社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－DBS－130 (16)
2003－ＦＩ－ 71 (16)
2003／5／23

キーワードマイニングによる文献の組織化と可視化機能の統合

－TermLinkerシステムの開発－

土橋 喜[†] 山内 平行[††] 立花 隆輝[†††]

[†]愛知大学現代中国学部
[††]前橋工科大学
[†††]日本IBM東京基礎研究所

個人が大量の文献情報を手軽に扱えるためには，情報の検索と収集，情報組織化，情報可視化，情報抽出などの要素技術を統合したシステムの研究開発が重要である．そこで主にインターネットから収集した文献を対象に，研究者の個人的な再利用を前提とした文献整理と内容検索機能を備え，情報可視化機能を統合した発想支援システムTermLinkerを提案する．このシステムは情報検索の非専門家にも手軽に使えるインタフェースを目標とし，マイニング機能と可視化機能を統合したものである．

Integration of Document Organization and Visual Function by Keyword Mining

－Implementation of TermLinker System－

Konomu Dobashi[†] Hiroyuki Yamauchi[††] Ryuki Tachibana[†††]

[†]Faculty of Modern Chinese Studies, Aichi University
[††]Maebashi Institute of Technology
[†††]IBM Research, Tokyo Research Laboratory

We have developed a system that integrated automatic document organization and visual functions. Our system named TermLinker which has a function to organize textual collections or documents with HTML. The TermLinker generate conceptual networks by eliciting the terms from textual collections and maps conceptual networks to visualize between keywords in two dimensional space automatically. The TermLinker proposes a method to elicit and visualize hidden relationships integrated with text mining from HTML documents and functions of the information visualization.

1．はじめに

個人が大量の文献情報を手軽に扱えるためには，情報の検索と収集，情報組織化，情報可視化，情報抽出などの要素技術を統合したシステムの研究開発が重要である．そこで主にインターネットから収集した文献を対象に，研究者の個人的な再利用を前提とした文献整理と内容検索機能を備え，これらに情報可視化機能を統合した発想支援システムTermLinkerを提案する（図1と図5を参照）．このシステムは情報検索の非専門家にも手軽に使えるインタフェースを目標とし，マイニング機能と可視化機能を統合したものである[5)6)]．

近年，電子化された研究論文などの文献が大量にインターネットに公開されている．検索エンジンを活用すれば，研究者にとっても有益な文献が手軽に手に入るようになった．中には研究成果を発表

した論文が電子ジャーナルの形式で公開されているものも数多くあり，このような電子的な文献の提供方法は今後も増加するであろう．

ところで収集した文献が比較的少ないあいだは，自分の記憶にたよりながら一つ一つの文献をブラウザなどに表示して手軽に読むことができる．ダウンロードしたときに自分でわかりやすいファイル名を付けておけば，混乱することなく目的の文献を見出すこともできる．しかし重要なキーワードがどの文献のどのあたりに現れていたかということになると，人間の記憶の限界を感じる場合が多くなり，いわゆる information overload の状態が起きやすくなる．

このような場合に収集した文献のタイトルや著者名が手軽に一覧できたり，キーワードがマウスのクリックだけ一覧できたり，それらの用語から元の文献が即座に検索できるようなシステムがあれば，収集した文献の整理や内容の検索などをより効率的に行うことができる．

インターネットの発展は，個人が大量の文献情報に自由にアクセスすることを可能にしてきたが，その情報量があまりにも多すぎるため，それらの文献情報を有効に利用するためは，新たな技術開発を必要としている．インターネットにおける情報の氾濫や大容量記憶装置の性能向上は，文献整理や内容検索の機能を持つシステムが，今後は個人レベルにおいて重要になっていることを示している．

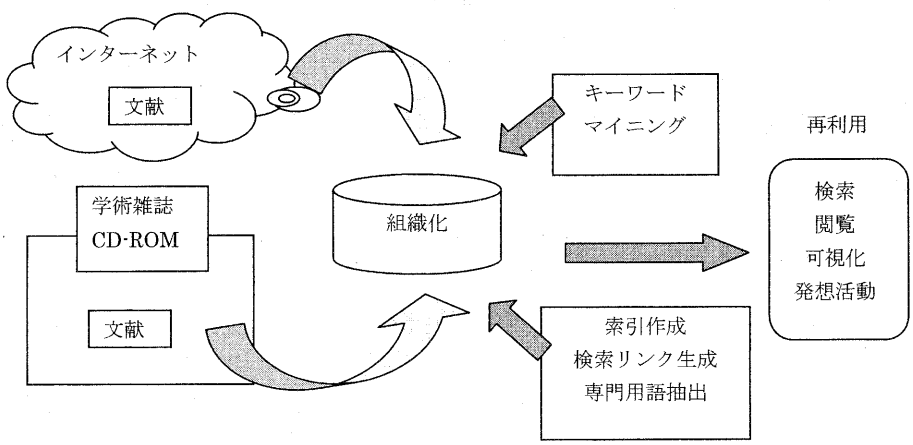

図 1　システムの位置づけ

2. システムのアプローチ

これまでに実用化された情報検索システムをはじめ WWW の検索エンジンなどは，単語を基本としている．しかしその単語が表現している概念が，元の文献の中で果たしている重要性や概念関係をわかりやすく表現できないところに大きな問題がある．この問題は利用者が検索のために用いる適切な単語の選択を困難にしているだけでなく，検索の際に質問として入力する単語と，検索の対象となる文献中の単語の不一致を引き起こし，必要な文献の検索ができにくいという重要な問題となっている[7]．この問題は個人的利用を目的に文献を収集した場合にも起こりえる．本論文では情報検索におけるこのような根本的な問題に対して，主にユーザの個人レベルで対策可能なシステム案を提案する．開発したシステムは大きく分けて３つのサブシステムを統合して構成されている．

2. 1. キーワードによる文献の組織化機能

キーワードを適切に抽出することができれば，それらと HTML のタグを使って文献内を検索するリンクを生成し，Netscape などのブラウザで閲覧できる検索機能を手軽に実現することができる．そこで収集した文献の内容を検索するためにキーワードを抽出し，HTML でリンクを張り巡らす組織化機能を開発した．

最近のインターネットにおける goo や Google などの検索エンジンでは，Web ページ内で使われて

いるキーワードから検索する全文検索機能が一般的に使われている．しかし各ページのタイトルについては，Web ページの作成者が常に適切なタイトルを付けるとは限らないため，抽出されたタイトルが意味不明のこともたびたびあり，このような点からもキーワードの重要性は高い．

ここでは科学技術関係の文献に近いものを想定しており，そこで専門用語と出現頻度の高い用語すなわち高頻度語および複合語の 3 つをキーワードとして抽出している（図 2）．

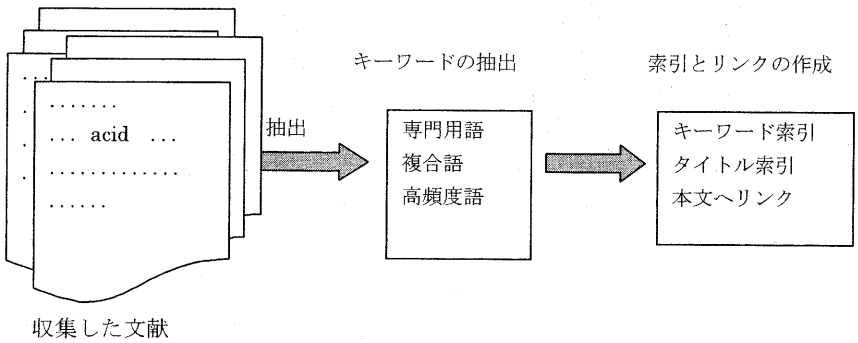

図 2　収集した文献のキーワードマイニングによる組織化

さまざまな分野において，分野に固有の用語が存在するが，それが専門用語であり，その分野の特徴を表している．文献に専門用語が数多く出てくる場合には，その文献の特徴を表していると考えられ，専門用語をキーワードとみなして取り出すことを考える．今回は地球環境問題の文献を対象としたテキストの専門用語辞書を構築し，その辞書を使って抽出している．

また文章上で複合語にすると意味内容が豊かになり，人間の発想を刺激する効果を高めることができる．専門用語では複合語が多数を占めるといわれる[4]．そのためここでは複合語もキーワードとして抽出する方法を取り入れている．複合語を抽出するときは，まず文献の先頭から 2 単語または 3 単語を読み込み，次に 1 語ずつずらしてパターンマッチングを行い，読み込んだ単語と一致する句が文献内に出現しているかどうか調べ，もし出現している場合は頻度をカウントして出現頻度が 2 回以上のものを複合語として取り出す．

さらに同じ文献の中で何度も繰り返し使われる高頻度語は，それなりに著者の重要な概念を表していると考えられ，それは重要な部分とみなすべきであり，キーワードとして抽出する必要がある．ここでは単純に出現頻度の高いものからストップワードを除いたものを抽出する．

2. 2. リンクファイルによる知識ベース化

キーワードマイニングによって抽出されたキーワードを使い，収集した文献の内容を検索する仕組みを作成した．検索といってもキーワードを入力して検索するような従来の多くの検索システムで行われているものではなく，索引や文献中のリンクをクリックしたら次の内容が見られるというハイパーテキストのリンク機能を基本にするものである．これらのリンク機能による検索は，操作がシンプルではあるが，目的とする情報に関連した付随情報が得られやすいなどの特徴があり，利用者の要求があいまいな表現やまとまりのない表現でしか言語化できないような状態のときや，具体的な言葉で明確に言語化できる状態のときにも効果がある．　文献を収集した段階では個々に独立して存在しており，そのままでは文献内部に潜む相互の関連性は把握しにくい．そこでそれらの文献に共出現するキーワードとリンクによって再構成し，検索機能を備えてハイパーテキスト化した．すなわち検索機能を持たない無秩序な文献集合に対して，ファイル名やタイトルを一見したところでは内部の関連性が見いだしにくいそれぞれの文献内部に関連性を作り出し，リンクによる検索機能を備えることによって，知識ベースとして再構築する機能を持たせた（図 3）．ユーザはここで述べた組織化機能とリンクによって，情報検索における試行錯誤的な発想を繰り返しながら，抽出したキーワ

ードから関連文献した文献の内容を読むことができる．なお検索手段としてキーワード以外にタイトル索引を作成した．収集した Web ページの<title></title>に記述がある場合はそのデータを取り出し，それらと文献をリンクさせタイトル索引から文献を検索できるようにした．

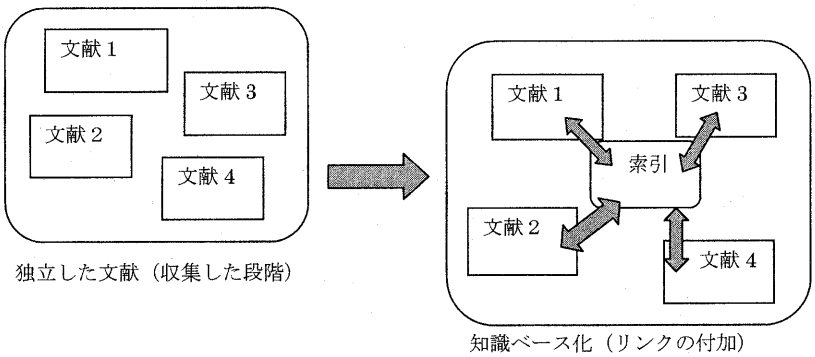

図 3　索引作成と文献の知識ベース化

3. 統合管理機能

統合管理機能はこれまでに述べた組織化機能とこの後で述べる可視化機能を統合するものとして設計されており(図 4)，次のような機能を持っている．

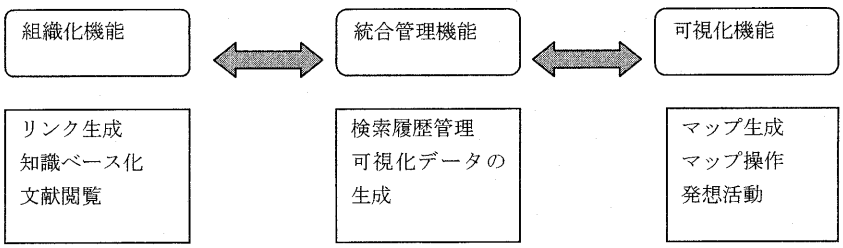

図 4　統合管理機能の位置づけ

3. 1. 可視化データの生成

httpd のアクセスログからユーザが検索した文献のファイル名を取り出し，そのファイル名をもとに収集した文献から可視化に必要なデータを生成する．例えば科学技術関係の文献では，ほとんどのセンテンスの中に，文脈を構成する上で重要な用語が含まれており，それらの概念関係は著者の問題意識を反映したものである．中でも専門用語や出現頻度の高い用語が複数同じセンテンス上に現れる場合は，それらのキーワードに潜む重要な概念関係に言及していると解釈される．主に名詞や動詞から構成されるキーワードがこれらに該当するが，このような重要なキーワード間に示される関係は，従来の意味ネットワークなどに代表される関係とは若干異なることから，ここでは概念ネットワークと呼ぶことにしている．

これらのキーワードをセンテンスごとに取り出すことができれば，ふたつずつ用語をペアにして 2 項関係を構成する組み合わせを生成できる．そしてこれらに共通の用語をつなぎ合わせていくと，文献の中に述べられたキーワードを連結した概念ネットワークとして描画できる．このことを地球環境問題の文献から取り出したひとつのセンテンスを具体的な例に取り上げれば次のように説明できる．

例えば「Both sulfur and nitrogen emission cause acid rain.」という文から，システムに用意した

専門用語辞書または複合語抽出あるいは出現頻度によって sulfur，nitrogen，acid rain という 3 つのキーワードを取り出す．次に sulfur - nitrogen，sulfur - acid rain，nitrogen - acid rain というようにキーワードをペアにした組み合わせを作る．さらにこれらに共通のキーワードを連結して概念ネットワークを生成し，画面上に描画して可視化する．

これを単一の文献で行なえば，その文献の著者の主題を概念ネットワークの形式で描くことになる．また複数の文献に対して行なえば，複数の文献に述べられた主題を合成して描画することになる．単一文献の場合も複数文献の場合も，文献に述べられた概念関係を表現する用語の使われ方によって，概念ネットワークによって描かれる構造図は，それぞれ異なるものが描画されることになる．またキーワードで該当するセンテンスを抽出して描画すれば，文献をまたがった概念ネットワークが描かれることにもなる．なおひとつのセンテンスからキーワードが 1 個だけ抽出される場合は，連結を行わずそのまま単独で画面上に描画している．

3. 2. 類似文献の提案

類似文献の提案機能は，可視化機能で概念ネットワークのマップを生成する場合に，文献の選定を支援するために使われる．文献の類似度は単純にキーワードが共出現した数によって求めているので，類似度の高い文献には共出現するキーワードが相対的に多くなる．類似したキーワードが共出現する文献では同じような主題が扱われている可能性が高く，専門用語の共出現が多い関係にある文献にこのような傾向が顕著といえる．

我々のシステムでは複数の文献を組み合わせてマップすると，共通のキーワードをはっきりさせ，それぞれの文献に固有のキーワードをマップ上でも分離することが可能になる．類似度の高い文献どうしを組み合わせてマップすると，キーワード間のリンクが多くなり複雑なマップが描画される傾向がある．また類似度の低い文献には共出現するキーワードが少ないので，文献どうしの関連性を表すリンクが少ないマップが描画される傾向がある．本研究における類似文献の提案機能は，可視化データを抽出する文献を自由に組み合わせることによって，概念構造を多面的にとらえる機能を支援するものとして開発している．

3. 3. ユーザのための関連情報の表示

ユーザインタフェースの観点から，システムが現在どのような動きをしているかを，利用者に知らせることが重要である．システムでは文献から生成した概念ネットワークをエディタに描画しているが，この描画を生成するために付随した情報を利用者に提供する．それによって利用者がマップを生成する視点を切り換えたり，またはマップの生成自体を行なうかどうかを決める支援としている．開発者側からすれば，単に操作上のエラーメッセージを表示するだけではなく，システムの目的や機能を無視した使いかたを防ぎ，システム操作上における利用者の適切な認知モデルを形成することが重要になる．

利用者はシステムからのメッセージを参考にしながら，文献の選択や概念ネットワークを描画するための文献の組み合わせなどを行なうことができる．システムでは利用者が検索した文献の履歴を管理しており一覧することができる．それらは概念ネットワークを生成する場合の文献の組み合わせとして利用される．選択した文献のうち不要なものの選択を解除して，マップの対象となる文献の組み合わせを変更することができる．また文献の中に含まれている重要な部分がある程度想定できるように，文献に含まれている専門用語を一覧表示したり，可視化機能でマップされる概念ネットワークの元になったデータを表示したりできる．

4. TermMapper の可視化機能

可視化機能（TermMapper）は統合管理機能によって生成された可視化データを受け取り，概念ネットワークのマップを生成する機能やマップの操作編集機能を備えている．このエディタは KJ 法を取り入れた発想活動を行うエディタであり，概念ネットワークをマップし，発想を行なうための作業領域

を提供するエディタである[8)11)]．利用者はこのウインドウ上でＫＪ法を利用して，概念ネットワークの操作と編集を行なうことができる．これらの操作と編集をとおして，問題を考える発想活動を支援するのが，このエディタが持つ機能のねらいである．このエディタの機能を活用すると，さまざまな発想活動を行なうことができる．また操作の仕方としては次のような状況を想定したものである．

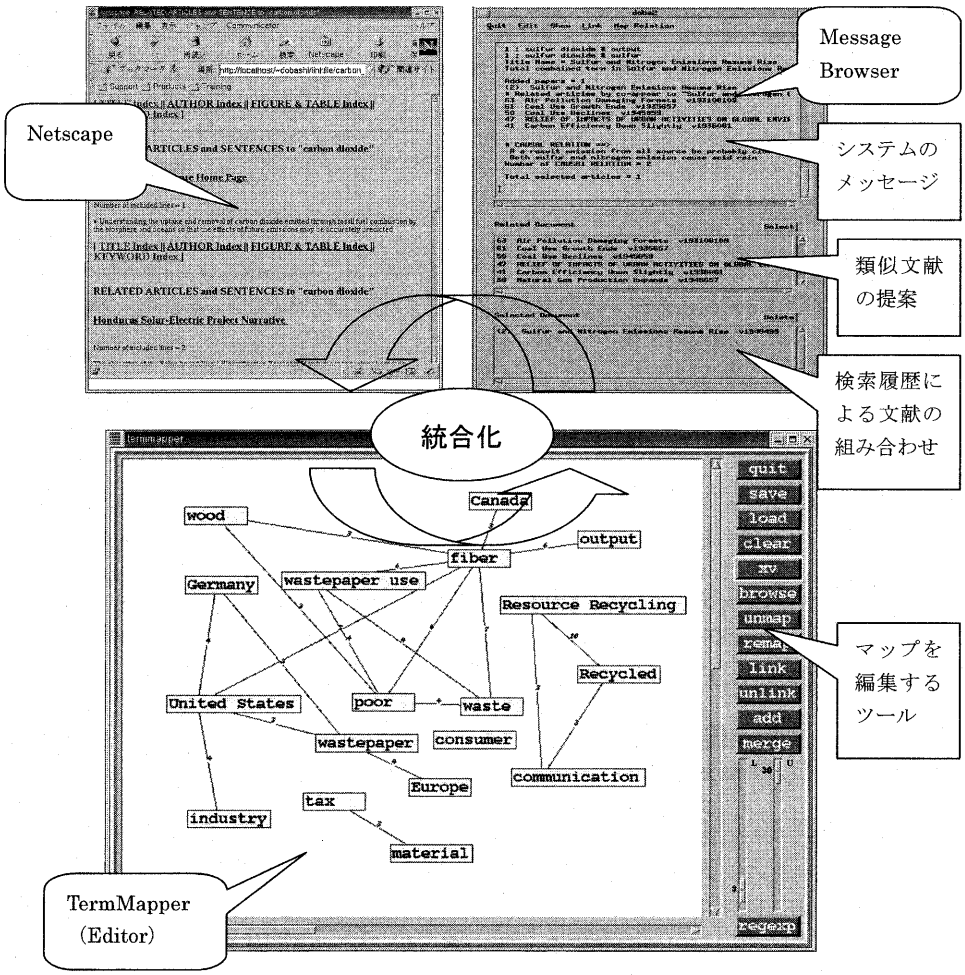

図５　TermLinker の全体

4．1．操作例と発想活動支援

ユーザが統合管理機能の画面からマップを描画するメニューを選択すると，TermMapper エディタが起動して該当する文献のマップが描画される．最初のマップが生成されたとき，抽出されるキーワードの数が多い場合，そのままでは見にくいことが多い．そのためスライダーを使って，とりあえず出現頻度の高い部分に絞ってみる．そうすると出現頻度の高い部分は，文献にとって重要な部分を示しているので，それを見ると何が主題となっているか，大体の見当をつけることができる．マップが見にくいときは，マウスのボタンを押しながらキーワードを選択して，TermMapper エディタのウインドウ上でいろいろな位置に動かしてみることを試みる．その間マップされた用語のつながりを考えながら，不要なキーワードを一時的に隠蔽したり，頻度の絞り込みによって隠された部分を再び表示したりしてみることができる．キーワードのつながりを考えながらそれらを動かしていくと，マップ

されたキーワードのつながりを文献上で確認したくなる場合もある．その時はキーワードが出現する文献とそのセンテンスを，キーワード索引を呼び出して即座にブラウジングすることができる．

同じ概念を指す類義語などがあれば，一つに統合することができる．またキーワードとキーワードの間に新たなリンクが必要な時は，その場で線分を引いてリンクを描画する．生成したリンクの上には，単語や数値を一緒に表示することができる．逆にリンクの線分が不要な場合は削除できる．キーワード自体も不要な場合は削除できる．

ある程度マップを見やすく整理して，保存の必要があれば作成したマップを保存しておき，後でまた呼びだして修正することができる．新たなキーワードがマップ上に必要になった時は追加することができる．複数の文献をマップした場合は，共出現の部分だけを表示することもできる．複数のウインドウに，複数のマップを生成して，比較しながら見ることもできる．

これらの機能を使いこなすためには，システムに慣れるまで多少時間が必要であるが，概念ネットワーク上でＫＪ法を行なうことによって，利用者の新たな認知構造の形成支援が期待される．

5. 関連研究

KJ 法に代表されるような概念構造を作図する手法は，多くの研究においてアイディアの発想や整理に活用されてきた [11]．しかし作図の前提となる問題のとらえ方や考え方が人によって千差万別なため，自由な発想を促すための情報の提供方法や種類にもさまざまな対応が必要であるが [2]，本研究では文献のテキストを対象とした概念ネットワークの生成と可視化方法が中心である．

最近になってユーザインタフェース研究との関連から，データマイニングの分野でも情報可視化技術の研究開発が盛んに行われるようになった．またテキストマイニングを用いた新しい手法も研究されてきている [14]．情報可視化の目的には，可視化そのものではなく，新たな用語や関連性の発見による情報検索支援や，新たなアイディアの生成支援などもある [1][12]．

文献情報をアイディア発想の源と考えたとき，文献に述べられた概念構造を可視化したい場合に，システム化を前提としたいくつかのパターンが考えられる．まずひとつの文献を読むだけでも，概念の構造を把握するのに十分な場合がありえる．あるいは複数の文献を読んで，それらに述べられた概念構造を関連付けて把握することが必要な場合も多い [13]．またひとつの文献を全部読まなくても，部分的に読めば十分な場合もある．

こういった状況を考えれば，システムによって文献から概念構造を自動的に描画する場合，文献を選択して内容を読めるようにするための機能が必要であるし，検索するためにはインデックスの生成なども必要である．このような機能を備えたシステムと概念構造を可視化する機能が連結していれば，文献を読みながら内容検索に必要なキーワードを提供したり，アイディアをまとめたりする支援システムが構築できると考えている．

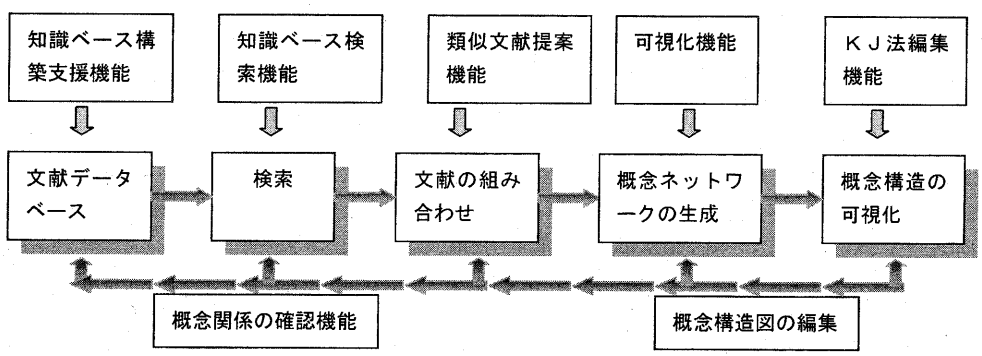

図 6　概念構造の可視化と発想活動支援のサイクル

図 6 は本システムを利用する場合，開発した機能がどの段階で使われるかを示したものである．文

献データベースの構築から，検索，文献の組み合わせ，概念ネットワークの生成，概念構造の可視化まで，研究論文に必要なアイディアをまとめる一連の流れに対応している．ユーザはこのサイクルの中で，発想活動の繰り返しを行なうことができる．システムが文献をそのまま提供するだけでなく，概念構造が見えやすくなるように可視化してくれることによって，「考えが及ばないような用語を見いだす」，「忘れていた用語を思い出す」，「思いも寄らない関連性に気づく」，「アイディアをまとめる」などの発想支援的な効果が期待できる．システムが単に検索結果を表示するだけでなく，問題構造を概念や概念間の関連性として表現できるならば，情報検索における上述した問題に効果が期待できるだけでなく，問題発見や問題解決における仮説生成を支援できる新たなシステムの開発が期待できる．

5. まとめ

インターネットに公開されている非定形で断片的な文書を知識ベース化して体系化すれば，再利用が促進される可能性が高まり，新たな関連性に気づくような効果も期待される．しかし現在のインターネットの情報量は既に人間の処理能力をはるかに超えており，これら知的資源の有効活用を支援できる技術が求められている．これは膨大な文献情報が公開される現在のインターネットが抱えている大きな課題でもあり，緊急に解決策の提案が必要とされている問題である．ここではこのような課題に対して，個人レベルで活用可能なひとつの提案を行っている．

謝辞

本研究の一部は愛知大学研究助成による．

参考文献

1) Card, S. K., Mackinlay, J.D., Shneiderman, B. : “Readings in Information Visualization Using Vision to Think”, Morgan Kaufmann, pp.686 (1999).
2) Chen C, : Information visualisation and virtual environments, Springer, pp.223 (1999).
3) Fayyad. U. M., Grinstein G.G., Wierse A.: Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, pp.407 (2002).
4) 林淑隆，中野英雄，獅々堀正幹，青江順一：”文字列照合マシンを利用した複合語キーワードの効率的抽出法”，情報処理学会論文誌, Vol.38, No.4, pp.815-825(1997).
5) 堀浩一：”システム統合のAIへむけて－発想支援系と知識処理系の結合の提案－”，人工知能学会誌, Vol.12, No.2, pp.85-89 (1997).
6) Hori, Koichi : “Concept Space Connected to Knowledge Processing for Supporting Creative Design”, Knowledge-Based Systems Vol.10, No.1, pp.29-35 (1997).
7) Ingwersen, P. : “Information Retrieval Interaction”, Taylor Graham, pp.246, 1992.(日本語訳：細野公男ほか訳，情報検索研究 －認知的アプローチー，トッパン, pp.378, 1995.)
8) 川喜田二郎：KJ法, pp.581, 中央公論社 (1986).
9) Lee, H. and Ong H. : “Visualization Support for Data Mining”, IEEE Expert, Vol.11, No.5, pp.69-75 (1996).
10) 那須川哲哉，諸橋正幸，長野徹：”テキストマイニングー膨大な文書データの自動分析による知識発見－”，情報処理, Vol.40, No.4, pp.358-364 (1999).
11) 杉山公造：”収束的思考支援ツールの研究開発動向－KJ法を参考とした支援を中心として－”，人工知能学会誌, Vol.8, No.5, pp.32-38 (1993).
12) 角康之：”情報可視化システムにおける適応的インタラクション”，人工知能学会誌, Vol.14, No.1, pp.33-40 (1999).
13) 辻井潤一：”ゲノム情報学と言語処理”，情報処理, Vol.43, No.1, pp.36-41 (2002).
14) 渡部勇：”ビジュアルテキストマイニング”，人工知能学会誌, Vol.16, No.2, pp.226-232 (2001).