Niwa, Shingo Nishioka, Toru Hisamitsu, Makoto Iwayama, and Osamu Imaichi. Associative information access using dualnavi. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pp. 771–772, 2001.

[8] Hideo Joho, Claire Coverson, Mark Sanderson, and Micheline Hancock-Beaulieu. Hierarchical presentation of expansion terms. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 454, 2002.