

杉田勝彦† 近藤雄裕† 石川大介† 池村匡哉† 森本貴之‡ 藤原譲‡

神奈川大学理科学研究科†

神奈川大学理学部‡

1 はじめに

現在インターネット人口は急激に増加し、企業、学校、家庭など様々な環境で、多くの人が利用している。これにより情報の発信利用は多種多様かつグローバルな広がりを見せている。現代社会において、管理すべき情報は大量で複雑である。これらに対し、迅速かつ正確な精度で様々な要求が求められる。この要求を満たすため、現在計算機によるデータベースが多く用いられている。

このように大量の情報は今後も更に増えつづけていくであろう。それに対し、計算機への要求も更に高度なものになる。それには単なる情報検索ではなく高度な情報処理を行う必要がある。計算機に人間特有の高度な情報処理、類推、帰納推論や仮説生成を行わせ、人の代わりに何かを発見させようとする試みは昔から続くエキスパートシステム、人工知能がそうである。これらの処理は、共通して情報の意味を扱うものであり、そのことは重要であるといえる。そのためこれらの実用システム構築の精度の向上には、この情報の特性をより的確に捉える必要がある。情報の豊富かつ多様な意味を理解するためには概念間の意味関係を十分に記述する必要がある。そして、情報を知識として構築することにより、より高度な情報処理の実現に近づくことができる。

本論文ではこれら情報の概念間の意味関係抽出による知識構造の構築に関する並列化の研究について述べる。

2 知識構造の構築

概念間の各種意味関係を自動的に結合、調整するためのシステムとして、同値関係を抽出する C-TRAN 法、階層関係と関連関係を抽出する SS-KWEIC 法、意味関係を抽出する SS-SANS 法、意味解析を行う SANS 法、そして、それらを統合・調整する INTEGRAL 法がある。情報を網羅的に収集し、それから C-TRAN 法、SS-KWEIC 法、SS-SANS 法、SANS 法を用いて意味関係を抽出し、INTEGRAL 法により統合、構造化することにより知識構造を構築する。[1]

大量の情報の管理と有効な利用のためには、その意味関係と目的に対応して構造化することが必要である。

Knowledge Construction based on Extracted Semantic relationships

Katsuhiko Sugita†, Takahiro Kondo†, Daisuke Ishikawa†, Masayuki Kikuchi†, Takayuki Morimoto†, Yuzuru Fujiwara†

†Graduate School of Science, Kanagawa University

‡Faculty of Science, Kanagawa University

情報の特性を考慮すると、概念構造は概念間の意味に対応して階層関係の他、部分的重なり、多項関係、再帰構造、内部構造、相対性、動的関係などを含み、グラフでは対応できない。そこで、ハイパーグラフの多項関係や双対性を更に拡張した相対性（概念－関係、概念－属性など）、その他の関係に対応できる概念記憶構造である均質化2部グラフモデル（Homogenized Bipartite Model: HBM）を用いることが望ましい。[2]

上記のいずれの手法も膨大な量の情報を対象とするため、またより情報を有効に活用するために均質化2部グラフを適用することを考慮すると、そのデータ量は更に膨大なものになるため、計算機の資源不足の問題は軽視できない。例えば、階層関係を抽出するために一用語辺り約 1KByte のメモリ領域を必要とする場合を考える。つまり 1 万用語では 10MByte、10 万用語では 100MByte、100 万用語では 1GByte のメモリ領域を必要とする。この例はごく単純なものであり、本来は知識構造の誤りを修正するための出典情報など、含むべき情報は他にも存在する。

このように単純なものでさえ大量データを扱う場合には多くの資源を必要とするため、単一のマシンのみで扱えるデータには限界がある。そのため多数の演算装置やプロセッサ、記憶装置を用いて相互結合した並列処理が求められる。次の章では、階層関係・関連関係の抽出を行う SS-KWEIC 法の並列化の検討を行う。本研究では C 言語を用いた MPI プログラミングにより並列実装した。

3 SS-KWEIC 法の並列化

SS-KWEIC 法 (Semantically Structured Key Word Element Index in terminological Context) は、専門用語の構成規則に基づいて、複合用語の基本構成用語の相互の関係を解析することによって意味関係（階層関係および関連関係）を自動抽出する手法である。

用語は単純語、疊語、擬音語、擬態語および合成語を含む。専門用語の造語規則はこの合成語に対する考察に由来する。合成語は主に次のようなものを指す。[2]

合成語::=複合語 | 派生語

複合語::=語基+語基 | 語基+連結要素+語基

派生語::=接辞+語基 | 語基+接辞

語基::=単純語基 | 複合語基

単純語基::=単純語

複合語基::=語基+語基

連結要素::=・ | / | の | な

接辞::=接頭語 | 接尾語 | 数詞 | 量詞

専門用語の特徴は、

- 大部分が名詞である。
- 後部分の体言類語基の性質や状態を前部分の語基が修飾、限定するなどの修飾関係が最も多い。
- 用語が複数の語基を含むことが多いこと。

今回は、大量データの扱いを並列化により実現することを検証した。

入力データは、学術情報センターの“NACSIS テストコレクション”(人工知能の分野の論文の題目と概要)を、日本語形態素解析システム“JUMAN”[3]を用いて解析した結果を用いる。並列化の場合とそうでない場合を考慮し、約 73000 語の用語を対象とした。

図 1 は 8 つのプロセッサそれぞれで SS-KWEIC 法により分散構築した概念構造の分布状態を表したものである。これは横軸に概念に含まれる用語の個数を取り、縦軸にそれらの総数を取っている。何も構造化されなかった用語はどのプロセッサも 2000 用語近く存在する。ただし、グラフ中では省略している。

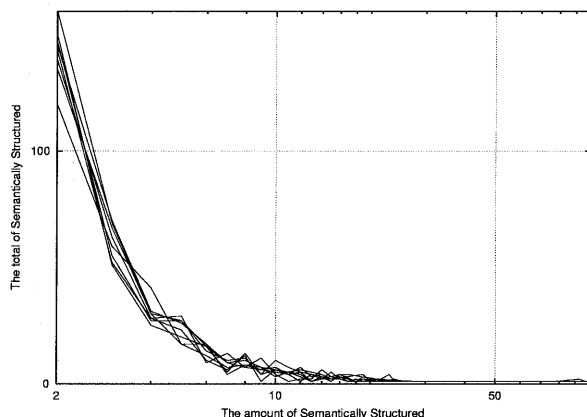


図 1: 8 プロセッサによる概念構造の分布

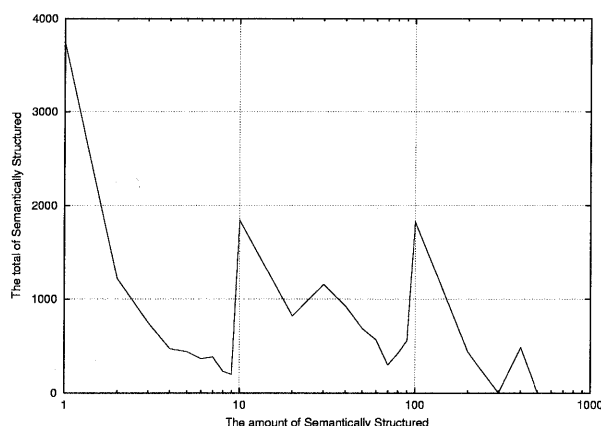


図 2: シングルプロセッサによる概念構造の分布

同じ入力データを使い、シングルプロセッサで分散せずに構築した場合は図 2 のようになる。

二つの違いから図 1 の状態では、明らかに各概念は構造化されておらず、同概念として構築されるべきもの

が拡散してしまっているのが分かる。そのためここから更にプロセッサ間で相互に概念構造のやり取りを行い、概念構造の拡散を防がなくてはならない。

分散化された概念構造の収束には、図 3 のようなシステムを実現することにより解決される。マスタプロセッサは、拡散した概念構造を保持するスレーブプロセッサに対し、概念構造の収束管理を行う。これにより同概念構造をまとめることができる。

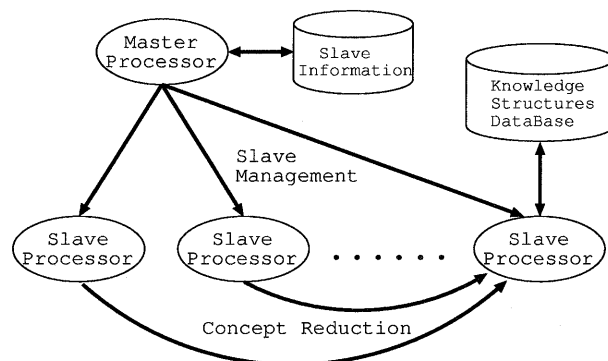


図 3: マスタによるスレーブ管理

4 むすび

知識構造構築には大量のデータ処理が不可欠である。そして、情報を表現するための構造もまた強力なものであるため、その処理量は膨大なものとなる。これら进行处理するための並列化・分散化は有効な手段である。

本研究では、SS-KWEIC 法における知識構造構築の並列化について検討してきた。現在行った並列・分散化には多くの問題点が残されており、今後それらを解決することが課題の一つである。本研究では触れなかったが、並列化による速度的な面も重要な問題である。計算機の資源不足と処理速度の問題は表裏一体である。そのため、負荷平均化などによる処理効率を考慮したアルゴリズムが求められる。

マルチプロセッサによる概念の並列・分散処理の結果、各概念構造は拡散してしまうことを述べ、これらの解決方法を示した。しかし、これは知識構造構築の初期段階が対象であり、知識構造の修正・更新などの機能に関してはより精微な後処理が必要である。

参考文献

- [1] JINGJUAN LAI, HANXIONG CHEN AND YUZURU FUJIWARA. An information-base system based on the self organization of concepts represented by terms, pp-316-325
- [2] 藤原譲. “情報学基礎論の現状と展望 —学習・思考機構と超脳計算機への応用—” 情報知識学会誌 Vol.9, No.1, pp-14-19, pp-23-24
- [3] 日本語形態素解析システム JUMAN <http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html>