意味検索のプロトタイプシステムの構築

大倉 清司　　潮田 明

（株）富士通研究所

1. はじめに

近年、検索技術を利用する機会が多いのはいうまでもない。例えばインターネット上のファイルを検索する検索エンジンのページビュー（日本国内）は2008年10月時点ですでに60億ページビューを超えている[1]。しかし、意図した検索結果が常に上位にランクされるとは限らず、検索精度はまだ十分とは言えない。

現状ではキーワードをクエリーとすることが多いが、通常は単語のみを指定するため、単語間の意味があいないなまま検索が実行されてしまう。その結果、意図しない内容が検索されてしまうことが多い。これに対して本研究では、自然言語文で検索したいことを入力すると、それを意味解析し、単語間の意味関係をふまえて検索ができる意味検索技術を開発した。インターネット検索のような汎用的な検索エンジンの開発は、膨大な文書が必要である上にその文書を解析しなければならないため、基礎研究としてのプロトタイプシステムの開発には向かない。最初は検索対象を限定し、従来技術と比較できるような検索エンジンを開発することにした。著作権やデータの入手の容易性から本研究では特許明細書のテキスト部分を対象にすることにした。

特許明細書を検索する場合、クエリーに近い文書だけではなく、文書でクエリーに近い文を出力する重要性もあると考える。今回開発したプロトタイプシステムでは、特許明細書を対象に意味解析を行いデータベースを構築し、文書だけではなく文単位の検索も可能である。

2. 従来技術

テキストを対象とする検索エンジンは、データベースの作成方法により大きく3つに分類できる。1つ目は、検索対象を解析しないでデータベースを作成する方法である。これには文字単位の統計情報などからインデックスを作成する方法が含まれる。2つ目は、検索対象文書を浅く解析する方法である。英語の場合は単語の語尾処理を行い正規化したり、日本語の場合は形態素解析をしたりした後にインデックスを作成する。3つ目は、検索対象文書を深く解析する方法である。深い解析には構文解析や意味解析が含まれる。この方法は検索意図に即した検索に有用といえるが、解析の精度が検索精度に影響してしまうという問題がある。

クエリーの入力方法には、キーワード入力と自然言語文入力の2種類がある。キーワード入力は、上で述べた全ての作成方法で作成されたデータベースを検索できる。検索対象文書を深く解析することにより文書内の単語間の構造を計算し、連想キーワードを表示する技術が提案されている[2]。また、文単位で依存構造を計算しておき、キーワード系列の入力に対して系列中のキーワードがデータベース中の各文において形成する依存構造パターンを同定し、パターンごとに検索結果を分類する用例文検索が提案されている[3]。

しかし、キーワード入力は複数の単語による検索であり、単語間の意味関係があいまいのまま検索が実行されてしまう。この問題は[2]でも指摘されている。例えば、特許検索において「機械翻訳結果で、訳文の単語を修正すると辞書が更新される」と言った内容の特許を検索したいとする。特許庁の公報テキスト検索[4]で、公開特許公報および特許公報の公報全文を対象に「単語　修正　辞書　更新」をAND検索すると、検索結果は2488件になってしまう。意図した検索結果が上位にランクされればよいが、キーワードにより意図される意味はあいまいである。単語を修正するのか、単語を更新するのか、またその両方か、あるいは、単語および辞書を修正するのか、キーワードの羅列だけではわからない。例えば、「辞書」「修正」というキーワードを含み、かつ、単語が自動的に更新されるという特許が検索されると、検索者からすれば意図しない検索結果になるが、検索エンジンの仕様としてはキーワードがヒットしてい

るため間違いとはいえない。また検索された各文書の評価値は、キーワードの出現頻度および出現位置からしか計算できず、ユーザーが意図しないランキング結果になってしまう。

これに対し、自然言語文入力においては、入力されるクエリーがキーワードに比べて情報量が多いので、適切に処理を行えばキーワード入力よりも高い検索精度が期待できる。入力された文を対象に形態素解析を行うだけのものと、構文解析に加えて語句と語句の意味的な関係を解析する処理などを含む意味解析まで行うものがある。例えば意味解析を使用すると、「単語を修正する」と検索したとき、「単語」が「修正する」の対象となっている文を含む文書をマッチさせることができる。これは、「単語」「修正」を同時に含む文書よりも大幅に絞り込まれる。

しかし、データベースの作成と同様に、自然言語文により検索するアプローチは、高精度に深い解析をするのが難しく、解析精度が検索精度に大きく影響してしまう。

特許検索などにおいては、単語間の関係（動作を表す単語とその動作の対象など）をクエリーとして規定したい場合が多い。その場合、文の意味構造をふまえた検索ができなくてはならない。そのため、本研究では、データベースの作成と入力キーの解析の両方に意味解析を使用することとした。

3. 「意味」の定義

文書が表す意味とは、文書中に含まれる文の意味構造の総和とみなすことができる。また今回対象とする特許明細書の検索では、クエリーに近い文書が検索されることだけでなく、クエリーに近い文が検索されることの重要性にも着目する。

本研究における意味構造とは格文法[5]に基づいたグラフ構造により表現される。すなわち、意味構造は、単語の概念を表すノードおよびノード間の関係を表すアークからなる有向グラフにより表される。図 1 は、「太郎は花子に本をあげた。」の意味構造を有向グラフにより表した 1 例である。

図 1 において、○で囲まれたものがノードを表し、□で囲まれたものがアークを表す。ノードは、”GIVE”,”HANAKO”,”TARO”,”BOOK”の 4 つである。アークは”中心”,”目的”,”動作主”,”対象”,”述語”,”過去”の 6 つである。アークはノード間の関係を表す。1 ノードにしかつながらないアークはそのノードの属性を表す。

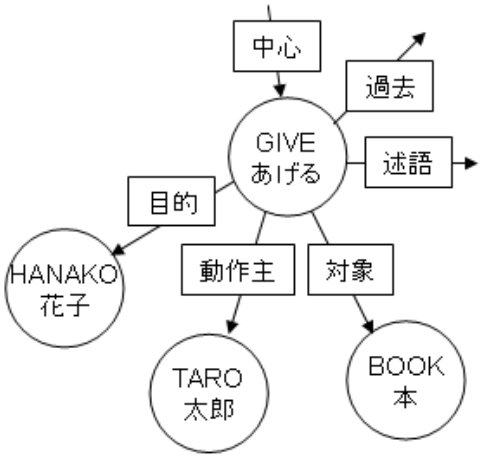

図 1. 意味構造

意味を有向グラフとして表現したとき、意味検索はグラフマッチングとして考えることができる。しかし、単純なグラフマッチングとした場合、以下の問題がある。

（1） グラフが複雑であればあるほどマッチ率が下がるため、グラフ全体をマッチさせるのは現実的でない。
（2） 意味解析処理において部分的に解析誤りがある場合にグラフ全体はマッチしない。

この問題に対処するため、意味構造を、アークとそのアークにつながるノード(1 つか 2 つ)の部分グラフに分解し、この部分グラフにより検索することにした。この部分グラフを「意味最小単位」と呼ぶこととする。部分グラフ構造においてアークにつながるノードが 1 つしかない場合はアークがつながらない箇所を”NIL”と記述することにする。以下は図 1 の意味構造から意味最小単位を抽出した例である。

(GIVE, HANAKO, 目的)
(GIVE, TARO, 動作主)
(GIVE, BOOK, 対象)
(GIVE, NIL, 述語)

(GIVE, NIL, 過去)
(NIL,GIVE,中心)

4. 意味構造の抽出

意味検索を実装するにあたり、意味構造を計算する仕組みが必要である。本研究では、日英翻訳エンジン ATLAS[6]の翻訳過程から意味構造を取り出すことにした。ATLAS は中間言語方式の翻訳方式を採用しており、原文を辞書と文法規則に基づき解析して意味構造を計算する。意味構造においては、ノードは概念記号と呼ばれる意味を表す記号で識別される。

5. 検索エンジンの設計

クエリー、検索対象文書を意味最小単位の総和としたときに、そのマッチング方法および各文書の評価値計算方法が重要である。

ATLAS の概念記号は形態素よりも抽象化されている。そのため、意味最小単位のマッチは完全一致とする。

評価値計算に関しては、意味最小単位の検索は文単位で行うため、文書(Dk)の評価値は文の評価値(Si)の総和とすることにした。すなわち、

Dk = Σi Si

文の評価値は、以下のようにして行う。データベース作成にあたり、検索対象文書を意味解析して意味最小単位を抽出し、それぞれの意味最小単位の逆文書頻度を計算しておく。検索時では、入力キーを意味解析して得られた意味最小単位 Mn につき、文中の Mn の出現回数に Mn の逆文書頻度をかけあわせる(=En とする)。さらに、En を合計したものに、文中に出現した M の数の 2 乗をかけあわせる。すなわち、文 i の評価値(Si)は以下の式で表される：

Si = (Σn Idf(i,n)×Freq(i,n)) × (Σn C(i,n))^2
ただし、
Idf(i,n) =文 i に出現する Mn の idf 値
Freq(i,n) = 文 i における Mn の出現回数
C(i,n) = 0 if 文 i に Mn が出現せず
C(i,n) = 1 if 文 i に Mn が出現する

文中に出現した M の数の 2 乗をかけあわせる理由は、入力キー内の意味最小単位が同時に 1 文中に現れれば現れるほど、その文が入力キーに類似した有向グラフを含むと考えられるからである。例えば、意味最小単位の同時出現数が 2 より 3 のほうがはるかに意味が類似する可能性が高い。これにより、解析誤りにより有向グラフが一致しなくても、意味最小単位が同時にマッチすれば評価値が高くなる。

自然言語文で入力したとき、意図しない意味最小単位で検索対象文書を検索してしまう場合がある。この問題に対しては、ユーザに、意味最小単位に相当するクエリー中の表現を提示して選択させることで解決することにした。

6. プロトタイプシステムの開発

以上で説明した検索方法により、意味検索のプロトタイプシステムを構築した。

検索結果は文書ごとにランキングされて、入力キーと類似しているものから順に表示される(図 2)。

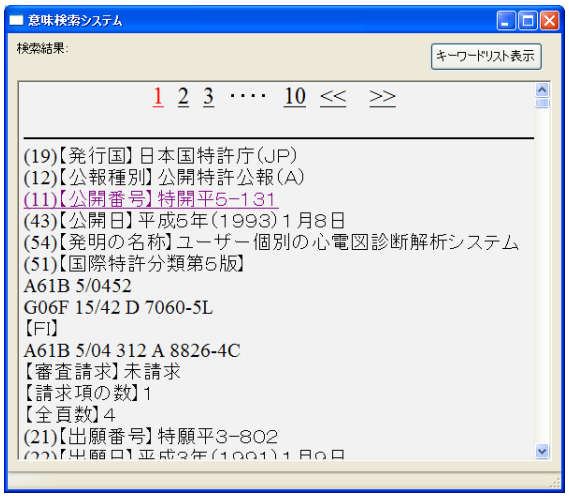

図 2. 検索結果画面

検索された文書の 1 つをクリックすると、明細書が表示されるが、評価値が高い文が強調表示される(図 3)。

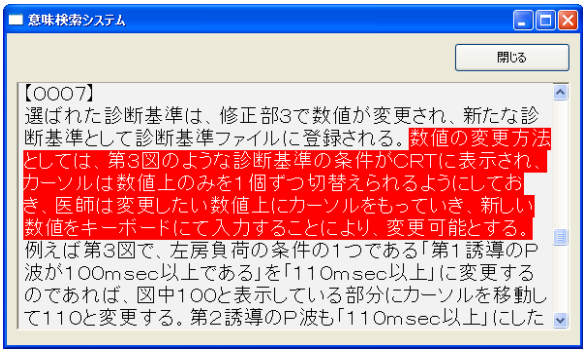

図 3. 明細書表示画面

評価値が高い文をクリックすると、その文に意味的にマッチする表現がハイライト表示される(図 4)。

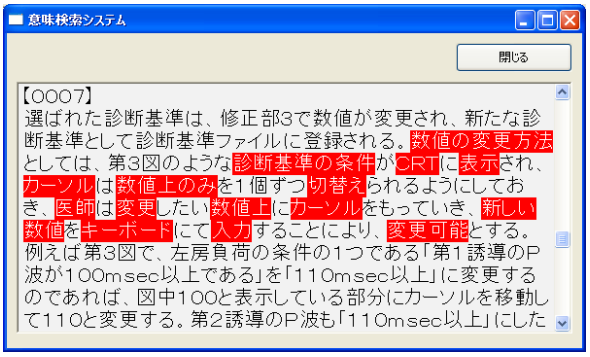

図 4. 明細書表示画面

意図した検索結果が得られない場合は、図 2 の「キーワードリスト表示」ボタンを押すと、入力画面の右側にキーワードリスト表示領域が現れる(図 5)。

7. 今後の展望と課題

今後の課題は、本技術を従来の検索技術と比較し、意味検索エンジンの精度・性能を検証することである。意味検索の精度評価のためには課題となるクエリーと検索対象文書を設定しなければならない。その上で、文書の評価値計算方法などを洗練させて精度向上を目指す。

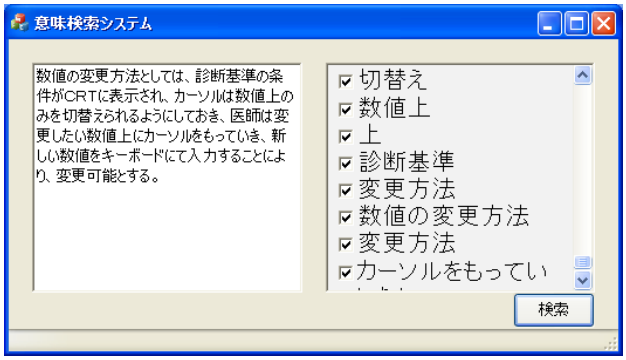

図 5. キーワード展開画面

参考文献

[1] http://www.netratings.co.jp/email_magazine/2009/01/NNR20090115.html
[2] 橋田浩一他(1999) 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. 情報処理学会研究報告 99-ICS-116, 13-16.
[3] 加藤芳秀他(2006) 依存構造に基づくコーパス検索. 電子情報通信学会論文誌. v.J89-D, n.12, 2006, p.2766-2770
[4] http://www.ipdl.inpit.go.jp/homepg.ipdl
[5] Fillmore, Charles J. (1968) The Case for Case In: E. Bach and R.T. Harms (eds) Universals in. Linguistic Theory. Holt, Rinehart and Winston, New York. pp. 1-88
[6] 富士通.英日・日英翻訳ソフト ATLAS. http://software.fujitsu.com/jp/atlas/