電子化辞書に基づき検索意図を現わす概念を作成し利用する情報検索システムの研究

北海道大学大学院 工学研究科
○吉岡 真治　原口 誠
e-mail:{yoshioka,makoto}@db-ei.eng.hokudai.ac.jp

1 緒言

現在、計算機により、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、我々は、検索語と関連文書群から検索者の意図を現わすのに適切な概念の抽象度のレベルを推定し、検索意図に応じた電子化辞書の概念階層の修正と、概念の抽象度に応じた検索要求の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 [1]。しかし、このシステムでは、検索要求に含まれる語に対してのみ、概念の汎化を行ったため、検索要求を短く表現している場合などにおいて、あまり、性能の向上が図られないという問題があった。

そこで、本論文では、関連文書群の一部を検索者の意図を表現したものと捉え、そこに特徴的に現れる概念を選択することにより、検索者の関連文書フィードバックによる検索拡張の方法を提案する。また、このプロトタイプシステムを作成し、NTCIR-1[2] と IREX[3] の2つのテストコレクションに適用し、その有効性について議論する。

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とその指標

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。検索者のこのような行動と、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。

例えば、同じ「ビデオ」というキーワードを検索語として用いたとしても、「ビデオ」や「DVD」などを含む「映像機器」の代表として用いたり、ビデオ一般 (「VHS ビデオ」、「8mm ビデオ」など) を指していたり、ビデオといえば VHS だと思っている人にとっては、「VHS ビデオ」を指していたりする。

つまり、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、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と考え、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。

このプロトタイプでは、検索語が持つトピックの分類能力に注目した、相互情報量に基づく以下の指標を利用した。

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

ただし、$p(w)$ は全文書中である単語 w が存在する確率

$p(w|r)$ は文書中のある単語 w が適合文書中に存在する確率

この指標は、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、汎化後の概念として 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、$p(A|r) \geq p(a|r)$ 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、$p(A|r)$ 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。

この性質により、過剰な汎化を行わずに、多くの文書に存在する概念が選択される。

2.2 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

プロトタイプシステムは、通信総研で作成されている BM25[4] を利用した情報検索のパッケー

ジ [5] (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。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[6] を利用した。

本システムでは、文書群に形態素解析を行い、各々の語に対応する EDR の概念 ID と組み合わせたインデックスを作成する。そして、検索時には、最初に検索文に含まれる全ての初期検索語群を作成する。この初期検索語群と関連文書群の組合わせから全ての語に対し、先ほど述べた検索語の指標を計算し、指標が増加する場合には、検索語を汎化した概念 ID で置き換えるという形で適合的汎化を実現した。

また、EDR の概念階層が、一般的な目的で作成されており、2 階層から 3 階層以上の汎化を行った場合に、ユーザの詳細な検索意図を表すのに適切な粒度の概念階層が無い場合が見受けられた。そのため、汎化した概念の内、関連文書群に含まれる語のみを含む中間階層に相当する概念を設定する方法を提案した。

本システムでは、関連文書の質が、検索意図の汎化の質に大きく影響する。そのため、初期検索語群で検索した結果の上位を利用する方法で検索語の汎化を行った場合に、最初の検索語に大きく依存した結果が出るため、適切な汎化が行われにくいという問題があった。そのため、検索者の意図をより明確に推定し、検索意図に応じた汎化を実現するためには、実際の正解文書群を与える必要があった。本システムでは、検索者の意図の推定という観点では効果があったが、検索性能の観点からは、より一層の性能向上が望まれている。

3 異なる検索拡張の手法を用いた情報検索システムの改良

3.1 関連文書に共通する概念の利用

検索者の検索意図を補完する検索拡張の研究が多くなされており、代表的なものとして、(仮想的な) 関連文書中に含まれる語を利用する関連文書フィードバックの方法やシソーラスを用いる方法がある。

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[7] を使った実験により確認されている [8]。本研究で提案している手法では、単純にシソーラス中にある概念をそのまま利用するのではなく、実際の関連文書に即して概念に属する語を選択的に生成していることが、性能の向上につながっていると考えられる。

一方で、関連文書フィードバックによる検索拡張は、関連文書中には検索者が直接指定していないキーワードであっても、検索に役立つキーワードがあるはずだという考えに基づいた方法であり、その有効性が確認されている。この観点から、前節で述べた適合的汎化に基づく情報検索システムを考えると、検索語と共通の検索に役立つ抽象概念を持つキーワードのみ追加する検索拡張を行っているとも説明できる。

実際に、ベースラインシステムにおいて関連文書フィードバックをした結果として拡張された検索語について分析すると、検索要求に含まれている語と語義的に関係ない語が検索において大きな役割を果たしていることが分かる。

一方で、ベースラインシステムでは、単純に関連文書に含まれる全ての語について検索要求に追加しているため、検索意図とは関係ないが偶然に関連文書に含まれている語を検索語に追加してしまうことにより、検索式が与えた関連文書にオーバーフィッティングしてしまう可能性がある。

よって、これらの検索語を排除するために、本研究では、関連文書に特徴的に共通して現れる概念に対応する語のみを検索語に追加する方法を提案する。具体的には、関連文書に含まれる語について、検索要求の語と同様に汎化の可能性を検討し、汎化した方が良い場合は、それらを検索語に追加することとする。

また、今回、用いている指標では、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必要がある。よって、関連文書に特徴的な語であっても、一つの表記でしか出現しない語に対しては、検索語への追加がなされない。また、電子化辞書に含まれない語 (特に、英語の略語など) で検索に役立つ語の追加も行われない。

これらの問題を踏まえ、汎化の対象とならない概念でも、式 (1) の指標が高い語は、検索語として追加する事とした。

最後に、本システムにおける検索拡張の方法についてまとめると以下のようになる。

1. 関連文書の選択
   初期検索の上位 5 件もしくは、ユーザが与えた文書を関連文書として利用する。
2. 検索語の汎化レベルの検討
   検索要求と関連文書に含まれている全ての語について、式 (1) の指標を用いて、概念の汎化を行うかどうかを検討する。概念の汎化を行う場合には、汎化した概念に対応し、関連文書に存在する語を検索語に追加する。
3. 関連文書に特徴的な語の追加
   関連文書に存在する全ての語について、式 (1) の指標を計算し、その指標が大きなもの、

表 1: 検索結果 IREX(Average Precision)

	long			short		
	normal	rel	rest	normal	rel	rest
BL	0.5206	0.5943	0.5533	0.4538	0.5424	0.4963
OLD	0.4759	0.5165	0.4983	0.4211	0.4677	0.4533
GEN	0.5010	0.5411	0.5081	0.4340	0.4971	0.4638
G+W(10)	0.5155	0.5820	0.5362	0.4764	0.5419	0.5054
G+W(15)	0.5241	0.5837	0.5511	0.4737	0.5409	0.5028

表 2: 検索結果 NTCIR (Average Precision)

	long			short		
	normal	rel	rest	normal	rel	rest
BL	0.5055	0.6295	0.4966	0.4110	0.6382	0.4503
OLD	0.4771	0.5287	0.4587	0.3543	0.4247	0.3587
GEN	0.4787	0.5386	0.4604	0.3540	0.4885	0.3821
G+W(10)	0.4789	0.5731	0.4742	0.3730	0.5654	0.4277
G+W(15)	0.4808	0.5787	0.4852	0.3737	0.5803	0.4322

上位数語を検索語に追加する。

3.2 検索実験の結果

本システムの有効性を検討するために、NTCIR-1 と IREX の 2 つのテストコレクションに適用した。今回提案した方法の有効性を検討するために、今回の実験では、次の 5 つのシステムについて検索実験を行い、平均精度を用いた評価を行った (表 1,2)。

ベースラインシステム (以下 BL) 検索語中の全ての語を検索拡張に利用するシステム

従来システム (OLD) 検索要求中の語の汎化による検索拡張のみを行うシステム

関連文書に含まれている語の汎化のみ (GEN) 先の手順 3 を行わないシステム

関連文書に共通する概念を利用 (G+W) 先の手順通りに検索拡張を行い、手順 3 で 10 語: G+W(10) もしくは 15 語: G+W(15) を追加するシステム

本システムでは、検索性能の評価の基準として、適合文書へのオーバーフィットの可能性というのを考えているため、各テストコレクションに対して、次の実験を行った。

初期検索の利用 (normal) 初期検索の結果を利用する。

正解文書を利用 (rel) テストコレクションで与えられた A 判定 (完全に一致) の正解文書から 5 件ランダムで抽出し利用する。

正解文書を利用し残りで評価 (rest) rel と同じ正解文書を利用するが、評価の際には、その 5 件は評価対象から除外する。

rel と rest については、10 回の試行を行って、その平均を取った。また、検索要求としては、検索要求を 1 文程度で示す short と数文にわたって表現する long を共に行った。また、評価は B 判定 (部分的に一致) までを正解として行った。

3.3 実験結果の考察

表1,2 において、おおむね OLD<GEN<G+W となっていることから、今回提案した改良手法は、全般において有効であると考えられる。また、ベースラインシステムとの比較では、IREX の場合では、正解文書を与えない場合において、良い性能を示しているが、NTCIR-1 の場合では、全般的にベースラインシステムよりも性能が悪い。ただし、正解文書を除いた評価における差は、正解文書を含んだ評価における差よりも小さく、オーバーフィットの度合いは小さいと考えている。

また、検索意図を検索者に分かりやすく提示するため、少ない検索語、もしくは、関連する概念で表わされることが望まれる。これに対し、今回の実験の内、IREX を例にとり、拡張した検索語の数について、その平均値を表 3 に示す。ベースラインシステムの検索語数に比べ、本システムは、20%程度以下の語数で上記の性能を実現しており、その点においては本システムの有効性は確認されている。

検索に役立つ指標の大きな語を 10 語加える場合と 15 語加える場合では、IREX の short の場合を除き、15 語加える場合の方が成績が向上している。元の検索要求における検索語の数と追加する語の数の比率の問題があるのではないかと考えているが、更なる考察が必要である。

表 3: 検索語の数 (平均値) IREX

	long		short	
	normal	rel	normal	rel
BL	247.5	320.4	180.2	316.3
OLD	13.1	14.9	3.3	3.8
GEN	34.2	53.5	18.1	44.8
G+W(10)	40.8	60.9	25.7	52.7
G+W(15)	45.0	64.7	30.3	56.5

今回の実験では、テストコレクション毎のばらつきが見られた。この原因の一つとしては、文書データセットの違いに依存するところが多いと考えている。つまり、IREX は新聞データを利用しており、EDR が持っている一般的な概念体系との親和性が高いと考えられる。逆に、NTCIR は論文の抄録データを利用しており、専門用語などがうまく扱えていないという事が考えられる。

逆に、IREX の事例では、関連文書として正解を与えない場合に、元のシステムよりも性能が向上している結果が得られている。関連文書を与えない場合の成績については、本手法の有効性が疑問であるという考え方と、正解文書中に含まれる語にはオーバーフィットの要因になる語が少ないからベースラインシステムの性能が上がっているという考え方があるが、個別課題への性能分析を通じて、詳細に検討する必要がある。これらについても、本システムの性能が辞書に依存することなども考慮にいれながら、システムがうまく働く場合の状況や、働かない状況について、より詳細に分析する必要があると考えている。

4 他の検索拡張の方法との比較

関連文書フィードバックによる検索拡張における検索語の選択手法として、村田ら [9] は、二項検定による検索語の選択を提案している。この手法は検索性能の向上という有効性も検証されているが、計算量も多く、処理に時間がかかる。

これに対し、本研究で提案する手法では、関連文書に特徴的に現れる概念を電子化辞書を知識源として選別することにより、適合文書に多く存在していなくても、同じ抽象化概念に属する語を利用することが可能であると言う点で、異なっている。

5 結言

本論文では、我々が提案している適合的汎化に基づく情報検索システムにおいて、検索者の意図を現わす概念を推定する方法として、関連文書に共通して現れる概念を利用する方法を提案し、検索拡張に応用した。

今回提案した手法は、すでに提案している適合的汎化に基づく情報検索システムの性能を改善させるものの、検索課題の性質によって、辞書などを使わない場合に比べて、必ずしも、性能が良いとはいえない。今後の課題としては、本手法がうまく働く検索課題のタイプの分析などを通して、更なる改良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。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 (C)(2) 課題番号 13224401) によって実施された。

参考文献

[1] 吉岡真治 他 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 –. 情報処理学会情報学基礎研究会, 2002-FI-67, pp. 151–158, 2002.

[2] 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.

[3] S. Sekine *et al.* IREX: IR and IE evaluation-based project in japanese. In *Proceedings of the Language Resource and Evaluation Conference*, 2000.

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

[5] 内山将夫 他 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.

[6] 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.

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

[8] 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.

[9] M. Murata *et al.* CRL at NTCIR2. In *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pp. 5-21–5-31, 2001.