

2X-4 構造化された知識の情報検索への応用

森本 貴之† 近藤 雄裕‡ 杉田 勝彦‡

石川 大介‡ 池村 匡哉‡ 藤原 譲†

†神奈川大学 理学部 ‡神奈川大学 理学研究科

1 はじめに

計算機はますます高速化、大容量化し、かつ低価格化が進んでいる。また、それに伴いインターネットによる情報化が加速度的に進んでいる。しかしながら、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、情報検索への応用に関する検討について報告する。

2 情報検索

情報化が加速度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 が挙げられる。実際の 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 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等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なう。

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なうためには多項関係や入れ子構造、さらには様相性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]。

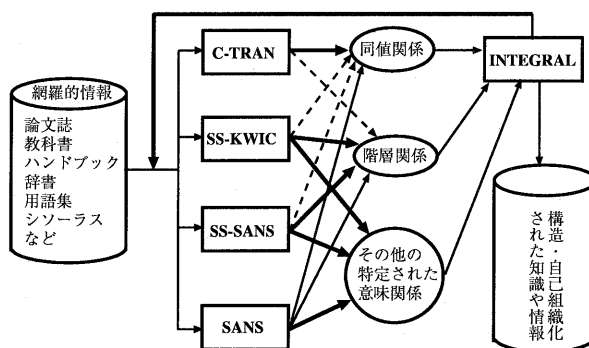


図1: 知識の自己組織化システム

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

Utilization of Structuralized Knowledge Resources for Information Retrieval

Takayuki Morimoto†, Takahiro Kondo‡, Katsuhiko Sugita‡, Daisuke Ishikawa‡, Masaya Ikemura‡, Yuzuru Fujiwara‡

†Faculty of Science, Kanagawa University

‡Graduate School of Science, Kanagawa University

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わす。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である。また、HBMにおいて、用語はそれぞれ概念を表わすが、“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“プロセッサ間結合ネットワーク”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はなく、その文献の中に“超並列計算機”と“プロセッサ間結合ネットワーク”の組み合わせられた内容が含まれていることを示す。

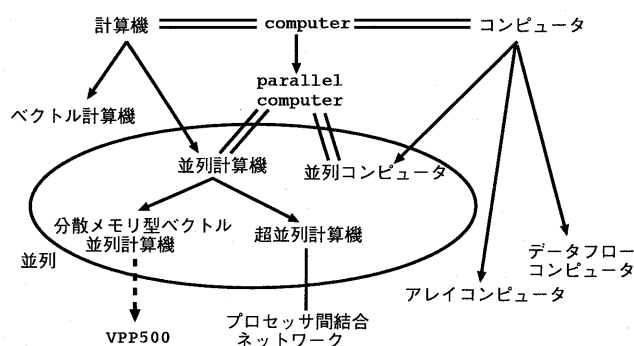


図2: HBMを用いた構造化の例

4 文献検索システム

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3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“情報処理用語大事典”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、各用語の後ろにある“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3: 構造化された知識(一部)

5 終りに

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情報検索への応用に関するものである。

今後は他の意味関係も考慮したより複雑な構造の実装と、より大量の知識構造を取り扱うための処理の並列化などに関する検討を行う予定である。

謝辞

本研究は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た。

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*, Proc. 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] <http://www.nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html>