データベースシステム 127－20
情 報 学 基 礎 67－20
（２００２．５．２２）

適合的汎化に基づく情報検索システムの研究(第1報)
– 検索語が持つ適合性判定への寄与度の利用 –

○吉岡 真治　原口 誠　大久保 好章
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。この問題に対し、検索者の検索意図を推定し、検索者の補完を行う検索システムが提案されている。しかし、補完した結果は複雑であることが多く、検索者が補完結果自体を評価することが困難である。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する。本報では、概念階層の汎化のレベルを決定するために、検索者により入力された検索語が持つ適合性判定への寄与度を考える。また、本システムの性能を向上させるためには、概念階層自体が目的に応じて再構成されることが必要であることについて述べ、目的指向の概念階層の修正の方法を提案する。

Research on Information Retrieval System based on Adaptive Generalization (1st Report)
- Generalization of Keywords based on the Analysis on Contribution of Keyword for Relevance Judgement -

Masaharu Yoshioka　Makoto Haraguchi　Yoshiaki Ohkubo
Graduate School of Engineering, Hokkaido University

Abstract: It is not easy for a user of Information Retrieval (IR) system to select appropriate keywords. Therefore, many IR systems have capability to modify keywords by estimating user's intent. However, since modified keywords are usually represented as complicated form, it is difficult to judge the appropriateness of them. 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 to use the contribution value of each keyword for relevance judgement to select appropriate generalization. In addition, we confirm general concept structure stored in a thesaurus is not sufficient for representing particular user's intent. Therefore, we proposed a purpose oriented method to modify concept structure.

1 緒言

近年のIT技術の発展に伴い、大量の文書情報が利用可能になっており、様々な情報検索システムが実用化されている。しかし、検索語の入力を中心とした現在の情報検索システムにおいて、一

般の検索者は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が困難である。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が提案されている。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意図の表現方法としては、シソーラスを使った検索拡張がある。しかし、単純なシソーラスによる検索拡張では、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、入力された検索語から検索意図を推定し、その意図に応じた検索語を汎化するかしないか、また汎化を行う場合に抽象度をどれだけ上げるかを決定する事により、検索者に理解しやすく、検索性能を維持したシステムの提案を行う。このシステムで行うような、目的に応じて概念階層の汎化レベルを決定する操作を**適合的汎化**と呼ぶ。

本報では、検索要求を表現する各検索語に対し、適合性判定への寄与度を示す指標を適合的汎化のための指標として、プロトタイプシステムを作成したので、それについて述べる。また、この検索実験の結果に基づき、一般的なシソーラスが持つ概念階層の問題について述べる。この問題を解決するために、目的思考の概念階層の修正という考え方を提案し、システムの改良を行ったので、それについても述べる。

2　情報検索における検索語の特徴

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と考えられる。ここでは、検索者のこの様な行動と、検索意図に応じて適切だと考えられる検索語の関係を考えることにより、検索意図の表現について考える。

図 1 の様な概念階層を考えたときに、次の 3 つの事例において「ビデオ」というキーワードが検索語として用いられたとして、その検索語が持つ意味について検討する。

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。

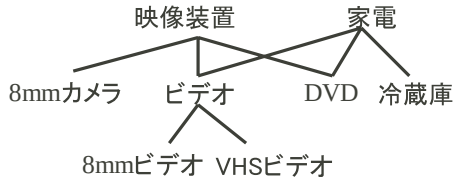

図 1: ビデオに関する概念階層

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。

このキーワードの汎化によって検索者の意図を外化させる方法の有効性については、文書のクラ

スタリングにおけるクラスタの意味付けを表すために EDR[3] の概念階層を用いた Fish Eye マッチングシステム [4] などで確認されている。本研究では、情報検索のための適切な汎化についてさらに考察を深めることにより、検索者にも理解しやすく精度の高い検索を実現するシステムを提案する。

3 検索意図に基づく検索語の適切な汎化

3.1 概念階層に基づく検索語の汎化

本研究では、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適合的に汎化する情報検索システムを提案する。ここで、適切な抽象度の汎化とは、検索キーワードが持つ正解判定の分別能力に関する情報を多く保存する汎化の事である。

よって、全ての検索語に対して、シソーラスによる検索拡張を行うのではなく、適切な検索語を選択し、汎化のレベルを設定するという手順が必要となる。本研究では、検索語が持つ様々な特徴量から、汎化レベルの判断を決定する関数を作成し、この判断を行う。

3.2 検索語が持つ適合性判定への寄与度

本報では、検索語の文書中の分布に注目して汎化レベルの判断を決定する関数を作成する。役に立つ検索語とは、その後の存在が適合文書の判別に役に立つ語であると考える。

この様な語が持つトピックの分類能力についての研究として、話題分類のための相互情報量に基づくキーワード選択 [5] の研究がある。この研究では、文書中のある単語が単語群 W 中のどの単語かが分かるかによって、T の同定に関して得られる情報量に相当する相互情報量 $I(T;W)$ を利用する。この時、式 (1) の [] 内を、「話題 T の決定のために単語 w が持つ寄与度」と定義し、$G(w)$ と記す。この $G(w)$ が大きな単語ほど話題同定により多くの情報を持っていると考える。

$$\begin{aligned} I(T;W) &= I(W;T) = H(W) - H(W|T) \\ &= \sum_{w \in W} [-p(w) log_2 p(w) + \sum_{t \in T} p(t)p(w|t) log_2 p(w|t)] \end{aligned} \tag{1}$$

情報検索においては、同定すべきトピックとは、適合文書 (r) と非適合文書 (全体文書から適合文書を取り除いたもの：$\bar{r}$) であると考えられるので、$T = \{r, \bar{r}\}$ と考えられる。これに基づいて、先ほどの $G(W)$ を式変形する。また、適合文書数に相当する r のサイズは、全文書に比べサイズが小さいと考え、非適合文書における語の分布は、全体の語の分布にほぼ等しい ($p(w|\bar{r}) \approx p(w)$) と考えると、式 (3) が得られる。

$$\begin{aligned} G(w) &= -p(w) log_2 p(w) + \sum_{t \in T} p(t)p(w|t) log_2 \\ &= -\{p(r)p(w|r) + p(\bar{r})p(w|\bar{r}) log_2 p(w) + p(r)p(w|r) log_2 p(w|r) + p(\bar{r})p(w|\bar{r}) log_2 p(w|\bar{r}) \\ &= p(r)p(w|r)\{log_2 p(w|r) - log_2 p(w)\} + p(\bar{r})p(w|\bar{r})\{log_2 p(w|\bar{r}) - log_2 p(w)\} \qquad (2) \\ &\approx p(r)p(w|r) log_2 \frac{p(w|r)}{p(w)} \qquad (3) \end{aligned}$$

ここで、$p(r)$ はトピック毎の定数である。よって、本研究では、以下の $G'(w)$ をあるトピックが与えられたときにその検索語の有効性を検討するための指標として利用する。つまり、汎化を行うことにより検索語の有効性が高くなる ($G'(w)$ が増加する) 汎化を適合的な汎化と考える。

$$G'(w) = p(w|r) log_2 \frac{p(w|r)}{p(w)} \tag{4}$$

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。

4　検索語の指標を用いた適合的汎化に基づく情報検索システム

前節で提案した検索語の指標を用いて汎化を行う情報検索システムを作成し、学術文献を元に作成された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] に適用して、その有効性を検証する。本研究では、検索システムを通信総研で作成されている BM25[7] を利用した情報検索のパッケージ [8] (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する。このベースラインシステムは、NTCIR-1 テストコレクションに適用した場合に上位のシステムと同等の性能を有しており、検索性能の比較のためのベースラインとしても利用する。

また、概念階層を与える電子辞書としては、EDR[3] を利用する。

4.1　特徴量を計算するためのインデックス構造と検索手順

今回提案する特徴量を計算するためには、式 (4) に示すように文書中における語や抽象概念の存在確率を計算する必要がある。本研究では、ベースラインシステムにおけるインデクシング作業を拡張した以下の手順によりインデックスを作成する。

1. 形態素解析ツール茶筅を利用し、文章の形態素解析を行い、ベースラインシステムと同様に、名詞を中心とした語をインデックス語として選択する。
2. インデックス語に対し、EDR の概念辞書を利用することにより、対応する概念とその概念の上位階層にあたる概念群をインデックスとする。このとき、頻出する概念 (例えば、最上位階層にあたる「概念」など) はインデックスから削除する。
3. 1. で選択した語と 2. で選択した概念群を全てインデックス語として登録する。

このインデックスを用いることにより、各語や概念が文書中にどのくらいの頻度で現れるかが計算可能となる。

本システムは、次の手順により、汎化のための指標を利用し、情報検索を行う。

1. 検索文を形態素解析し、インデックス作成時と同じフィルタリングをする事により、初期検索語のリストを得る。
2. 初期検索を行い、関連文書の候補を得る。
3. 各検索語について、汎化可能な概念の候補を選び出し、各々の $G'(w)$ を計算する。本システムでは 2 で得た関連文書の候補から、上位 5 件を関連文書と考えて、関連文書中の語の分布と全体の文書中の語の分布を計算する。

4. 検索語の $G'(w)$ と汎化概念の $G'(w)$ を比較し、汎化概念の $G'(w)$ の少なくとも一つが検索語の $G'(w)$ より大きい場合は、$G'(w)$ の最大値を持つ概念に汎化する。汎化が行われた場合は、汎化された概念を検索語とみなして 3 からのステップを繰り返す。
5. 全ての検索語について汎化のプロセスを繰り返し、汎化を行った概念については、検索語を汎化概念に置き換えて検索式とする。

4.2 検索システムによる実験と評価

本システムを評価するために、NTCIR-1 テストコレクションへの適用を行った。本システムは初期検索の精度に大きく影響を受けると考えられるため、初期検索として、ベースラインシステムがオートマチックフィードバックに利用しているのと同じ初期検索の文書を使う場合と、類書検索的な要素が強くなるが、あらかじめ正解文書を 5 件与える場合について実験を行った。

表 1 にベースラインシステムの結果、初期検索にベースラインシステムを利用した結果、正解文書を利用した結果と比較対象となる検索式拡張なし、ベースラインシステムの結果を示す。long と short は検索語を抽出するために使う検索文の違いで、short は検索要求のタイトルのみから検索語群を作成する検索で、long は検索要求の説明文や、関連するキーワードが含まれた文から検索語群を作る検索である。また、今回の判定では、部分的適合以上を正解として評価を行った。

表 1: 検索結果 (Average Precision)

システムのタイプ	long	short
関連文書としてオートマチックフィードバックで利用する文書群を利用	0.479	0.3505
関連文書として正解文書を利用	0.483	0.3608
検索式拡張なし	0.468	0.3454
ベースラインシステム	0.488	0.4097

また、表 2 に示すように、初期検索にベースラインシステムを利用した場合は検索語の汎化もあまり行われず、検索性能も悪い。正解文書を与えた場合には、語の汎化を行う数も増加し、検索性能も向上している。本システムが初期に与える正解文書の品質に影響を受ける事を示している。よって、現時点のシステムの性能では、オートマチックに適切な汎化を行うのは不十分であり、ユーザとのインタラクションあるいは、類書検索的な形での利用が適切であると考えられる。

表 2: 行われた汎化の数

システムのタイプ	long	short
関連文書としてオートマチックフィードバックで利用する文書群を利用	9	120
関連文書として正解文書を利用	29	140

また、本システムは、ユーザには理解しやすい検索式にはなっていると考えられるが、ベースラインシステムに比較して検索性能が悪い。その原因を調べるために検索課題毎に、システムの検索性能を比較した結果、検索性能を落としている課題では、2 段階以上の汎化が行われていた。この様な多段階の汎化が行われると、一つの概念に対応する語の数が多数存在し、必要のない語が多く検索に利用されている事が解った。

5　目的指向の辞書の概念階層の修正

5.1　概念階層の修正を伴う情報検索システム

前節で述べたシステムの問題は、シソーラスとして用いたEDRにおいて、概念階層は一般的な目的で作成されており、詳細な検索意図を表すに十分でないことが原因で起きている。

例えば、2節で述べたビデオの事例 (ビデオで映画を見たいと考え、ビデオを検索語に利用) 考えたときに、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。そのため、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。

本研究では、汎化した概念であっても、関連文書に含まない概念は検索に関係しない概念と考え、汎化した概念に対応する語の中で関連文書に含まれる語のみで暫定的な中間概念を設定し、検索に利用する。

具体的には、前節で述べたシステムの手順4を次の手順4'に置き換える。

4'. 全ての検索語について汎化のプロセスを繰り返し、汎化を行った概念については、関連文書中に存在する語のみを列挙し、検索語として利用する。

5.2　検索システムによる実験と評価

前節のシステムの比較のため、初期検索としては、正解文書を利用することとした。表3にベースラインシステムの結果、正解文書を利用した結果、概念階層の修正を含むシステムの結果を示す。この表から、概念階層の修正を行うことにより、検索性能の向上が確認できる。また、表4に、今回のシステムによる汎化を通じて検索式に追加された語の数を示す。

表 3: 概念階層の修正を行うシステムの検索結果 (Average Precision)

システムのタイプ	long	short
概念階層を修正	0.5093	0.4029

表 4: 汎化の対象になった語の数と追加された検索語の数

	long	short
汎化の対象になった語	29	140
追加された語	38	208

今回のシステムは正解文書の情報を一部利用しているため、そのままベースラインシステムと比較するのは不適切である。よって、ベースラインシステムに初期検索結果として正解文書を与えた場合、正解文書の情報により検索式を作成し、利用した正解文書以外の文書を検索した場合の実験を行った。表5に示すように、ベースラインシステムに正解文書を与えた場合には、大幅に検索性能が向上する。

表 5: ベースラインシステムに正解文書を与えた場合の検索結果 (Average Precision)

システムのタイプ	long	short
ベースラインシステム＋正解文書	0.6258	0.5903

次に、オーバーフィットの可能性を検討するために、システムに与える正解文書を全体の文書集

合から取り除いて作成した文書集合を作成し、検索実験を行った。各々のシステムに対し、以前の実験で作成した検索式を新しく作成した文書集合に適用した。

ベースラインシステムに正解文書を与えた場合には大きく検索性能を落とし、long では今回提案した概念階層を修正するシステムよりも性能を落としている。この事は、ベースラインシステムが作成した検索式は、オーバーフィットしているのに対し、今回提案しているシステムが作る検索式は、より一般的な検索意図を反映した検索式になっていることを示していると考えられる。short の場合には、本システムでは検索拡張に使える語が少ないのに対し、ベースラインシステムでは、long とほぼ同じ検索拡張が行える事が大きな違いであると考えられる。

表 6: 検索結果 (Average Precision)

システムのタイプ	long	short
ベースラインシステム＋正解文書	0.4359	0.3867
概念階層を修正	0.4463	0.3248

5.3 目的指向の概念階層の修正の効果

今回の実験結果より、シソーラスの概念階層を特定の検索に応じて修正するという事の、情報検索における有用性を確認することができた。

シソーラスの概念階層の構成が検索結果に影響を与えることに注目して、シソーラスの内容を修正して利用する方法が提案されている。[9] では、複数のシソーラスの共通部分を用いることにより、より一般性の高いシソーラスを利用する方法を提案しており、シソーラスの修正が検索性能の向上に役立つことを示している。

本研究で提案している目的指向の概念階層の修正方法は、情報検索以外の分野にも適用可能であり、データマイニングにおいて、精度落ちが少なく人間が理解しやすいルールを作成する場合などにも応用 [10] を行っている。

6 結言

本報では、概念の適切な抽象化のための指標として、キーワードと正解文書の相互情報量に基づく関連性の指標を利用し、情報検索システムを構築した。また、シソーラスが持つ概念階層を目的に応じて変更することが、検索性能の向上に役立つことを示した。今回提案したシステムは、5 件の正解文書を与えることにより、検索語の汎化を行い、検索者の意図が理解しやすい検索式が作成可能である。また、この検索式による検索は、他の検索システムと比較して十分な検索精度を持っていることを確認した。

今後の展望としては、他のテストコレクションに適用することによる分野依存性の検証や、検索語を適切に汎化するための異なる指標の検討などを進め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。本研究の一部は、文部科学省科学研究費補助金 (特定領域 (C)(2) 課題番号 13224401) 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] 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.

[4] 高間康史, 石塚満. Fish Eye マッチング：概念体系を利用した視点抽出に基づく文書整理支援機能. 人工知能学会誌, Vol. 14, No. 1, pp. 93–101, 1999.

[5] 恒川俊克, 山下洋一, 溝口理一郎. キーワードスポッテイングに基づくニュース音声の話題分類. 情報処理学会音声言語情報処理研究会 ,98-SLP-20, pp. 61–68, 1998.

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

[8] 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 ー 63, pp. 57–64, 2001.

[9] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. Improving information retrieval system performance by combining different text-mining techniques. *Intelligent Data Analysis*, Vol. 4, pp. 489–511, 2000.

[10] Yoshiaki Okubo, Yoshimitsu Kudoh, and Makoto Haraguchi. Constructing appropriate data abstractions for mining classification knowledge. In *Proceedings of the 14th International Conference on Applications of Prolog - INAP01*, pp. 275–284, 2001.