理学総合研究所 平成12年度 報告書

2001-3-16

U.「脳における意味記憶構造シミュレータの産学協同開発研究」

神奈川大学　理学部　情報科学科　藤原　譲

1．はじめに

情報化の加速度的進展に伴い多種多量の情報が提供されるようになるともに、検索や数値計算のみでなく意味処理に関する高度な要求が高まり、国際的に種々の研究・開発が行われている。それには意味理解、思考機能のメカニズムが必須である。ここでは概念間の意味関係の解析から概念表現構造のモデルを確立することに成功したので、その結果を用いることにより概念意味記憶が可能となり意味理解や情報生成、思考機能などの実現の見通しが得られ、本研究はその方式の実証と実用化の第一ステップとして行ったものである。

研究の特徴は良く知られているように計算機の始まりとともに、人工知能、データベースの分野で意味処理は最大の課題であったが、エキスパートシステムで典型的に示されたようにルール主導型では辞書的支持を追加しても意味処理ができないことが、歴史的にも理論的にも明らかとなってきているので、本提案では意味内容の表現構造の新モデルによる意味理解、情報生成、思考などの実現機構に基づいていることである。

とくに重要なことは脳における記憶構造はニューラルネットワークまたはそれに関連した構造と思われているが、本研究では意味関係の解析からネットワークすなわち2項関係に対応するグラフ構造では多項関係、双対関係、相対関係、入れ子関係、再帰関係などが扱えないのでグラフを拡張したハイパーグラフを、さらに拡張した均質化2部グラフ型構造を予見しその基本機構を実証したことである。

その成果として現在の計算機が人間の脳に比し劣る点である意味関係の処理機能が実現可能となり、計算機の活用可能範囲が著しく拡張できる見通しが得られたことである。

具体的には従来から行ってきた研究成果を拡張して、

1）高分子、化学反応などの分野の専門知識を網羅的に収集、解析、評価し、意味関係を自動抽出して自己組織的に機械学習を行わせた。

2）意味構造化知識に基づき情報生成、類推、機能推論、仮説推論などの思考機能システムを設計、構築した。

3）その実用化のため脳における意味記憶構造に対応する均質化・動的・多重N次元配列・ハイパー結合プロセッサーのシミュレータの設計、試作、評価を行った。

2．意味関係抽出による知識構造と構築

2.1　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。[?]

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。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）、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。[?]

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、含むべき情報はこの他にも存在する。

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ー、記憶装置を用いて相互結合した並列処理が求められる。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。

2．2　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、畳語、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。[?]

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｜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｜畳詞

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”(人工知能の分野における論文の題目と概要)を、日本語形態素解析システム“JUMAN”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。

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。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整理すると何も構造化されなかった用語はどのプロセッサも 2000 用語近く存在することが示された。

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には構造化は著しく進むことが示された。

二つの違いから並列分散処理では、予想されたように各概念は断片的にしか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、共通に関連する部分を統合する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。

結局知識構造構築には大量のデータ処理が不可欠であるが、情報を表現するための構造もまた複雑なものであるため、その処理量は膨大なものとなる。これらを処理するためには並列化・分散化は有効な手段である。

本研究では、SS-KWEIC 法においての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。本報告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・分散処理は第一段階の高速化には有効であり、一方各概念構造は拡散してしまうが、意味関係の統合によるこれらの解決できることが示された。

3．情報化された知識の情報意味検索への適用

3．1情報意味検索

現在の計算機では数値計算やキーワード検索、演繹推論がその根底であり、豊富な情報や知識の内容を十分に活用できるとは言難く、情報や知識の意味内容に対する高度な機能の要求も強く認識されるようになってきている。

このような情報・知識の内容に関する、より高度な処理を行うためには意味理解が必要である。そして、意味理解のためには、意味関係を表現する構造が要求される。本研究では、構造化された知識の利用手段の一例として、情報意味検索への応用に関する検討について報告する。

情報化が加速的に進む現代において、情報検索の重要性は非常に高いものである。しかしながら、膨大な情報の中から目的に合致したものを効率よく検索することは非常に困難である。典型的な例としてはWebのSearch Engineでは、検索要求を厳しくすると見つからず、要求を甘くすると大量の結果が現れ、ユーザ自身による詳細な調査が要求されるといったことが非常に多い。そこで、情報検索の一例として文献検索を土台に、情報の持つ意味を考慮した検索について検討する。

一般的な文献検索の要求としては、“ある概念（用語）について記載されている文献の検索”が挙げられる。しかし、このような検索は実際には対象である概念の持つなんらかの特徴・事象の記載の有無を調査するために行われるものであり、“複数の概念がある関係を持った形で記載されている文献の検索”が本来の形である。したがって、従来から研究されている概念の出現頻度等の統計的情報ではこのような関係を示すことはできない。また、抽象的な概念による検索結果の中から興味深い文献を探すといったこともよく行われる。この場合、膨大な検索結果となることが多く、さらなる絞りこみを行うための指針が必要となる。そこで、本研究では意味関係に基づいて構造化された知識を用いることによってこれらの問題に対処する。

3．2　知識の構造化

知識を有効に活用するためには、その意味などを含めた多角的な面からの理解が要求される。そしてそのためには、以下に示す3点を実現する必要がある。

1．知識の特性とくに意味関係の解析

2．属性、特徴、意味、構造に関する基礎理論の確立、利用技術、手法の開発：体系化

3．各分野の情報への具体的な応用のためのアルゴリズム、システムの整備

また、知識の意味内容は媒体を通して表現された文字や記号を解釈するといった間接的な方法をとらざるを得ない。科学や技術の分野においては、用語、特に専門用語は抽象概念を表現する最も便利かつ強力な媒体である。そこで、概念を表現する最小単位として用語を取り上げ、この用語の体系化を行う。

このような用語の体系化において、意味関係が表現可能な構造化を行うためには多関係や入れ子構造、さらには様相生や相対性等についても表現可能でなければならない。しかし、木構造やグラフ、ハイパグラフといった従来の情報構造ではこれら全てを表現することはできない。そこで、新しい情報構造表現として均質化２部グラフモデル（Homogenized Bipartite Model:HBM）を提案している[1][2]。また、用語を基にした概念間の各種意味関係を自動的に統合、調節するためのシステム（図１）の開発も進めている[3][4]。

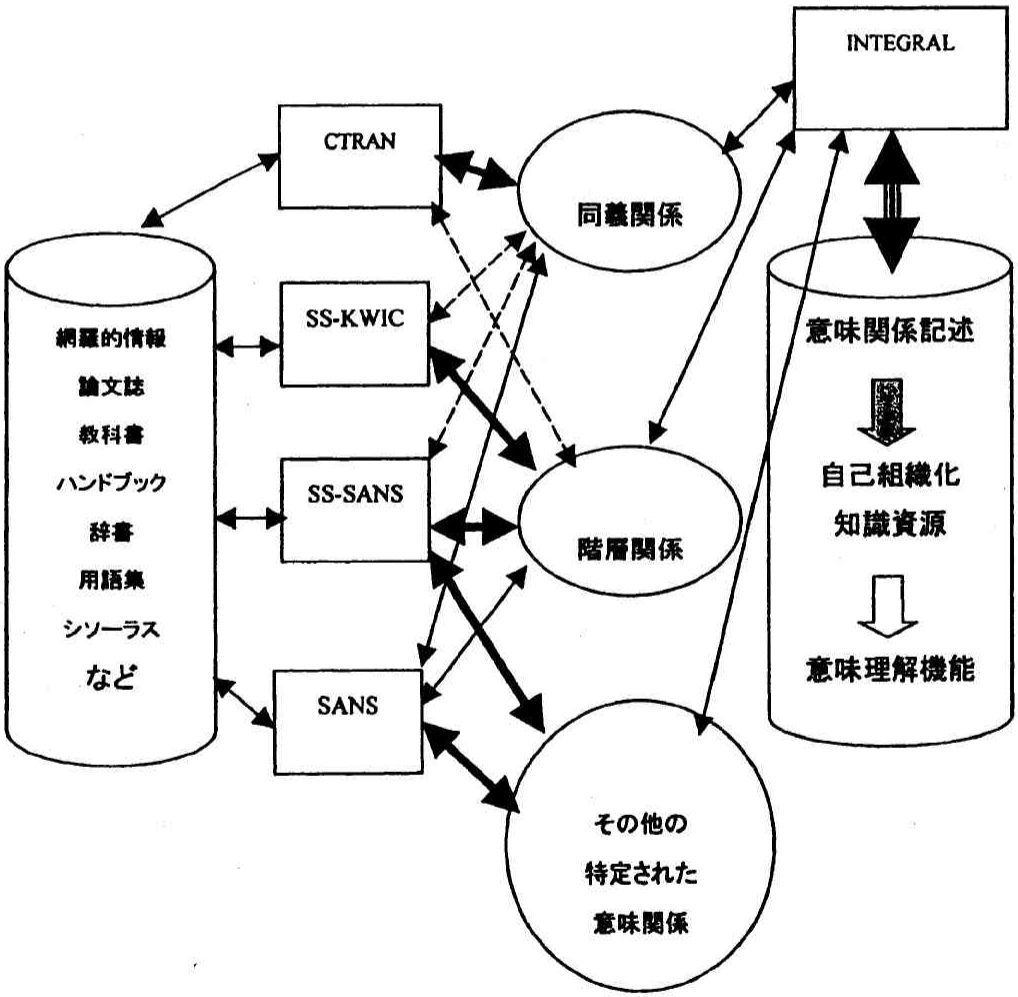

図１意味関係に基づく知識の自己組織化：学習機構

HBM は概念構造間の多種多様な意味関係を表現するために開発した情報構造である
この HBM による構造化された知識の例を図 2 に示す。図中の実線矢印は階層関係、2重の実線は同値関係、波線矢印は包含関係、1本の実線と円で囲まれた部分はそれぞれ関連関係を表す。この2つの関連関係の違いは、1本の実線が SS-SANS 法{3}によってある文献から抽出された関連関係であるのに対して、円は“並列”をキーワードとする関連関係も概念の一つである。

SS-SANS 法によって抽出された“超並列計算機”と“プロセッサ間統合ネットワーク”の関連関係は単に両用語がある文献中に含まれるというだけでなく、その文献の中に“超並列計算機”と“プロセッサ間統合ネットワーク”の組み合わさった内容が含まれていることを示す。

3．4　意味検索システム

文献検索システムは知識の構造化と検索処理に大きく分けられる。知識の構造化は図1の自己組織化システムを用いて以下の手順で行う。

1．文献データからの用語および意味関係の抽出
2．用語の語基分割（日本語形態素解析システム“JUMAN”[5]を使用）
3．知識の構造化

構造化された知識においては各用語および意味関係はそれぞれ出典情報を持ち、検索処理は各種関係のナビゲーションを行い、検索要求を満たす文献ならびに関連知識とその文献に関する情報が示される。現在、プロトタイプシステムが完成しているが、実装されているのは階層関係と同値関係のみである。

図2に構造化された知識の一部（用語“並列コンピュータ”に着目）を示す。この結果は、国立情報学研究所のテストコレクション NTCIR-2（文献データ）およびオーム社の“情報処理用語大辞典”の対訳（同値関係）を入力データとしている。

この図では、インデントは階層性を表わし、矢印は階層の方向（上位概念から下位概念へ）を、等号は同値関係を表わす。“『』”で囲まれた用語は検索要求を示す。また、用語の後にある“gakkai-j-XXXXXXXXXX”は文献のタグ情報を示す。（ただし、複数の文献に現れる場合は“...”で省略）

コンピュータ：gakkai-j-0000341470...
　→脳型コンピュータ：gakkai-j-0000342489
　→『並列コンピュータ』：gakkai-j-0000340080
　　→超並列コンピュータ：gakkai-j-0000345205
　＝計算機：gakkai-j-0000343562...
　　→ベクトル計算機：gakkai-j-0000342091
　　→並列計算機：gakkai-j-0000340082

→仮想並列計算機：gakkai-j-0000345206
→クラスタ型並列計算機：gakkai-j-0000342073
→分散メモリ型並列計算機：gakkai-j-0000342206
→超並列計算機：gakkai-j-0000340098...

図２：構造化された知識（一部）

4　むすび

加速度的に進む情報化において要求される計算機の新しい機能として、情報の意味内容に対する高度な機能の実現に向けて知識・情報の構造化に関する研究を行っている。本研究は意味関係に基づき構造化された知識の構造化、情報意味検索への応用およびそのための意味記憶構造シミュレータの設計・構築・評価に関するものである。

地域産業との連携：本研究で開発したシステムは「意味関係の学習可能な概念記憶構造の新しいモデル」に基づくもので、これからの高度情報化社会を先導する方式と製品を産み出すための研究であり、情報関連企業の多い当地に適したものである。
科学技術庁の方針に添って国-地方連携による産業政策の一環として神奈川県では先端技術の開発と産業振興を目的として川崎市にサイエンスパークを設けている。その中核機関として「かながわサイエンスパーク：（株）ＫＳＰ」のコーディネートにより、神奈川サイエンスパーク内の神奈川高度技術支援財団およびベンチャー企業、大企業と大学との連携により本研究の共同推進を図っており、本プロジェクトにも強い関心を示し、今後の展開を図ることになっている。

謝辞

本研究は理学研究所から助成された研究費に加えて科学技術庁の「地域研究開発促進拠点支援事業（略称：RSP 事業）」の一環として財団法人神奈川高度技術支援財団による「神奈川県地域研究開発促進拠点支援事業（研究成果育成型）」の助成金により研究が支援されました。またデータとして国立情報学研究所で作成された NTCIR-2 を使用させていただきました。これは科研費報告書および国内学会の提供する学会発表要旨の一部を利用して作成されました。以上のことにつきここに記載して関係機関および担当の方々に感謝申し上げます。

参考文献

[1]Y.Fujiwara and Y. Liu, The Homogenized Bipartite Model for Self Organization of Knowledge and Information, IFID 2 (1), pp13-17, 1998

[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-13-29,1999

[3]T. Morimoto, T. Maeshiro, Y. Fujiwara, Extraction of Semantic Relationships among

Terms to Construct Organized Knowledge Resources, Pro. Of 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp459-465, 1999

[4]森本貴之、真栄城哲也、藤原譲、用語間の階層・関連関係の抽出と情報の構造化、情報処理学会第 60 回全国大会講演論文集(3)、pp93-94,2000

[5]Jingjuan Lai, Hanxiong Chen and Yuzuru Fujiwara An information-base system based on the self organization of concepts represented by terms, pp-316-325

[6]日本語形態素解析システム JUMAN

[7]http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html