論文内容の要旨

論文題目　　仮説推論法と文書主題抽出法に関する研究

氏　名　　松尾　豊

本論文は，大きく第I部と第II部から構成される．第I部は仮説推論の高速解法とその周辺であり，大量の知識が与えられたときにどのように高速に推論処理を行うかを対象とする．第II部では，文書からのキーワード抽出を扱う．ここで開発したいくつかのキーワード抽出手法は，文書からどのように知識を取り出せばよいのかという問題の基礎的な要素技術になると考えている．以下，順を追って説明する．

推論の枠組みのひとつとして，観測から原因を予測する仮説推論がある．仮説推論は，真か偽か不明な事柄をとりあえず真と考えて推論をすすめ、矛盾なくゴール（観測事象）を導くことができれば立てた仮説は正しかったと考える推論形式である。ゴールを証明する仮説の組は複数ある場合もあるため、コストに基づく仮説推論では各仮説に重みを付与し、その重み和を最小とする解（仮説の組）を最も望ましいものとする。

コストに基づく仮説推論は，NP完全またはNP困難である．したがって、コストに基づく仮説推論の最適解を見つけるには、最悪ケースで問題規模に対して指数オーダの推論時間がかかる。そこで、コストが最小に近い準最適解を、平均として多項式オーダの推論時間で求める研究が行われてきた。仮説推論問題を，最適化問題に変換する方法は今までにいくつか提案されてきたが，それらは最も基本的な次の2つの置き換え法にまとめられる．

次のような節があるとする．

$$p1 \vee \neg p2 \vee \neg p3 \tag{1}$$

この節は3つのリテラルのうちどれかが真になればよいという要請を表わすので、真＝1，偽＝0を表す

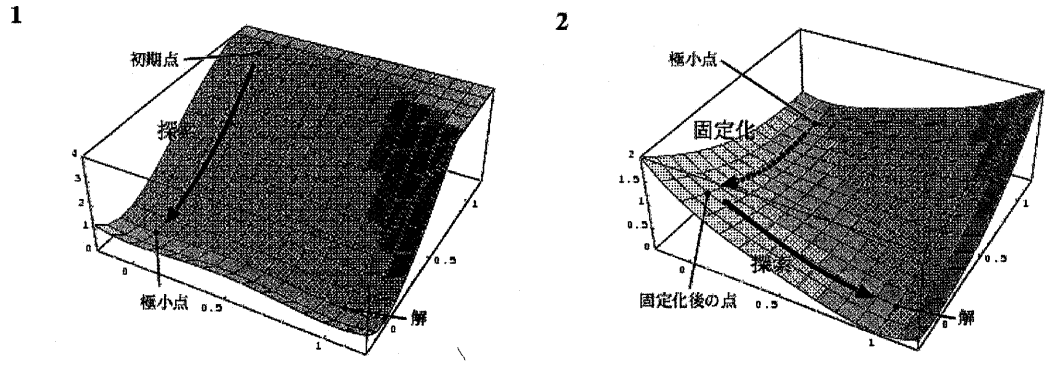

図 1: 探索の様子

とすれば，以下の不等式制約と等価である．

$$x_{p1} + (1 - x_{p2}) + (1 - x_{p3}) \geq 1$$

このような線形不等式制約をへの置換を置換 L と呼ぶことにする．目的関数を

$$\text{Minimize} \quad f = \sum_{i \in H} w_i x_i$$

とし，置換 L による制約に基づく問題を問題 L と呼ぶ．(ただし，w_i は仮説 i のコスト．)

一方，式 (1) は，

$$\neg(\neg p1 \wedge p2 \wedge p3)$$

と変形できるので、$\neg p1 \wedge p2 \wedge p3$ が偽となればよい．したがって、以下のように等式制約でも表現することができる．

$$(1 - x_{p1}) x_{p2} x_{p3} = 0 \tag{2}$$

これを置換 NL とよび，この制約に基づくコスト最小化問題を問題 NL とよぶ．

これらの置き換え法を利用して仮説推論の準最適解を得る SL 法を開発した．SL 法は，まず問題 L を解くことで実数領域の最適解を得た後，問題 NL をペナルティ法で解くことで，実数最適解近傍の 0-1 解を得ることができる．ただし，探索が局所解に陥ることが発生するので，その際には非充足な節を矯正するように変数の値を変え，その値で固定するという処理を行う．つまり，非線形関数の降下 (Slide-down) と持ち上げ (Lift-up) を交互に行う (図 1)．この方法では，仮説数 n に対し，$n^{1.8}$ の多項式オーダの探索時間で準最適解を得ることができた．

また，置換 L と置換 NL という性質の異なる 2 つの制約を同時に扱う手法も開発した．拡張ラグランジュ法という枠組みを用い，複数の種類の異なる制約を扱う．探索の初期フェーズでは置換 L による制約を利用し，コストの低い領域を探索する．充足していない節があれば，置換 NL による制約を順次付加していくことで，高速に質の高い解を探索することができる (図 2)．SL 法よりもさらに高速にコストの低い解を探索することができた．

さて，アルゴリズムの性能を正しく評価するには、用いる問題に対する考察が不可欠である．そこで，同じタイプの問題のパラメータを変えることで，問題の難しさがどのように変化するかを考察した．同じタ

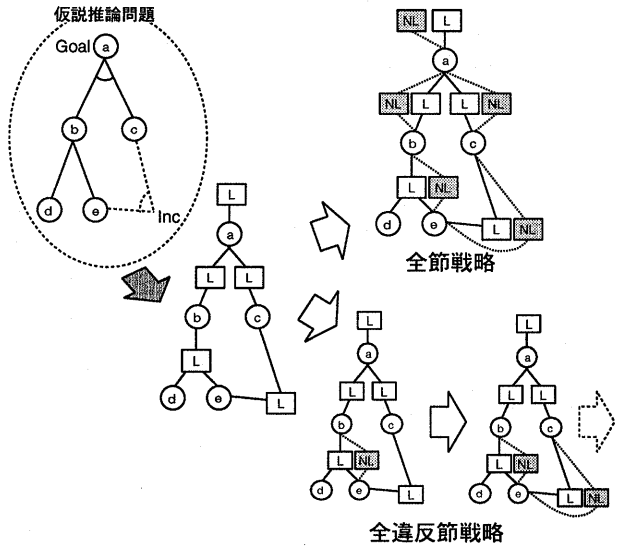

図 2: 制約プロセッサ NL を加えていく 2 種の戦略

イプの問題でも，制約を徐々に厳しくしていくと，置換 L の有効性がある領域で急激に悪くなることを明らかにした．また，問題を解く前に，どの程度難しい問題であるかを，解の数を見積もることで計算する方法についてもいくつか考案した．

仮説推論は，人間がある知識を記述し，計算機が推論して解を提示するという仕組みであるが，これをひとつのシステムと考えたときには，どのように簡単に知識を表現すればよいか，どのように結果を分かりやすく提示すればよいかという視点も重要である．仮説推論の目的関数が複数の場合や，重要な制約を解と同時に提示するといった，仮説推論をより使いやすくする拡張についても考察した．

第 II 部では，主に文書からのキーワード抽出というテーマを扱う．近年では大量の電子文書が入手可能であり，テキストマイニングの研究もさかんである．大量の文書から知識を抽出することができれば，推論の枠組みを十分に生かすことができるだろう．

まず，ひとつの文書内における語の共起を考える．同じ文内に同時に出現した場合に 1 回共起し，頻出語と各単語の共起回数を数える．仮に、語 w が頻出語 $g \in G$ と全く独立に生起するなら、語 w と語 $g \in G$ が共起する確率は頻出語単独での生起確率と同様の分布になるはずである．一方、語 w と頻出語 $g \in G$ の間に何らかの意味的なつながりがあれば、この確率は偏ることになる．“achieve” や “case” のような一般的な語は偏りが少ないが (図 3)，“non-linear function” や “hypothesis” のような文書の主題と特に関連した語は偏りが大きい (図 4)．したがって，この偏りが大きい語をキーワードとして取り出す．ここでは χ^2 検定を用いて偏りを計る．頻出語単独での生起確率を理論確率 $p_g(g \in G)$ とし、語 w と頻出語群 G の共起の総数を n_w、語 w と語 $g \in G$ の共起頻度を $freq(w, g)$ とすると、統計量 χ^2 は以下の式で与えられる．

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w,g) - n_w p_g)^2}{n p_g}$$

$\chi^2(w)$ の大きな語 w が理論確率分布からのずれが大きな語である．この手法により，単一文書だけから従来手法を上回る性能でキーワードを取り出すことができた．

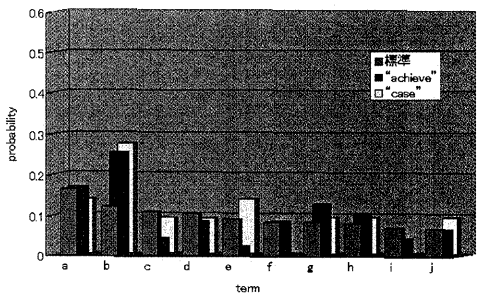

図 3: 語 *achieve, case* の頻出語との共起の確率分布

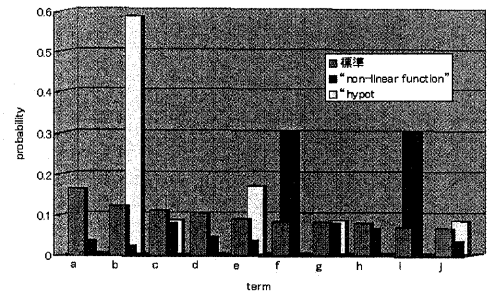

図 4: 語 *"non-linear function", "hypothesis"* の頻出語との共起の確率分布

共起関係をもとに，単語の共起グラフを作ることもできる (図 5)．各ノードは語を表し，エッジは単語間の共起を表す．このグラフは，近い概念の語同士はたがいに結ばれクラスタになっており，そのクラスタ同士が緩くつながっている "Small World" の特徴を備えたグラフとなっている．Small World とは，ノードがクラスタ化されているにも関わらず，任意の 2 点間のパス長が短いグラフである．ひとつの文書から得られた語の共起グラフもこのような構造をしており，Small World 構造に対する貢献の高い語をキーワードとして抽出する．つまり，ひとつの語を取り除くことで，パス長が大きく減少するような語をキーワードとして取り出す．この手法では，重要な語と同時に一般的な語もキーワードとして抽出してしまうが，idf の指標を同時に用いることにより，よい性能を得ることができた．

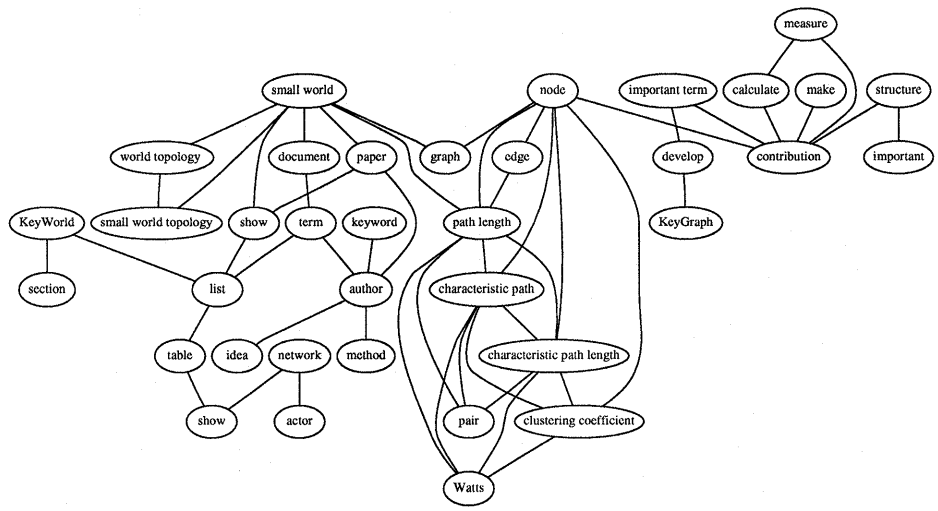

図 5: ある論文の語の共起グラフ

最後に，全体のまとめと今後の人工知能の方向性について述べ，結論とする．