構造化された知識を基にした情報検索システム

○森本 貴之, 近藤 雄裕, 杉田 勝彦, 石川 大介, 池村 匡哉, 藤原 譲

Information Retrieval Systems Based on Organized Knowledge resources

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

Abstract

The global flow of information is being developed at unprecedented speed, and an importance of information retrieval become higher. However, users may not make a good use of huge amount of information by using conventional computers whose major functions are numerical calculation, symbol matching in information retrieval and deduction. Especially, an importance of information retrieval becomes higher. However, it is difficult to search relevant information that is satisfied its purpose from a vast information efficiently.

To solve these problems, a new intelligent method of information retrieval using organized knowledge resources based on semantic relationships is proposed.

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、数値計算やキーワード検索、演繹推論が根底である現在の計算機では、豊富な情報や知識の内容を十分に活用できるとは言難い。そのため、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。

このような情報や知識の内容に関する、より高度な処理を行うためにはそれらの持つ意味を理解することが必要である。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 1)2)3)4)。本研究では、構造化された知識を情報検索に利用する試みについて報告する

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益々益々高まっている。一方、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。Web の Search Engine を例にとると、検索要求として用いる概念(用語)によって検索結果の量が大きく異なる。特に、大量の検索結果の中から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。また、統計的情報を利用して検索の精度を高める研究も行われているが、統計情報は出典の内容を表わすことは本質的にできない。それらに対して本研究は、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」が考えられる。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、本来の形は「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」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。一方、ある特定分野の中から興味深い内容について書かれた文献を見つけるといった検索もよく行われる。このような場合、検索結果の中からの絞り込みを行うことが多いが、効率よく行うためは絞り込みのための指針(情報)が必要である。しかし、元々絞り込みのための情報はないかあるいは明確でないため、情報自体を検索結果の中からえる必要がある。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。

- 知識の特性、特に意味関係の解析
- 知識の体系化(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)
- 各分野の知識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

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えない。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 5)6)。

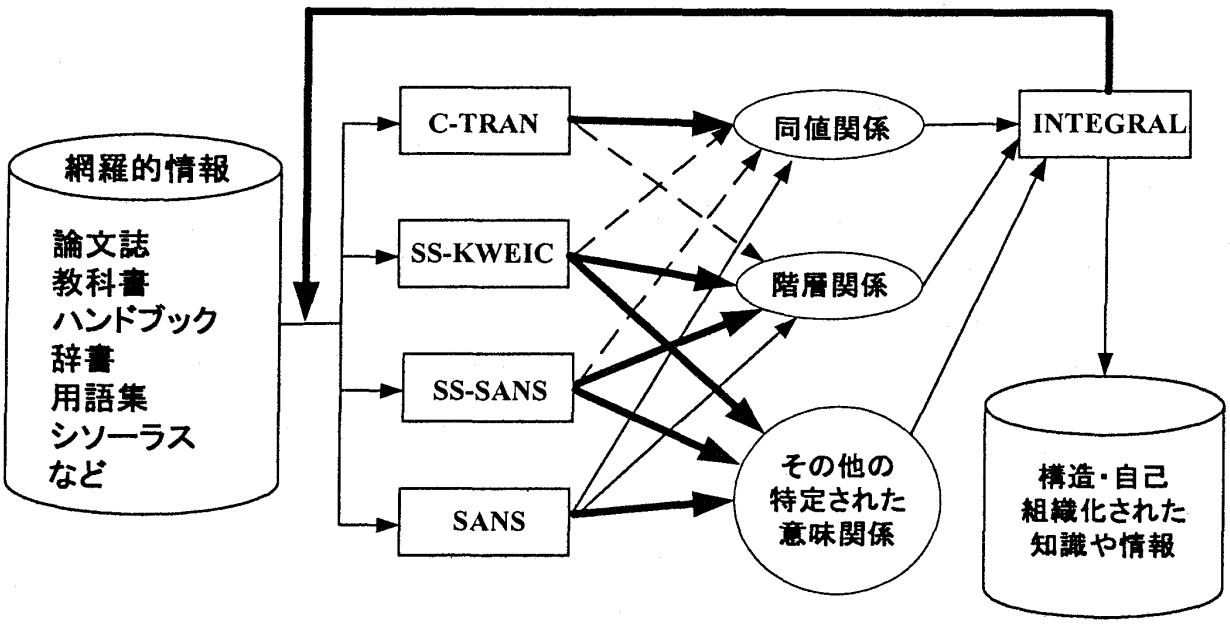

図 1. 知識の自己組織化システム

HBM は概念構造間の多種多様な意味関係を表現することを目的としている。HBM による構造化された知識の例を図 2 に示す。図 2 には関連関係と階層関係がそれぞれ 2 種類の表現があるが、それは抽出法の違いによるものである。関連関係の円で囲まれた部分(「並列」をキーワード)と階層関係の実線矢印は SS-KWEIC 法によるものである。一方、実線の関連関係と破線矢印の階層関係は SS-SANS 法によるものである 5)。特に、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」の組み合わさった内容が含まれていることを示す。

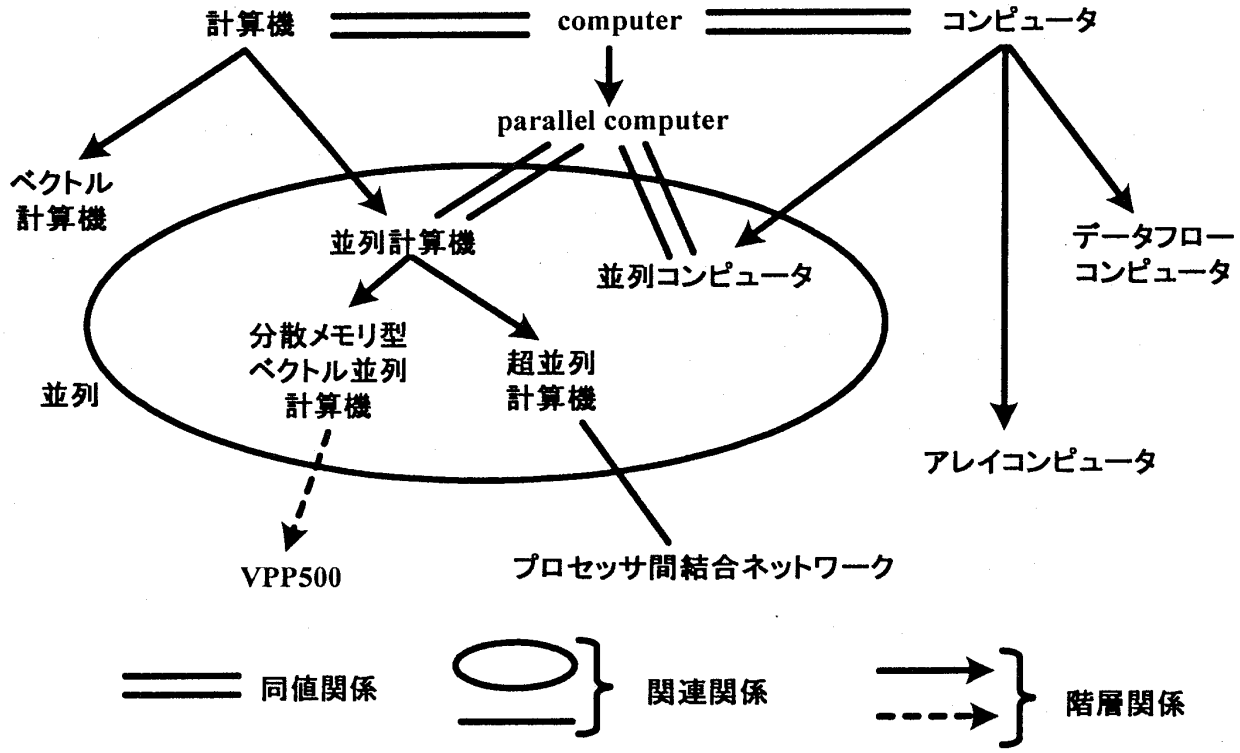

図 2. HBM を用いた構造化の例

4. 情報検索システム

4.1 システム概要

情報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. 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”7) を使用)
3. 意味関係に基づく構造化

検索処理は意味関係のナビゲーションによって行われ、検索要求を満たす出典ならびに関連知識とその出典に関する情報(各用語および一部の意味関係はそれぞれ出典情報を持つ)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“並列コンピュータ”を検索要求とした結果の抜粋を示す。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)の一部を入力データとしたものである。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”は出典のタグ情報を示す。(ただし、この図では四つ以上の出典がある場合は“... ”で省略)

用語と意味関係を用いて構築された概念構造とそのナビゲーションによって本システムは以下の三つの特徴を持つ。

1. 意味関係を基にした出典の内容の反映
2. 再現率の向上
3. 検索の絞り込のための方向性の提示

概念構造は辞書や教科書、論文誌等といった各種出典から抽出された用語および意味関係を基にして構築されているため、通常の全文検索や統計情報では本質的に無理な出典の内容を反映させることができる。また、概念構造のナビゲーションを利用することで再現率が向上するが、これは意味関係に基づくものであるため適合率をそれほど低下させるものではない。さらに、ナビゲーションの際に辿った意味関係を検索絞り込のための情報と

して利用することができる。

コンピュータ ：gakkai-j-0000341470 gakkai-j-0000342371 gakkai-j-0000343309 ...
　→ノートブック型コンピュータ ：gakkai-j-0000342168
　→携帯型コンピュータ ：gakkai-j-0000348016
　→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 gakkai-j-0000344216 gakkai-j-0000345141
　→アナログ計算機 ：gakkai-j-0000343604
　→64 ビット計算機 ：gakkai-j-0000342940
　→ベクトル計算機 ：gakkai-j-0000342091
　→メッシュバス計算機 ：gakkai-j-0000342093
　→モーバイル計算機 ：gakkai-j-0000342157
　→分散共有計算機 ：gakkai-j-0000342728
　→移動型計算機 ：gakkai-j-0000342159
　→SMP 型計算機 ：gakkai-j-0000343419
　→並列計算機 ：gakkai-j-0000340082 gakkai-j-0000340084 gakkai-j-0000340086 ...

図 3. 検索結果の抜粋(検索要求 ： 並列コンピュータ)

4.2 並列化

このシステムでは検索は概念構造間の意味関係をナビゲーションすることによって実現される。従って、各種意味関係による概念構造の繋がりが非常に重要であり、繋がりが少なければ意味関係に注目した効果的な検索結果を導き出すことはできない。また、知識として利用するためには網羅性が非常に重要である。したがって、実用性を考慮すると、大量の用語およびそれらの間の意味関係が取り扱えるシステムでなければならない。そこで、MPI による並列実装を行う。

並列化においは大きく分けて以下の 2 点について考慮する必要がある。

- 概念構造の分散配置
- 分散配置を考慮したナビゲーション

テキストではなく概念構造を使用する本システムでは概念構造は各種意味関係で繋がっており、検索処理は概念構造のナビゲーションによって成し遂げられる。したがって、処理の効率等を考慮すると概念構造は特定の意味関係を基準として分散させることが望ましく、一般的な情報検索をその土台とすると階層関係が標準的と考えられる。 しかし、この基準となる意味関係は分野や使用環境によって異なることも考えられる。図 4 は概念構造の分散の概略を示したものである。Master プロセスは Slave プロセスから送られた局所的な概念構造の分散情報を用いて Slave プロセス間の概念構造の転送を制御することによって、階層関係に基づいた概念構造の分散を行う。

特定の意味関係を基準とした概念構造の分散配置によって他の意味関係で繋がっている概念構造は、実際には空間的に離れたものになることが非常に多い。したがって、これらの空間的に離れた概念構造を跨いだナビゲーションが要求される。そのためのアルゴリズムの概略は以下の通りである。

1. Master プロセスは検索要求を全ての Slave プロセスに broadcast
2. 各 Slave プロセスは個々に検索を行い、検索要求を満たす場合はそこから辿ることのできるすべての概念構造を Master プロセスに転送
3. Master プロセスは概念構造の再構築を行い、その中から階層関係以外の意味関係で繋がっている用語を新しい検索要求として 1. の処理に戻る。

以上の処理を辿ることのできる概念構造がなくなるまで繰り返すことで Master プロセスにユーザの検索要求から辿ることのできる全ての概念構造が集約される。図 5 は“並列計算機”をユーザの検索要求とした例で、Master プロセスは Slave n プロセスから転送された概念構造を元に“コンピュータ”を二度目の検索要求として broadcast する。

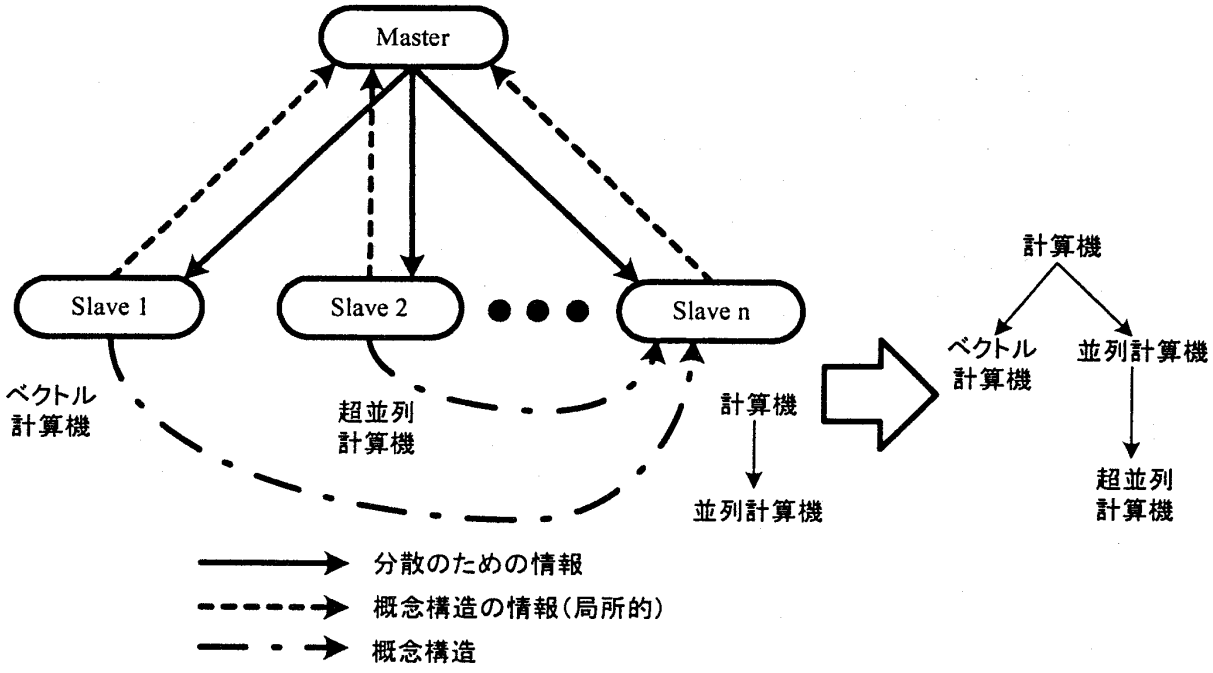

図 4. 構造化の並列化

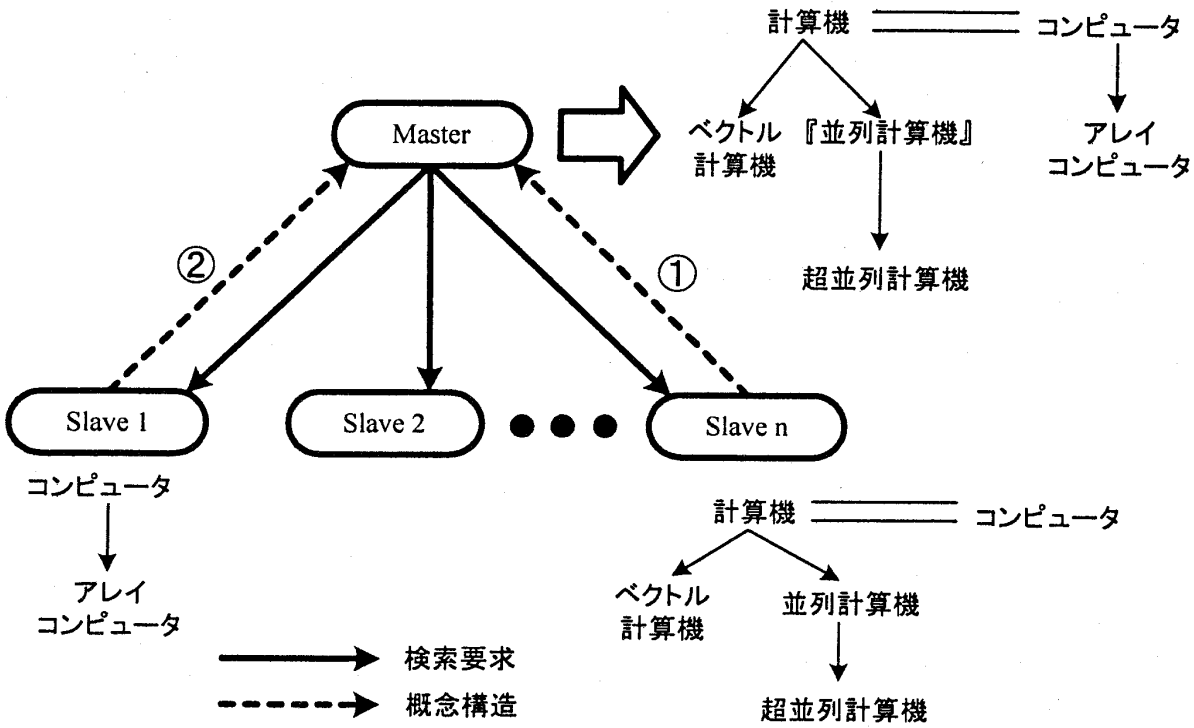

図 5. 検索処理例

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利用として、情報検索システムへの適応に関するものである。

今後は 以下の 2 点について研究を進めていく予定である。

- 現在まだサポートしていない意味関係(因果関係等)の実装
- 並列実装に関する検討

6. 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) 森本貴之, 真栄城哲也, 藤原譲, 情報の構造化 -- 学習・思考機能実現に向けて --, 情報処理学会第 59 回全国大会講演論文集(3), pp75-76, 1999.
4) 森本貴之, 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 61 回(平成 12 年後期)全国大会講演論文集(3), pp111-112, 2000.
5) 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.
6) 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.
7) http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html

---

森本 貴之　神奈川大学 理学部 (〒259-1293 神奈川県平塚市土屋 2946)
近藤 雄裕　神奈川大学大学院 理学研究科 (同上)
杉田 勝彦　同上
石川 大介　同上
池村 匡哉　同上
藤原 譲　独立行政法人 工業所有権総合情報館 (〒100-0013 東京都千代田区霞が関 3-4-3)

Takayuki Morimoto (morimoto@info.kanagawa-u.ac.jp) Faculty of Science, Kanagawa University
Takahiro Kondo (s965521@educ.info.kanagawa-u.ac.jp) Graduate School of Science, Kanagawa University
Katsuhiko Sugita (s965731@educ.info.kanagawa-u.ac.jp) Graduate School of Science, Kanagawa University
Daisuke Ishikawa (dais@goto.info.kanagawa-u.ac.jp) Graduate School of Science, Kanagawa University
Masaya Ikemura (ike@goto.info.kanagawa-u.ac.jp) Graduate School of Science, Kanagawa University
Yuzuru Fujiwara (fujiwara-yuzuru@ncipi.jpo.go.jp) National Center for Industrial Property Information