知能と複雑系 116－3
（1999．5．24）

# 構造化文書に基づくインタラクティブな意味的情報検索

橋田浩一
電子技術総合研究所知能情報部
*hasida@etl.go.jp*

豊浦潤　津高新一郎
RWCP情報ベース機能三菱研究室

文書全体ではなく、質問への答えなどの情報をピンポイントで抽出する技術が求められている。しかし、検索対象である文書集合の内容をよく知らない多くのユーザにとって、初めから適切な検索要求を発行するのは不可能である。また、ユーザの検索要求と文書集合中の言語表現とのギャップを自動的に埋め合わせるのも現在の技術では難しい。そこで、探索空間を効率的に絞り込むように検索要求を改訂するためのヒントを次々にユーザに与えることにより、インタラクティブに検索を進める方法を考える。そのヒントは、文書中の意味的な依存関係に基づいて生成することができる。そのような依存関係を推定する方法と検索効率の関係について検討する。

# Semantic Information Retrieval from Structured Documents

HASIDA Kôiti
Electrotechnical Laboratory

TOYOURA Jun　TSUDAKA Shinichiro
RWCP Information-Base Functions
Mitsubishi Laboratory

Increasingly needed is a technology for pinpoint retrieval of particular pieces of information rather than entire documents. Without enough knowledge about the set of documents to search, however, the user cannot issue an appropriate query from the outset. It is also very difficult to automatically plug the gap between the user's query and the linguistic expressions in the documents. In interactive information retrieval, the user should be provided with hints on how to revise the query so as to efficiently narrow down the search space. Such hints can be generated from semantic dependencies in the documents. The efficiency of retrieval is discussed in relation to the methods to obtain such dependencies.

## 1　はじめに

従来の情報検索は、正確に言えば情報の検索ではなく、文書の検索であった。しかし実際には、文書全体ではなく、特定の質問（たとえば「程度や分量が小さいために効果がないことを簡潔に言うにはどうするか」、「男子生徒用と女子生徒用に異なる規格の椅子を使っている学校はあるか」、「他のファイルへのポインタであるファイルを何と呼ぶか」など）への答えのようなピンポイントの情報を検索したいという需要も無視できない。いや、むしろそのような需要の方が潜在的はに多いのではないだろうか。

以下では、ピンポイントの情報検索のためにインタラクティブに検索を進める方法について述べる。それはつまり、探索空間を効率的に絞

り込むように検索要求を改訂するためのヒントを次々にユーザに与える方法である。語句の間の意味的依存関係に基づいてそのヒントを生成する方法について論じ、検索の効率について考察する。

## 2 インタラクティブな検索

ピンポイントの検索に限らず、普通のユーザがいきなり適切な検索要求を発行するのは無理である。これはしばしば、何を検索したいかが自分でもよくわかっていないからだが、たとえそれがわかっていたとしても、検索対象である文書集合の性質 (たとえばどのようなテクニカルタームが使われているか) がよくわかっていないことが多い。したがって、たとえば自然言語による検索要求をせよと言われてもユーザは途方に暮れてしまうことになる。そのような状況で無理に自然言語の検索要求を行なったとしても、その検索要求と文書集合中の言語表現との乖離を自動的に埋め合わせるには、シソーラスによる展開などの比較的扱い易い手法だけではなく、組合せ的な推論を行なう必要があり、それを現在の技術で精度よく実現するのはまず不可能だろう。

ユーザの知識の欠如を自動的に補完するのが無理なら、ユーザの知識を増しながらインタラクティブに検索を進めるのが有効と考えられる。それには、検索範囲を適切に絞り込むように検索要求を改訂するための具体的なヒントをユーザに与える必要がある。

インタラクティブな情報検索における探索範囲の変更には、従来は関連フィードバックなどの統計的手法が用いられてきた。たとえば Yang et al. (1999) は、複合名詞に基づく検索要求の改訂と関連フィードバックに基づく探索空間の重み付けの変更を行なっている。また Excite や Infoseek では、候補として挙げられた各文書に対してそれに似た文書を検索することができる。こうした統計的な方法は、それぞれある程度まとまった大きさを持つ文書が検索の対象である場合には有効だが、ピンポイントの情報検索には向かない。また、Excite のように統計的な方法で新たなキーワードを提案してくるシステムもあるが、やはりピンポイントの検索の場合には適切なキーワードを統計的に推定するのはほぼ不可能である。

たとえば「他のファイルを指しているだけのファイルを何と呼ぶか」という質問の答え (「エイリアス」、「ショートカット」など) を求めたい場合、「指す ファイル」または 'point file' のようなキーワードの列によって検索を始めることになるだろう。当然、このような一般的なキーワードの組合せだけで探索空間を十分に絞り込むのは不可能であり、多くの文書が候補に挙がる。AltaVista を 'point+file' ('point' と 'file' を両方含むという条件) で検索してみると，1,571 個の文書が照合する (日本標準時 1999 年 4 月 29 日 21 時; しかも、その中には求める情報が含まれない)。その上で各文書をチェックして目的の情報に到達するには膨大な手間がかかる。

そこで検索範囲を適切に変更しつつ絞り込んで行く必要があるが、そのように検索要求を改訂するためのヒントを統計的な方法で自動生成することはピンポイントの検索においては不可能である。たとえば、Excite で 'point file' によって検索すると、files、server、upload、imagemap、ncsa、lines、cgi、m-x、ftp、kilobyte という新たなキーワードの候補が提案されるが、この中のいずれも、「他のファイルを指しているだけのファイルを何と呼ぶか」に対する答えを求める役には立たない。

## 3 意味的依存関係の利用

ピンポイント検索におけるヒントは、各文書の統合的性質だけでなく、意味的な依存関係に基づいてヒントを生成することができる。たとえば、「他のファイルを指すだけのポインタであるファイル」という表現は「他」から「ファイル」への (「の」を介した) 依存関係、「ファイル」から「指す」への (「を」を介した) 依存関係、および「ポインタ」から「ファイル」への (「である」を介した) 依存関係を含む。この表現が探索範囲の中に現われるとき、「ファイル」という検索要求に対して、それと依存関係にある「他」、「指す」、および「ポインタ」とを新たなキーワードの候補として提案することができる。

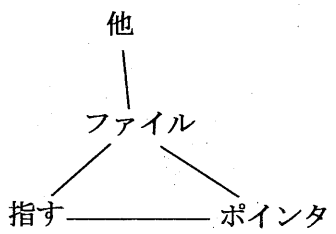


図 1: 依存関係のネットワークの一部

そのための計算には、文書集合全体から得られる依存関係のネットワークを用いる。このネットワークの各節点は語に対応し、文書集合中でのその語のすべての生起を表わす。また、各結線は 2 つの語の間の依存関係に対応し、文書集合中でのその依存関係のすべての生起を表わす。たとえば「他のファイルを指すだけのポインタであるファイル」という表現が文書集合中に現われれば、依存関係のネットワークは図 1 のような部分を含む。図の中の「ファイル」は、「他のファイルを指すだけのポインタであるファイル」の中の 2 つの「ファイル」の生起を含む。このようなネットワークの中で結線を辿ることによって新たなキーワードに関するヒントを生成することができる。言うまでもなく、検索の精度を上げるためには、依存関係は共参照を考慮に入れた上で定義することが望ましい。

ここで言う依存関係としては、言語学の研究用にはあらゆる種類の語の間の依存関係を考える必要がある場合もあろうが、実際の多くの用途においては、上記のように付属語を介した自立語の間の依存関係を考えるのがよいだろう。すると、自立語の間を仲介する付属語 (の列) は意味的な 2 項関係 (たとえば「を」は被動作対象など) を表わす。検索においては、その関係の種類と向きは無視するのがよいかも知れない。これは、各語に対してそれと結線で繋がっている語は一般に非常に多く、関係の種類と向きでさらにそれを分類するとユーザの選択肢が多くなり過ぎると思われるからである。なお、依存関係を用いた検索に関しては新美他 (1998)、Hyoudo et al. (1998)、立石他 (1999)、中山・松本 (1999) などの研究があるが、インタラクティブな検索に関する検討はあまり行なわれていない。

この方法における検索要求は、単なる語の集合ではなく、検索要求に含まれる語の間の依存関係の集合からなる部分ネットワーク (これを検索要求ネットワークと呼ぼう) と考えるのがよいだろう。つまり、検索要求に含まれるある語と依存関係を持つ他の語を検索要求に含めることにより、それらの間の依存関係が検索要求に含められるわけである。その場合、検索要求の改訂とは、この検索要求ネットワークを変化 (たいていは成長) させることである。これによって、検索要求ネットワーク中の依存関係の生起を含む文書ではなく、検索要求ネットワークと照合する文書中の構造を検索することができる。これによってピンポイント検索が可能になる。

検索要求の改訂の際に新たに検索要求ネットワークに取り込むことができる語の候補は一般には多いので、それらの語の間に何らかの順位を付けることによってユーザによる選択を容易にする必要がある。その順位は tfidf のような評価値に基づいて定義するのが良いだろう。ただしこの場合、tf は term frequency ではなく語 $a$ と語 $b$ との間の依存関係の頻度であり、idf は inverse document frequency ではなく $a$ と依存関係を持つ $b$ の生起の頻度から求まる (実際には、$a$ と $b$ との間の依存関係の頻度は、$a$ と依存関係を持つ $b$ の生起の頻度と等しいと考えてよい)。ここで $a$ はすでに検索要求ネットワークに含まれている語、$b$ は検索要求ネットワークに取り込まれる語の候補である。もちろん、ここで考慮される $a$ の生起は、検索要求ネットワークと照合する文書の部分に含まれるものに限られる。

このように依存関係のネットワークの各部分に多数の文書を多数の部分を代表させることにより、文書集合 (正確には文書の部分の集合) 全体の様子を見渡すことができることに注意されたい。従来のほとんどの検索法では、正解文書の候補の集合が得られたら後はそれらを個別に調べざるを得なかったが、その候補の個数が非

常に多くて正解が少ない場合には手間がかかり過ぎる。これに対し、正解の候補の集合全体の性質を検索要求に関連する範囲で要約したものをユーザに提示し、それに基づいてインタラクティブに検索を進めることにより、その手間を劇的に減らすことができるだろう。

## 4 解析と検索の性能

依存関係のネットワーク自身のサイズは文書集合全体よりもはるかに小さい。また、依存関係に関する解析の曖昧性がなければ、各節点および各結線から文書集合中の語および依存関係の生起へのポインタに必要な記憶容量は、依存関係の曖昧性がなければ文書集合の大きさにほぼ等しい。これは、各語が統語的には高々1個の他の語に係ると考えられるからである。

解析が曖昧性を含む場合に文書の各部分について多数の解析結果を考慮すると、このポインタの個数が多くなり、索引のサイズがもとの文書集合全体の数倍に及ぶことになる。したがって、検索の精度と記憶容量とのトレードオフを考える必要がある。ただし、依存関係を求めるために必要な (制御や非有界依存の解析を含む) 統語解析と共参照の解析の精度は現在はいずれも 80%以上になってきている。したがって、解析に曖昧性がある場合でも、上位の解析結果だけを用いることにより、索引のサイズを抑えつつ十分な精度が得られる可能性がある。

しかし、意味関係 (主題役割や修辞関係) などを含むさらに詳細な意味構造に関しては自動的な解析の精度はまだかなり低いので、そうした意味構造を利用した情報検索においては索引のサイズと検索の精度および効率との関係が問題になる。自動解析した構造に基づいてそのような情報検索を行なうのはまだ無理かも知れないが、人手修正済みの GDA タグ (橋田他, 1999) のようにかなり細かくかつ精度の高い構造を用いれば高度な意味検索も十分可能だろう。

## 5 おわりに

現在、以上で述べた方法の妥当性を検証するための実験を行なっている。特に、依存関係を語の間の単なる共起関係から推定した場合、自動的な解析から推定した場合、および人手修正の結果から求めた場合について、検索要求ネットワークの変更の際の新たな語の選択の効率を比較している。

また、GDA タグなどによって高度に構造化された文書とそうでない文書が共存する文書集合からのインタラクティブな情報検索について考えたい。正確にタギングされた文書とそれ以外の文書に同様に適用可能であり、しかも前者に関しては高い効率と精度の検索を可能にするような方法があれば、タギングを一般の利用の場面に普及させるための有力な手段となるだろう。

## 参考文献


立石 健二・峯 恒憲・雨宮 真人 (1999). 係り受け構造や語の意味情報を利用した WWW 上での日本語テキスト検索システム. 『言語処理学会第 5 回年次大会論文集』, pp. 317–320.

橋田 浩一・長尾 確・内山 将夫・Christoph J. Neumann・高橋 直人 (1999). GDA タグ集合の設計と応用. 『言語処理学会第 5 回年次大会論文集』.

Hyoudo, Y., Niimi, K., & Ikeda, T. (1998). Comparison between Proximity Operation and Dependency Operation in Japanese Full-Text Retrieval. *Proceedings of 21th ACM SIGIR Conference*, pp. 341–342.

中山 拓也 ・松本 裕治 (1999). 文節共起を利用した文章検索支援. 『情報処理学会研究報告 99-NL-130』, pp. 33–40.

新美 和彦・兵藤 安昭・池田 尚志 (1998). 係り受け情報を用いた全文検索とその評価. 『第 11 回デジタル図書館ワークショップ』, pp. 27–34.

Yang, K., Maglaughlin, K., Meho, L., & Robert G. Summer, J. (1999). IRIS at TREC-7. *The Seventh Text REtrieval Conference (TREC 7)*. http://trec.nist.gov/pubs/trec7/t7_proceedings.html.