研究論文

自己組織型情報検索システム：SIR^{*1}

仁 木 和 久

電子技術総合研究所情報科学部認知科学研究室^{*2}

Self-organized Information Retrieval System : SIR

Kazuhisa Niki

Cognitive Science Section, Information Science Division, Electrotechnical Laboratory

A Self-organized Information Retrieval System, called SIR, is being developed to contribute to a Self-organized Information Base Research Project. SIR is especially designed to be a user-friendly information retrieval system that can be used at an early stage without a database that has been specifically designed for particular information retrieval (IR) tasks. For this purpose, SIR must become intelligent by means of self-organization during use. In order to achieve these goals, SIR is designed using a connectionist information retrieval system. In this article, we describe the design philosophy and the configuration of SIR, and some experimental performance measurements.

1. は じ め に

自己組織型情報ベースの基盤技術の研究の一環として、将来型の情報ベースに必要な自己組織化能力と、ユーザにとって使い易い情報検索に必要な情報処理技術に関する研究・開発を行なっている。本論文では、情報検索用の情報の集積が十分でない状態で、素人のユーザでも使い始めることができ、使うことにより情報検索用の情報の集積が行える情報検索システムをその開発目標とした自己組織型情報検索システム SIR (Self-organized Information Retrieval System)¹⁾ について、コネクショニスト情報処理を利用した狙いや、システムの構成法、その試作システムを用いた評価等について報告する。

2. 自己組織型情報検索システム SIR の狙い

従来の論理型情報検索では、質問を形式的に詳しく記述でき、理想的な検索ができる。しかし、そのように使いこなすためには、十分吟味され構造的に蓄積された情報検索用の情報の存在と、さらにユーザがその検索用情報の使い方を十分熟知していることが必要である。比較的自由な検索を可能にする自由キーワード方式の検索を実現するベクトル型の情報検索システムにおいてすら、

検索の指標となるベクトル表現を与える必要があることから、事前の情報検索用の情報の構築が必要であるという事情は変わらない²⁾。

しかし、この情報検索用の情報の蓄積が一般に問題である。情報の溢れている現代社会では、原情報があってもその情報の整理が行われていない。あるいは、検索用情報を管理するサーチャのような専門家がおらず、当事者が整理をする必要がある。このように考えると、検索用の情報が不十分な状況でも、溢れる情報を利用／検索できる情報検索システムが望まれることがわかる。

「検索用の情報の蓄積が不十分な状態からでも使える情報検索システム」という、一見パラドックスを乗り越えられるシステムとは、どのようなものであろうか？このような情報検索システムが備えているべき機能や特徴は、次のように列挙できる。

- 1) 情報検索用の情報の蓄積が十分でなくても、情報検索システムの構成ができること
- 2) 情報検索システムの素人でも直感的に使えるような、使い易いシステムであること
- 3) 情報検索システムが、学習あるいは自己組織化機能を持つこと

もちろん、「情報検索用の情報の蓄積が完全に無い」状態から始めることはできない。集積された情報／データの中から機械的に取り出せる情報に、ある程度は使え

^{*1} 1994年4月20日受付

^{*2} 〒305 茨城県つくば市梅園 1-1-4

る情報蓄積があることを仮定している。また、情報検索の素人といっても、計算機の使い方や図形からの意味の解釈手法などの知識や常識を持ったユーザを想定することは必要である。そして、この情報検索システムが実際に使われる状況で、ユーザと自己組織型情報検索システムとの相互作用によって、誤りは修正され、有用な情報構造が蓄積されることが必要である。

本論文で紹介する自己組織型情報検索システム SIR では、技術マニュアルや文献、技術情報等の分野で使うことを想定し、原情報は豊富にあるが情報検索のための情報の集積が不十分な状態においても情報検索を可能にし、使っている間に情報検索のための情報の集積が行えるシステムの実現を目標としている。このように、「情報の自己組織化」の初期段階をサポートする情報検索システムとして SIR を位置付ける^{1,3)}。

3. コネクショニスト自己組織型情報検索システム

自己組織型情報検索システム SIR に要求される機能と特徴を満たすためには、次の3つの理由でコネクショニスト情報処理技術の適用が適している。

- 1) 関係記述を論理的に記述できない情報の表現手法として、関係記述のコネクショニストネット表現が自然である。また、関係記述のネットワーク表現は原情報の統計的処理等により機械的に作成できる。

- 2) 視覚的な表現を用いた、人間との対話の手段としてこのネットワーク表現が有効に使える。
- 3) 確立された情報の蓄積の無い状態での学習/自己組織化機能に優れている。

このようなことから、SIR をコネクショニスト情報処理システムとして設計した。特に、本システムを情報検索システムの素人が容易に使いこなせるシステムとするため、関連情報のネットワーク表現をグラフィック表示するユーザフレンドリなインターフェースを開発した。さらに、本システムが適用できる対象は広いと考えられるため、特定の情報検索システムを作るのではなく、多種の情報検索システムの生成を行えるメタシステムとして本システムを設計した。また、一般に多量の計算量を必要とするため実用性が低いと考えられているコネクショニスト情報処理ではあるが、適切な設計を行えば十分実用に供することが可能であることを示すことも本システム開発の目的の1つである。

自己組織型情報検索システム SIR の環境の全体像は、Fig. 1 に示すように、検索情報の前処理部（キーワード抽出および統計処理部）、SIR-Gen と呼ぶ情報検索システムの生成部、および関係情報をノード間の結合情報として生成・管理するためのユーティリティからなる。本章では、検索情報の前処理と、情報検索で使われるコネクショニスト情報処理の活性化伝播と学習について述べる。SIR の根幹をなす SIR-Gen については、実現手法

Fig. 1 Overall configuration of Self-organized Information Retrieval System: SIR
This figure also shows the construction flows of the SIR.

を含めて次章で詳細に述べる。

3.1 ネットワーク表現

文献の検索を想定すると、情報検索とは、文献に関するキーワードを入力して関連文献を見つけることである。これに対応したネットワーク表現を Fig. 2 に示す。キーワードノードからなるキーワード層と文献名ノードからなるドキュメント層の2層からなり、その間は何らかの関連性を表すリンクで結合されている。このようなネットワーク表現においては、キーワード入力により、活性伝播を用いて関連する文献名を探すことができる。

このネットワーク表現の優れた点は、情報検索用の明確な知識が無くても、次節に示すような手法によりネットワーク表現の機械的生成が可能であることと、ネットワーク間の関連性を用いた活性伝播による情報検索状況を見せることにより「情報検索の可視化」が可能で、優れたユーザインタフェースを提供できる点である。

3.2 キーワード抽出およびネットワークの生成

文献全文を対象に各単語の原形をキーワードとし、文献 i 毎のキーワード j の出現頻度 F_{ji} を計算し、収集する。全文で高頻度に出現するキーワードや、前置詞、定冠詞、および各文献において出現頻度の低いキーワード等は、情報量が低いため除く。この出現頻度情報等を用いて、キーワード層とドキュメント層の2層相互結合ネットワークの初期結合荷重を次のように計算する。文献 i からキーワード j への結合荷重は、次式で与える^{1,4)}。

$$W_{ji} = \frac{F_{ji} \log\left(\frac{N}{DF_j}\right)}{\sqrt{\sum_{k=1}^N \left(F_{ki} \log\left(\frac{N}{DF_k}\right)\right)^2}} \quad (1)$$

ここで DF_j は、キーワード j が存在する文献の総数であり、 N は文献の総数である。実際には、さらに文献 i からキーワードへ出ていく荷重を総和が1になるように正規化する。(1)式に出現する関数は、自動インデクシングにも使われる関数²⁾でもあり、妥当性があり、試みた他の多くの手法の中でも概ね良い結果を示した。

また、キーワード j から文献 i への結合荷重は、次のように与えられる。

$$W_{ij} = \frac{1}{DF_j}$$

このように計算した SIR 用のネット結合荷重を用いて、ネットワーク生成部 (SIR-Gen) が読み込む初期設定ファイル (Fig. 1 のコネクショニストネットデータ) を生成する。

3.3 活性伝播

ノード i の活性値 a_i は、周囲のノード j からの活性値 a_j を空間加算し、 $[-1, 1]$ のシグモイド関数 f で出力変換した次式で計算する。

$$a_i(t+1) = f\left(\sum_{j=1}^n W_{ij} a_j(t) + a_i(t)\right) \quad (2)$$

キーワードが入力されると対応したキーワードノードの活性値を1に固定して、(2)式に従い関連した文献名ノードを活性化させる。さらに、活性化された文献ノードに関連したキーワードノードを再帰的に活性化させ、この再帰的活性伝播の計算を繰り返す。結合荷重は正規化された小さな正值であるため、この拡散的伝播の収束は非常に早く、5, 6回から多くても10回程度の繰り返し計算で収束する。活性領域の全てのノードの活性値のステップ毎の変化がある閾値 (0.002) より小さくなったときに、収束とみなし計算を打ち切る。

(2)式は、本システムのために経験的に採用された関数であるが、収束した状態で活性値 a_i は $\sum_{j=1}^n W_{ij} a_j$ を変数とするシグモイドに似た関数関係を持つ。また、記憶作用を持つが、その様な機能は現在のところ利用していない。

閾値 (th1; 通常 0.005 に設定) より大きい活性値を持つノード集合を「活性領域」(注目する情報の存在領域; Fig. 2 参照) として切り出す事が出来る。この活性領域は、入力キーワードに関連を持つキーワードと文献名の集合であり、現在注目すべき事象からなるワーキン

Fig. 2 Connectionist Network Representation for Information Retrieval and Visualization of IR Process.

ゲメモリ（作業領域）となっている。この活性領域の活性化の様子とノード間の結合関係とをユーザに直接見せることにより「情報検索の可視化」を実現できる。これについては、4.2 で再び述べる。

3.4 学 習

学習は、入力キーワードに対して目的の文献が正しく想起され易くなるように結合荷重を修正するものであり、正しい文献検索に成功したときユーザの評価により起動される。機械的に生成された初期設定ファイルの情報は修正が必要であり、また情報検索の個人適応の為に必要な機能である。

学習アルゴリズムとしては、“正しく選ばれた文献の活性化に寄与したキーワードとの結合を強化し、更にもそのようなキーワードの活性化に寄与した文献との結合を寄与度に応じて強化する”「因果探索学習」を採用した。アルゴリズムの概要を次に示す。

ステップ 1: 選択された文献名ノード（学習活性値 = 1）に活性値を送っている「活性領域のキーワードノード」との間の結合を“ ϵ * キーワードの活性値 * 文献名ノードの学習活性値”だけ強化する。

ステップ 2: 選択された文献名ノードから「活性領域のキーワードノード（複数）」に、学習活性値をその間の結合値を掛けて逆伝播させる。キーワードノードと、そのキーワードノードに活性値を送っている文献名ノードとの間の結合を“ ϵ * 文献名ノードの活性値 * キーワードノードの学習活性値”だけ強化する。

ステップ 3: ステップ 2 と同様に、各キーワードノードから学習活性値を文献名ノードに逆伝播し、文献名ノードとキーワードノードとの間の結合を“ ϵ * キーワードの活性値 * 文献名ノードの学習活性値”だけ強化する。

ステップ 4: この逆伝播に基づく学習ステップ 2, 3 を繰り返す。

ステップ 5: 探索を打ち切った後、結合荷重を正規化する。

ループを余り深く辿っても意味が曖昧になるので、因果探索は 2 回のループで打ち切っている。また、学習の係数 ϵ は 0.1 とした。学習した結合は、情報検索の終了後、次回に使う初期設定ファイルとして保存できる。

4. SIR-Gen の構成

SIR-Gen は、初期設定ファイルを読み込み、情報検索用のネットを構成し、一連のコネクショニスト情報処理を行う「コネクショニストネット部」と、ネットワーク

を情報検索用に使うための「制御部」、ユーザのキーワード入力の際の誤字を発見し訂正する「誤字訂正部」¹⁾、さらにはネットワークの状況や文献の内容提示等を行う「インタフェース部」とからなる (Fig. 1 参照)。本章では、オブジェクト指向プログラムによる実装手法とグラフィックインタフェースについて紹介する。

4.1 コネクショニストネット部のオブジェクト指向による実現

SIR-Gen 自身のネットワーク表現モデルとしては相互結合の全結合モデルを与え、ノード間の結合関係と結合荷重を記述した初期設定ファイルの情報の従い、層構造等の任意のネットワーク構造を作り出すように設計した。このような柔軟性は、任意の構造を持つより複雑なネットワーク表現を可能にするだけでなく、本システムのような粗な結合を持つコネクショニスト情報処理システムの効率的な実現を可能にする。

コネクショニストネット部は、二つのクラス Node と Network から成る C++ で書かれたオブジェクト指向システムとしてプログラムした。各ノードは、各種の属性情報と共に結合先のノードに関する情報を持つ。また、コネクショニスト情報処理に関する活性伝播や学習アルゴリズム等のメソッドを Node クラスと Network クラスで定義する。このオブジェクト指向によるインプリメンテーションにより、コネクショニスト情報処理が実用的な速度で、効率良く計算できる。つまり、オブジェクト指向の記述によりリンクを直接辿るため、粗な結合を持つ情報検索用のネットワーク表現の効率が良い。また、活性値を持ったノードだけからの伝播だけを考慮して、活性領域（とその隣接領域内へ）の局所計算で良いため、計算量を大幅に減少させることができる。この計算の局所性は、扱う文献量が多くなった時に、さらに威力を発揮する。

次に述べるネットワークの可視化の実現を含め、本システムのようなメタシステムに於いて「書きやすく」「良いコードを生成できる」手法としてオブジェクト指向のプログラミングを採用した。

4.2 グラフィックインタフェース部

Fig. 2 のようなネットワークの表示により「情報検索の可視化」が可能になることを述べた。「情報検索の可視化」では、ネットワークの状況（活動状態の大きさや結合荷重の大きさ、ノードの結合関係等）を見せるだけでなく、表示ノードのクリックによるキーワードの入力や、文献の内容の参照等もできる事が重要である。本システムでは、この「情報検索の可視化」をグラフィックインタフェースを用いて実現した。Fig. 3, 4 にそのグ

Fig. 3 A SIR of Unix User Command Manual (before learning)
Input keywords, “delete” and “file”, activate many document nodes and keyword nodes, but the target document “rm” is not the winner.

Fig. 4 A SIR of Unix User Command Manual (after learning)
Input keywords, “delete” and “file”, activate the target document “rm” at the top.

グラフィックインタフェースの主要な部分の様子が見られる。(5.1で紹介するように、Fig. 4はキーワード delete と file に対してコマンド rm を正しく連想できるように学習した後であり、Fig. 3は学習以前の様子を示す)。

画面左下の主ウィンドーは、キーワードのキーボード入力および様々のサブウィンドーの表示制御を行う。画面左上は、活性領域のネットワークを表示するウィンドーで、本システムで最も重要な役割を担っている。表示した2層の上段はキーワード層で、下段はドキュメント(この場合、Unix コマンド)ノード層である。表示されている円は、各々のノードの活性値の大きさを表現している。活性領域に属するノードを活性値の大きさに従い、中央から周辺に配列しているため、中央部に重要なノードが配列される。Fig. 4のケースでは、delete と file のキーワード入力に対し rm コマンドが第一候補として想起されている。

ユーザは、キーワードを入力し、そこから連想される活性領域のグラフィック画面を見て、関連するキーワードを再認しながら、グラフィック上のノードを“クリック”することにより新しいキーワードを入力することができる。また、文献ノードを“クリック”することにより、その文献を見ることが出来る。Fig. 4右下のウィンドーは、rm ノードの選択により、rm の文献(この場合マニュアル)を表示している。

上下ノード間の線は、コマンド(文献)ノードから出てキーワードノードに入る結合を表示しており、これもノード間の関連性をユーザに視覚的に伝える。線の太さは、結合荷重に比例した太さで表示されている。右上のウィンドーは、コマンド(文献)層の活性ノードの活性値等の情報を表示している。

なお、グラフィックスプログラムは X11 のアテナウィジェットを使って書かれているため、十分移植性がある。

5. SIR の適用例および評価

2, 3章で示した SIR の狙う機能と特徴を実証するため、SIR の試作システムを作り3つの情報検索システムを構築し、その使用結果に基づいて簡単な評価を行った。

5.1 情報検索システムの生成

Fig. 1のキーワードの自動収集プログラムの数行を手直しするだけで、次のやや性格の異なる3つの情報検索システムを生成した。

5.1.1 Unix コマンドマニュアルの検索システム

Unix (SunOS 4.1.3) の80のユーザコマンドのマニ

ュアルファイルに出現する単語 1672 個をキーワードとし、文献(この場合、コマンドマニュアル)名を探す検索システムを構築した。Unix ユーザコマンドの使い方の解説は、通常/usr/man 等にファイルとして存在し、手軽に入手できる。1つのコマンドに対し、通常 A4 で1, 2枚の分量であるが、プロの手により詳しく書かれたものから、数行のものまであり、実際の世の中にある文献の性格も持ち合わせている。また、コマンド rm と rmdir とが同じ文章で紹介されているなど、文献検索を混乱させる要因も色々を持ち合わせている。

既に 4.2 で紹介で示した Fig. 3, 4が、その例である。Fig. 3のように、その初期では、キーワード delete と file に対して、コマンド rm と rmdir を同等に連想していたものが、通常のユーザのコマンドの使い分けを反映し学習により、Fig. 4のように、rm を連想するようになる。

このような使用と評価による自己組織化能力は、SIR のようなシステムには本質的に不可欠な能力である。何故なら、機械的に生成した情報検索用の情報は、本システムのようにユーザに直感的に分かり易く、使い易いシステムに於いては十分に役立つものの、常に正しい情報である保証は無い。使用と評価とにより正しく検索が出来るように修正が必要である。

5.1.2 JICST 研究速報の情報検索

JICST の公開している情報関連の技術報告アブストラクト集(99 文献)から検索システム(Fig. 5)を生成した。100 字から多くて 200 字、通常 120 字程の文献のアブストラクトであるが、著者が各自の視点で書いたものなので、用語や書き方も統一が取れていない。

SIR では、1つのキーワードを入力すると関連したキーワードが連想されるため、次に入力すべきキーワードを連想リスト(あるいは表示ノード)から再認/選択すればよい。さらに検索状況が分かるなど、ユーザの感覚として論理型検索より使い易いとの報告があった。

5.1.3 Neuroprose 文献検索

神経回路の研究文献の計算機アーカイブ Neuroprose に付属したインデックスファイルを用いた文献検索に SIR を適用した。JICST のアブストラクトよりも更に文章の量が少なく文献当たり 1, 2 行で、10 文字程のキーワードしか含まない。162 文献で、1206 種のキーワードが含まれていた。

このような状態でも、JICST と同様、論理型検索の唐突さがないこと、さらに連想機能が働き、キーワードの少なさを補うことが分かった。

5.2 SIR の情報検索の特徴

SIR の活性伝搬を用いた情報（文献）検索の特徴を、学術用語でまとめると、次のようになる。SIR の文献検索は、原理的に連想機能により関連したノードを広く辿る（検索する）機構となっており、高再現率検索方式²⁾を実現している。また、厳密な照合を必要としない最適照合方式になっている。SIR では、文献に出現する単語は全てキーワードとなりうるので、自由キーワード方式を実現している。

前節で SIR の情報検索が、従来の情報検索と多少異なっていることを述べた。この事情をもう少し詳しく述べる。Unix のマニュアルの検索を、キーワードが文献に含まれるか否かで全文について検索すると、キーワード file に対し 58 の文獻候補リストが得られるが、キーワード delete に対しては僅か 6 個の文獻しか得られない。検索結果が多過ぎたり少な過ぎたり極端になることが多い。また、重要さの順位が付いていないため、ユーザは戸惑う。これに対し、SIR では Fig. 3, 4, 5 に見ることのできるように関連の有るキーワードと文獻名が順位を付けて連想される。ユーザはそこから注目したいキーワードを「思い出し」て「選択」すればよい。一般的には活性値が高いほど「相応しい」あるいは「適合」した文獻／キーワードであるという判断の目処にも

なる。しかし、SIR を使ってユーザが行っている検索／探索の内容は、実際には次のように多様である。

- 1) 基本的には、入力キーワード群に直接関係する文獻の OR 検索
- 2) ノード活性値を高めめの閾値で切り取る AND 検索
- 3) ノード間の関連性を辿って、間接的な関連文獻まで想起する連想検索
- 4) 活性値の大きさを目安に、検索に利用できる適切さの順位付けをも判断の材料に用いる検索
- 5) キーワードと文獻との結合関係とその強さもグラフィック表示から読み取り、意味的な関係を使った検索

このように、SIR は、「情報検索の可視化」により、人間にとって分かり易く、操作しやすい情報検索の状況を作りだし、様々なモードでの探索の機会をユーザに与えている。

5.3 文獻の増大に伴うシステムの規模の問題

SIR が文獻検索の対象数の増大に対しても、十分対応できる実用的なシステムに成りうることを示す。これは、一般的にコネクショニスト情報処理に対して、文獻等の処理対象数の増大に伴いネットワーク規模が増大し、計算量も莫大なものになり、実用に耐えなくなる可能性が指摘されている為である。

Fig. 5 A SIR of JICST Technical Report Database.

Table 1 Scale measurements of SIR

文 献 数	10	20	33	50	81	89
キーワード数	534	988	1110	1594	2164	2257
リンク数	1560	4100	4800	7600	12600	13910
最大リンク数	10680	39520	73260	159400	350568	401746
リンク数/文献数	156.00	205.00	145.45	152.00	155.56	156.29
キーワード数/文献数	53.40	49.40	33.64	31.88	26.72	25.36
リンク数/最大リンク数	0.15	0.10	0.07	0.05	0.04	0.03

Fig. 6 Relationship between numbers of documents and numbers of links/keywords for a SIR of Unix manual pages. The size of the SIR is measured by numbers of links/keywords per a document.

SIR に於いて、文献の増加とコネクショニストネットワーク規模の増大の関係を、Unix コマンド検索の例で示したのが Table 1 と Fig. 6 である。図と表に表れる「リンク数」は実際に SIR で生成されたリンク数であり、「最大リンク数」は 2 層間の完全結合マトリックスとしたときに生ずるリンク数である。

この表や図から分かるように、文献の数の増大に対し、文献当たりのリンク数やキーワード数は、フラットあるいはやや減少気味であり、システム規模の爆発的な増大はない。ここに示された性質は、文献が持つ情報の性質を反映したものであるが、この性質をコネクショニスト情報処理に持ち込むためには、オブジェクト指向のネットワークの表現やシステムの実装が重要であった。活性伝播の計算も、活性領域内での局所計算で済むため、やはり計算量の増大を抑さえるのに貢献している。

6. お わ り に

試作システムを実際に構成して、自己組織型情報検索

システム SIR の狙いである「情報検索用の情報が不十分な状態での情報検索システムの構築、直感的に使い易い情報検索、および学習・自己組織化機能」が、実用性を考慮した形で実現できる事を示した。このような特徴を持つ SIR は、自己組織型情報検索システムの初期の自己組織化を実現する為に不可欠なシステムであると考えている。

今後の SIR システムの研究開発に於いて、次の 2 つの重要な方向がある。コネクショニスト情報処理の特徴を一層伸ばす方向と、論理型の情報検索システムと SIR との狭間を埋める為の研究の方向である。特に後者は、将来の自己組織型情報検索システムを構成する上での重要な課題であり、SIR のようなシステムはその土台としても必要である。

謝 辞 本研究は、科学技術振興調整費総合研究「創造的研究支援のための自己組織型情報ベースの研究」の一環として行われた。本研究の開始時に尽力し、研究の機会を与えてくださった諏訪基情報科学部長、本研究プロジェクトに参加し日常協力頂いた認知科学研究室の皆さん、および本システムの開発、特にプログラミングに関して貢献した東京理科大の実習生であった渡辺泰之君と杉崎正之君に対して感謝します。

参 考 文 献

- 1) 仁木和久, 渡辺泰之, 杉崎正之 (1993): コネクショニスト情報検索システムの開発, 電子情報通信学会ニューロコンピュータ研究会, NC 92-160, pp. 247-254
- 2) 伊藤哲郎 (1986): 情報検索, 昭晃堂, 東京
- 3) 仁木和久, 渡辺泰之, 杉崎正之 (1993): コネクショニスト & 論理の統合型情報検索システム: CLIP, 電子情報通信学会 1993 年春季大会, 分冊 6, pp. 55
- 4) Rose, D., Belew, R. (1991): A connectionist and symbolic hybrid for improving legal research, Int. J. Man-Machine Studies, Vol. 35, pp. 1-33