98-01095

電子図書館における問題構造可視化支援機能の開発

土橋　喜　愛知大学現代中国学部助教授

1　はじめに

これまでに実用化された情報検索システムにおける検索機能は、単語を基本としており、その単語が表現している概念が文献の中で果たしている役割や、問題構造をあらわす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用い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極めて重要な問題となっている[9]。

最近のインターネットを利用した情報検索システムでは、検索結果が数十万件になることも珍しくない。インターネットに公開される情報量があまりにも膨大なことから、不要な情報も数多く検索されてしまい、この問題はさらに深刻になっている。情報検索システムにおいてこのような問題が起こる原因には、現実の問題構造に即した用語によるインデックスの生成が行われていないことや、問題構造をわかりやすく可視化する機能が備わっていないこと、および検索結果を絞込むための適切な検索語を選定する支援機能が不十分であることなどが指摘される。最近では情報の可視化技術を活用して検索機能に様々な改善が試みられているが、いまだに十分な解決に至っていない。

本論文では情報検索におけるこのような根本的な問題を解決するために、テキストマイニングの考え方を取り入れ、文献のテキストから問題構造を構成する概念関係を取り出して可視化する機能の提案を行っている[15,22,23]。さらに電子図書館機能と情報可視化技術を統合することによって、複数の視点から視覚的に問題構造を描画する機能を開発し、これによって問題構造に則した検索語の決定支援を行なうシステムの開発と評価を行っている[2]。

2　専門用語の抽出による問題構造図の生成と可視化

まず文献から問題構造図を生成して半自動で再編成し、可視化する方法について述べる。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に対する考え方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、それらの概念間に潜む重要な関係に触れていると解釈される。主に名詞から構成される専門用語がこれらに該当するが、このような重要な用語間に構成される関係は、従来の上下関係などに代表される概念関係とは異なることから、ここでは問題構造と呼ぶことにしている。

これらの専門用語や出現頻度の高い用語を、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問題構造の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に関する具体的なひとつの文を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、専門用語辞書または出現頻度によってsulfur、nitrogen、acid rainという3つの専門用語を取り出す。次にsulfur‐nitrogen、sulfur‐acid rain、nitrogen‐acid rainというように専門用語をペアにした組み合わせを作る。さらにこれらの共通の文字列を連結して組み合わせの出現頻度を付与して概念ネットワークを生成し、コンピュータの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、その文献の著者の問題構造をネットワーク形式で描くことになり、複数の文献に対して行なえば、複数の文献に述べられた問題構造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、概念ネットワークによって描かれる問題構造図は、文献に述べられた問題構造を表現する用語の使われ方によって、それぞれ異なるものが描画されることになる。ここでいう問題構造や概念関係は、言い換えれば広い意味において文献に述べられた著者の知識の構造である。最近では文献などのデータベースからこのような知識の連鎖構造のルールを発見するための手法の開発が盛んである[5,18]。例えば本研究における概念ネットワークの自動生成は、データベースから隠された関連性や規則を見出す知識発見やテキストマイニングとその可視化の研究［4,14］および人間の創造性の増幅をめざす発想支援［6,7,8,12,13］の分野で行なわれている研究と関係が深い。

概念ネットワークの要素となる専門用語の関係を明確に定義するためには、文献の中で著者が表現している重要な概念を取り出し、それらの概念関係がどのようにつながっているかを識別しなければならない。しかし全ての概念関係を識別することは膨大な知識を必要とするため極めて困難である。加えて既に明白となっている一般的な用語の関係を大量に識別しても、インデックスとしては使えるが専門家にとって新たな発想の刺激となるかどうか疑問が多い。そのため本システムの目標を、問題をもれなく説明する構造の可視化ではなく、ユーザが何らかの発想の刺激を感じられるように、断片的な知識を結び付ける構造を自動生成することにおいた。

そしてこれまでの概念ネットワークは、主に意味的な上下関係や連想関係など一定の構造を表現する方法として、様々な分野で利用されてきたが、本システムでは概念間の関係を記述することは行なわず、同一センテンス上に現れた専門用語のペアを作り、これによって概念関係を専門用語の組み合わせとして自動的に抽出することを考案した。これはKJ法を行なう場合にカードの初期配置をランダムにしたほうが良いという考え方を取り入れたもので [11]、連結された用語と用語の関連性をユーザが自分で推測したり、あるいは文献に戻って用語間の関連性を確認したりすることによって、発想を促す支援を行なうためである。

3　インデックス間の隠れた関係の可視化

図1はシステムによる概念の組み合わせと基本的な概念ネットワークの生成方法を概念図で示したものである。この後の図3で示すような複数文献による問題構造図を生成するためには、図2に示したようにいくつもの手順を踏んで、概念ネットワークを生成することが必要になっている。図2では収集した文献をデータベース化する段階から、問題構造図を生成するまでの流れを示している。本研究ではシステムによってこれらの自動化を行なっている。またここでは生成された概念ネットワークを可視化し、ユーザ自身よって見やすく整理したマップを問題構造図と呼んでいる。

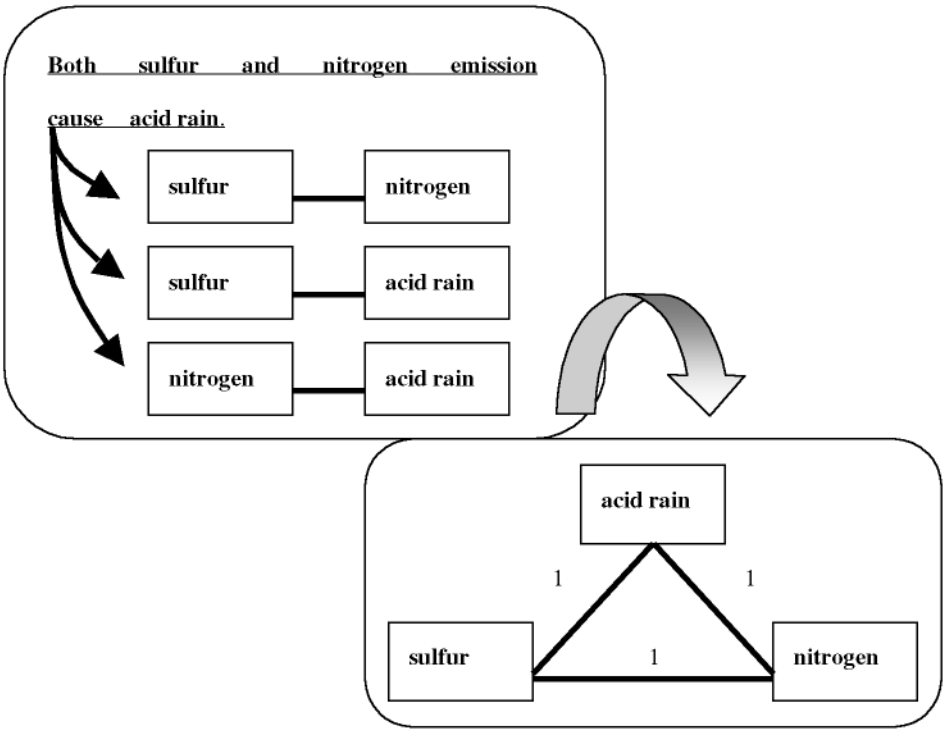

図 1 の左上は本システムが文章から抽出した概念の組み合わせ方の例を示し，右下はそれを最も基本的な概念ネットワークの形式にマップしたところを示している．リンク上の数値は，組み合わせの出現頻度を表わす．

図1　システムによる概念の組み合わせと基本的な概念ネットワークの生成

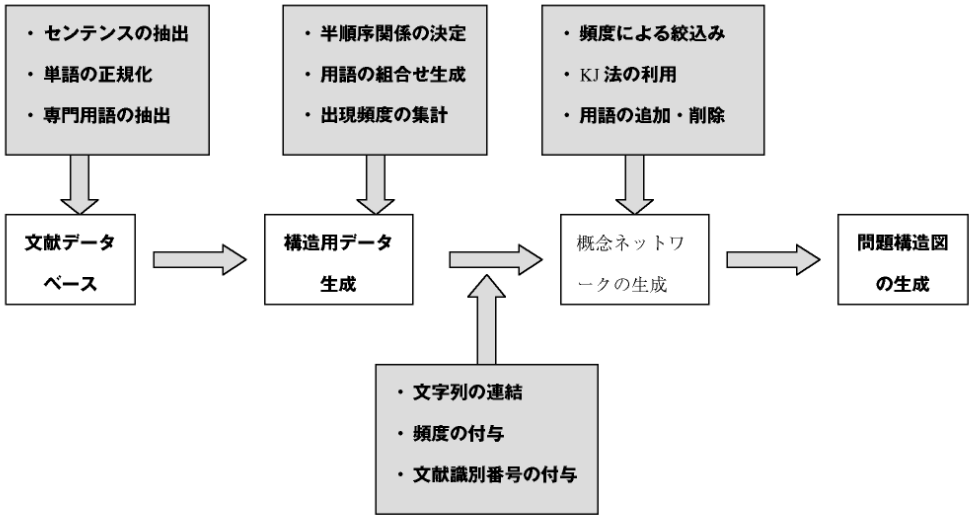

図2　テキストからの問題構造図の生成課程（概念図）

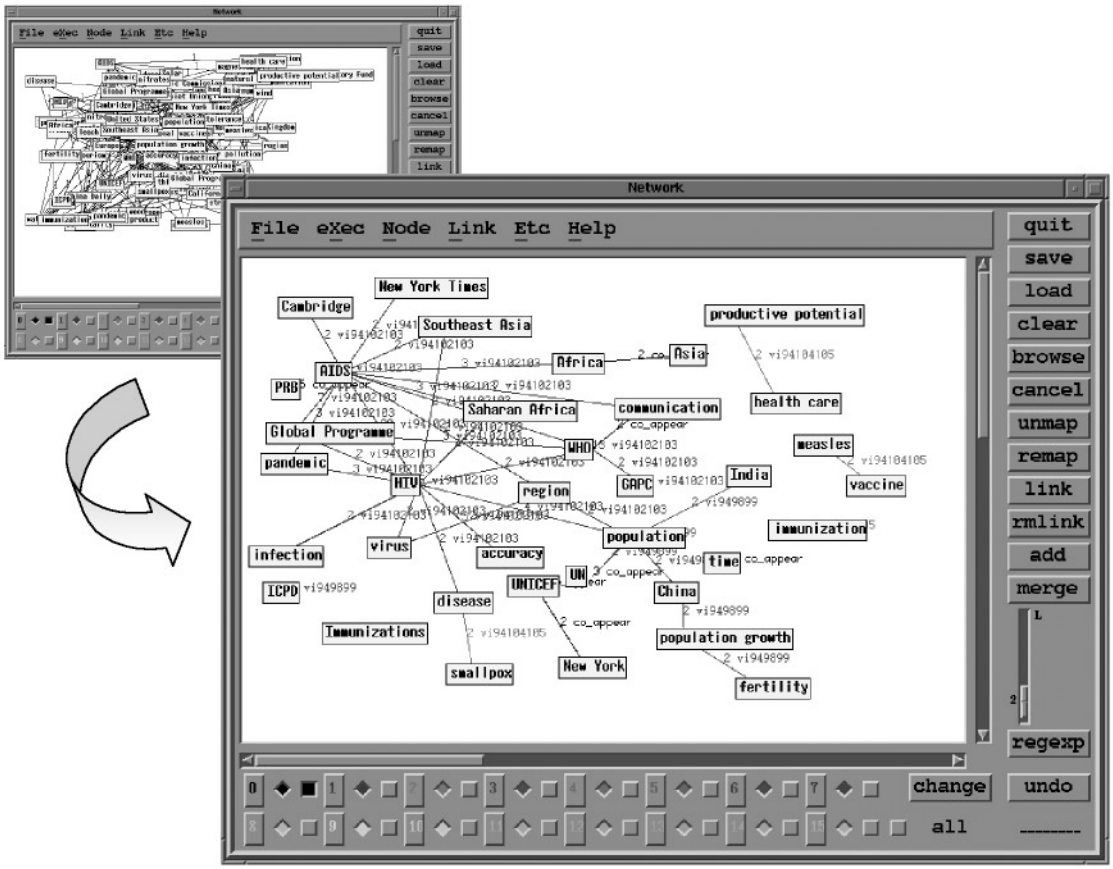

図3　複数文献のマップによる問題構造図の可視化例

本研究では地球環境問題に関する文献を例に取り上げており、図1の英単語は実際にシステムによって抽出された専門用語の例を表わしている。実際のシステムでは専門用語の組み合わせの出現頻度を集計し、専門用語と専門用語を結び付けるリンク上に出現頻度を表示している。専門用語を抽出する場合、地球環境問題の専門辞典のインデックスから作成した専門用語辞書（実験では9,980語）を使う方法と、システムによって出現頻度の高い用語を自動的に抽出する方法を併用している。このふたつの方法を組み合わせることによって文献に表現されている重要な部分を取り出す効率を上げることが

できる。これは出現頻度が低くても重要な用語は取り出す必要があり、専門家向けの効果を期待した工夫である。

概念ネットワークを生成する場合、元になった文献に含まれる問題構造を表わす重要語を抽出してそれらを連結して描画している。生成された2項関係の組み合わせは、文献のセンテンスごとに行っているので、文脈を反映したものとなり、それぞれの用語は文献データベースを検索するためのインデックスとして使われている。地球環境問題の文献などでは、これらの組み合わせを2次元平面に描画すると、あたかも問題とその構成要素からなる概念ネットワークのように描画される。概念の連結は2項関係を生成した場合に見出される専門用語の文字列の一致を手がかりに行っているので、このような連結方法を取れば、文献が複数の場合でも、異なる文献間にまたがる概念ネットワークを自動生成することができる。

生成された概念ネットワークは、概念間のつながりを自動的に線分で連結し、ユーザの自由な発想を促すため、2次元平面にランダムに描画される。ユーザはランダムに描画された専門用語を整理しながら、それらが連結された背景に潜む関連性を考える発想活動を行なうことができる。またそれらの用語に張られたハイパーテキストのリンクをたどりながら、文献の検索を繰り返すことができる。

図3は複数文献による問題構造図の可視化例を示したものである。図3おいてに左上の図はシステムによる初期マップであり、まだユーザが手を加えていない段階である。それに対して同じく右下の図は、ユーザ自身の手によってマップされた専門用語を移動して、問題構造図を編集したものである。ユーザはこの問題構造図を作成するために3つの文献を選択した。選択された文献は2冊の論文集(Vital signs, 1993, 1994)中で、数ページにわたり掲載されたエイズ（AIDS）に関する論文である。ユーザは図3の問題構造図を編集するために、専門用語の組み合わせの出現頻度を2以上に絞り込んで、見やすくする工夫をしている。

この問題構造図からは、AIDSとHIVウイルスとの関係や、AIDSの症状の一部がうかがわれる。最近ではアフリカと東南アジアにおけるAIDSの流行が大きな問題となっていることがマスコミなどでしばしば報道されるが、この問題構造図にもこれらを示すキーワードが抽出されている。さらにAIDSによる若年層の死が、これらの地域における人口構成に問題を起こし、出生率にも何らかの影響を与えていることがわかる。そのほかマップされた用語のつながりを追って行くと、WHOなどの国際機関がAIDS問題に関して活動を行なっていることも容易に把握される。

問題構造図にマップされた専門用語の関係をどのように推測するかはユーザの判断によるが、システムではこれらのマップされた用語から元の文献に戻り、専門用語が出現しているセンテンスを見て、マップされた関係を確認することができる。

4　電子図書館と可視化支援機能の統合

KJ法に代表されるような問題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[11,17]。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と種類にもさまざまな対応が必要であるが、ここでは文献のテキストを対象とした問題構造図の生成と可視化方法を中心に取り上げている。

最近になってユーザインタフェース研究との関連から、情報可視化技術の研究開発が盛んに行われるようになり、新しい手法も開発されてきている[16,20,21]。情報可視化の目的のひとつには、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。

文献情報をアイディア発想の源と考えたとき、文献に述べられた問題構造を把握する場合に、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。まず主にひとつだけの文献から、問題の構造を把握すれば足りる場合がある。次に複数の文献を読んで、それらに述べられた問題構造を関連付けて把握することも頻繁に行なわれる。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問題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問題構造を可視化する機能が連結していれば、文献を読みながらアイディアをまとめる支援システムが構築できる。最近になってインターネットブラウザを活用した電子図書館システムの開発が行なわれており、これらの研究を参考に本研究でも電子図書館機能と密接に連結させる方法を開発している[1]。

例えば複数の視点から概念ネットワークを生成する機能や、生成した概念ネットワークの描画をインタラクティブに操作して問題構造図に仕上げる機能などを開発した。ユーザはこれらの機能を使い分け、文献データの問題構造を複数の視点から可視化し、インタラクティブに観点を変えながら、描画された問題構造図を操作することができる。

4.1　問題構造の生成における視点の転換機能

問題を考えるときに全体的な視点の転換をしてみることも必要になる。視点を全体的に転換する目的は、今まで考えていたものと全く異なる問題を考えたいとき、あるいは別な著者の考え方を参考にしたいとき、さらに複数の著者の意見を

統合して考えてみる場合などである。また全くことなるキーワードで問題を考えるなどもこれに相当する。これは問題の構造を表わす概念ネットワークのマップを全く新たに生成しなおすことを意味している。これらの点を考慮して概念ネットワークによる問題構造図の生成は、次の3つの視点から行なうことができる。

（1）単一文献からの生成

ユーザがインターネットブラウザで検索したひとつの文献を対象にして、概念ネットワークを生成する。従って描画される問題構造の範囲は、選択した文献の内容だけに限定される。

（2）複数文献からの生成

本システムはユーザが検索した文献を記録しており、そのなかからユーザが自由に文献を組み合わせて、概念ネットワークを生成する。描画される問題構造は選択した複数の文献にまたがったものとなる。

（3）キーワードにもとづく生成

ユーザが指定したキーワードで、データベース全体を検索して、キーワードを含むすべてのセンテンスから、概念ネットワークを生成する。問題構造は指定したキーワードとその数に依存するが、キーワードを複数入力すれば、それらに関連した概念ネットワークとなる。複数の文献にわたり、キーワードに関連した用語を網羅的に調べたいときに効果がある。

4.2　知識発見の支援機能をめざして

図4はシステム構成図である。本システムはインターネットブラウザとテキストマイニングの統合によるインターネットからの知識発見を目標に、インターネットに公開されたテキストまたはHTMLによってタグ付けされた文献が分析の対象である。本システムはインターネットブラウザ（図ではDocument Browser）、類似文献の提案などシステムのメッセージを表示するMessage Browser、概念ネットワークを描画するNetwork Editorの3つの部分から構成される。なおインターネットブラウザは、収集した文献の検索および内容を読むために使われる。システム全体ではインターネットブラウザとそのアクセスログを使った電子図書館を構成しており、システムを統合するため次のような機能を開発した。

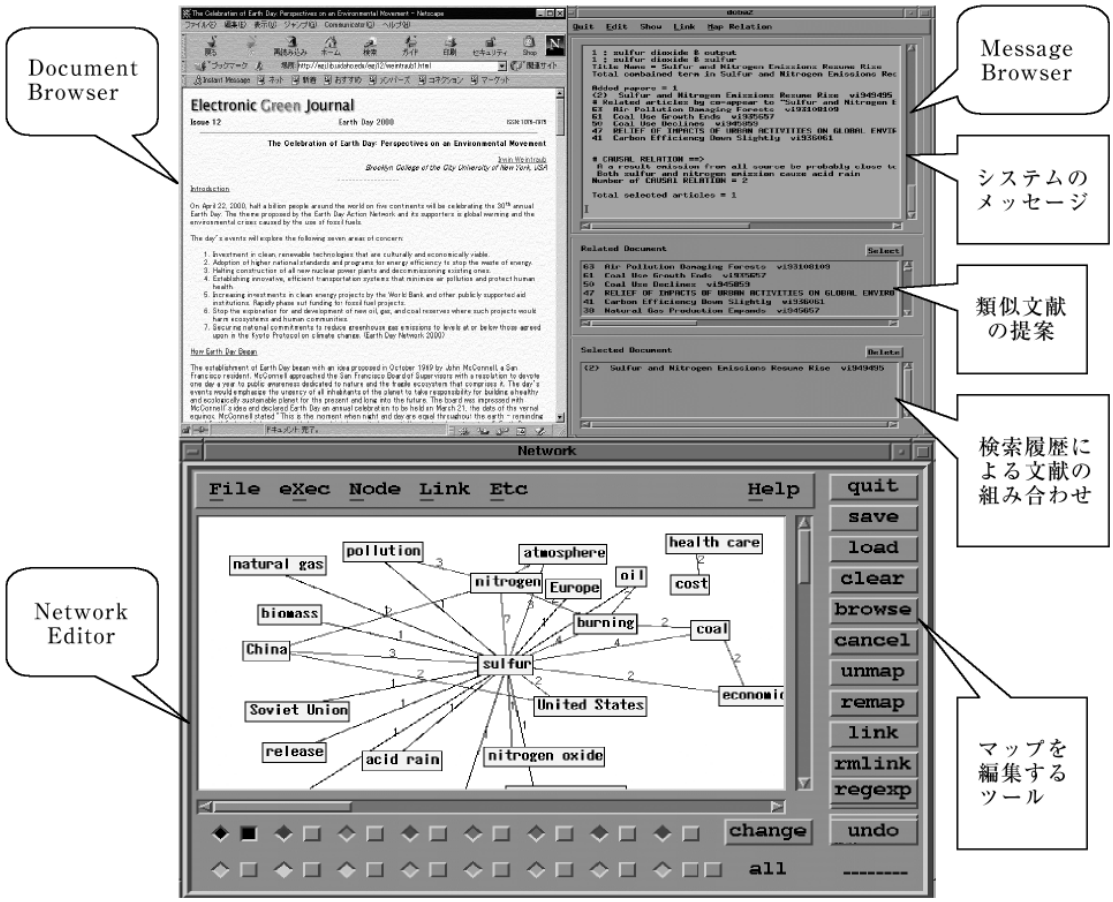

図4　システム全体のイメージ

（1）知識ベース構築機能

本システムはインターネットなどを利用して、ユーザが自分で収集した文献が分析の対象であるが、テキストまたはHTMLによってタグ付けされた文献は、知識ベースに追加することができる。

（2）知識ベース検索機能

知識ベースを検索するため、HTMLを使ってタイトル、著者、図表にリンクを生成する。著者や図表などHTMLのタグのないものは手作業で付与した。文献から抽出した専門用語によるキーワードインデックスは、用語が出現する文献および文献中の該当するセンテンスがブラウズできるように、リンクを自動生成する。キーワードインデックスはHTMLのタグを利用して、ユーザのキーワードからその場でリンクを生成することが可能であり、検索機能を同時に実現したものになっている。

（3）類似文献提案機能

ユーザが文献を一つ選択すると、システムは自動的に内容の類似した文献を提案する。文献間の類似度は、共出現した専門用語の出現回数の合計で求めている。これによってユーザは、類似した文献が提案されるウインドウから文献を選択することによって、内容的に関連した文献を容易に収集することができる。

（4）文献組み合わせ機能

複数の文献から概念ネットワークを生成する場合、それらの組み合わせを自由に変更することができる。

（5）概念関係の確認機能

開発したシステムでは複数の意味を持つ用語を区別して連結することが困難であるため、生成された関連性が妥当かどうか、確認する機能が必要である。開発したシステムによってマップされる用語は、文献のキーワードインデックスに自動的にリンクしており、用語が出現する文章を、インターネットブラウザを利用して一覧することができる。この機能によって用語がどのような文脈で使われているかを即座に確認することができる。

図5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、問題構造の可視化まで、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、問題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[3]。

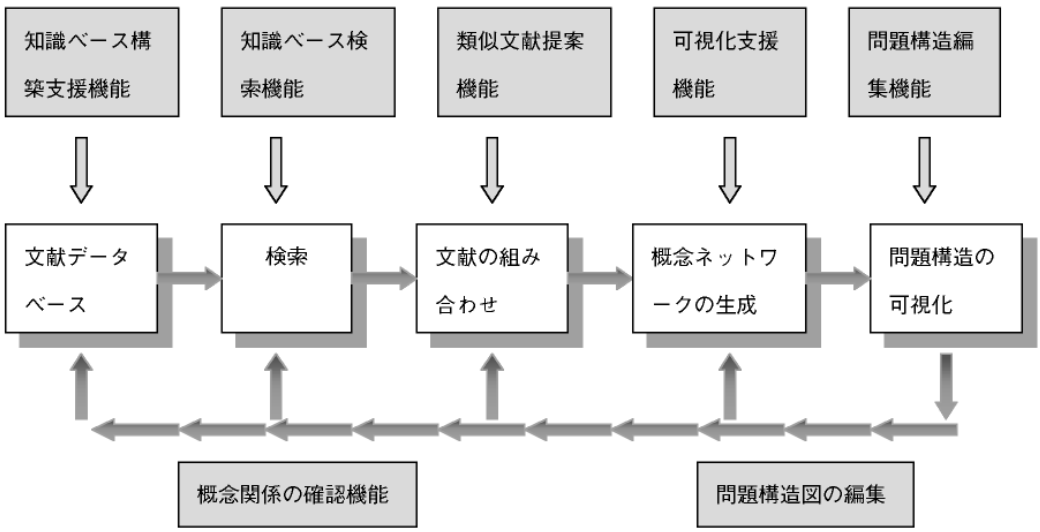

図5　問題構造の可視化による発見支援のサイクル

5　実験と評価

我々は日常的にいろいろなところに分散した情報資源から情報の収集を行っている。インターネットに公開された情報と検索システムを活用すると、個人的なデータベースを構築して活用することができる。このような場合に収集した文献を有効に活用するためには、収集した文献やデータをよく理解し、有用な知識を迅速かつ的確に発見する支援システムが必要である。インターネットからの知識発見を支援する新たな情報可視化技術の開発が必要になっており、本システムはこのような社会的な背景をも考慮して開発した[15]。

実験に用いた文献は、地球環境問題に関連したものをインターネットや学術雑誌から収集した。文献は全て英語である。データベース全体では、文献のタイトル数は133、文献に含まれている図表は241である。

実験ではインターネットブラウザを使って、あらかじめ収集した文献データベースの検索をしながら、内容を読んでもらい、システムによって生成された問題構造図が与える効果について意見を求めた。被験者は東京大学先端科学技術研究センターの大学院生(22人)をお願いした。被験者の内訳は地球環境関係の大学院生7人と情報工学関係の大学院生15人である。被験者の意見をまとめると、生成された概念ネットワークによる問題構造図には次のような特徴と効果があることが明らかになった。

（1）専門用語の組み合わせを生成すると、頻繁に利用されるものとそうでないものの間に出現頻度に差が生じる。この出現頻度の差を利用して、問題構造を表している可能性の高い部分と、可能性の低い部分をおおよそ描き分けることができる。
（2）出現頻度の高い部分は、著者によって頻繁に利用された専門的な用語の組み合わせであり、文献の内容から見れば相対的に重要な部分である。この部分は常識的な用語で問題構造を説明するようなつながりを構成しているので、その領域の専門家ユーザが本システムを使っても、新たな知識に気づく可能性は少ない。しかし文献の主題を視覚的に把握することができるので、文献の要約を読むことと似たような効果が得られる。
（3）逆に出現頻度の低い用語の組み合わせは、著者が文献中で言及した回数が少なく、システムが新たに組み合わせを生成した部分が多い。文献の内容から見れば相対的に重要性が低いと言えるが、システムが新たな概念関係を提案している部分である。専門家にとっては新たな着想のきっかけや異なる視点の発見につながるような部分は、この頻度の低い部分に含まれている可能性が高いという意見が得られた。
（4）出現頻度の高い専門用語を抽出すると、名詞だけでなく動詞なども抽出される場合がある。非専門家は名詞と動詞が連結されると、専門用語の関係がより明確に把握できる。また名詞だけの場合でも、専門用語と専門用語の間にどのような関係があるかを考えるようになる。
（5）画面では文献の全部を表示できないことが多いので、画面から隠れた部分にどのような重要語が隠れているかなどを把握しやすい。

6　まとめ

現在の文献情報検索システムは、文献における概念の重要性や概念間の関係を正確に表現することはできない。そのためテキストマイニング、情報可視化、システム統合、インターネットからの知識発見、発想支援などの研究成果を取り入れて、問題構造を可視化支援するシステムを提案した。被験者による試用実験の結果から、生成された概念ネットワークによるマップが問題構造を反映しており、文献の内容が把握しやすいことが明らかになった。また気づかなかった用語に新たに気づいたり、複数の文献にまたがる問題の関連性を想起したりしやすいなどの効果が認められた。

謝辞

本研究は東京大学先端科学技術研究センターの堀 浩一教授、中須賀真一助教授、山内平行氏、日本アイ・ビー・エム東京基礎研究所の立花隆輝氏に多大なご指導とご支援を受けた。ここに感謝の意を表す。

参考文献

[1] “ Toward a World Wide Digital Libraries ”, Communications of the ACM, Vol.41, No.4 (1998).
[2] Card, S. K., Mackinlay, J.D., Shneiderman, B., “ Readings in Information Visualization Using Vision to Think ”, pp.686 (1999).

[3] Davis, W. H., Peirces's Epistemology, Nijhoff, 1972.（日本語訳：赤木昭夫訳、パースの認識論、産業図書、pp.288, 1990.）
[4] Fayyad, U. M. (et al eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, pp.611 (1996).
[5] 福田剛志、”データマイニングの最新情報－巨大データからの知識発見技術－”、情報処理、Vol.37, No.7, pp.597-603 (1996).
[6] 堀浩一、”発想支援システムの効果を議論するための一仮説”、情報処理学会論文誌、Vol.35, No.10, pp.1998-2008 (1994).
[7] 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).
[8] Hori, Koichi, “Concept Space Connected to Knowledge Processing for Supporting Creative Design”, Knowledge-Based Systems Vol.10, No.1, pp.29-35 (1997).
[9] 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－、トッパン、pp.378, 1995.)
[10]“特集：ディジタル図書館”、情報処理、Vol.37, No.9, pp.813-864 (1996).
[11] 川喜田二郎、KJ法、pp.581, 中央公論社 (1986).
[12] Knowledge-Based SYSTEMS, Vol.10, No.1 (1997).
[13] 國藤進、”発想支援システムの研究開発動向とその課題”、人工知能学会誌、Vol.8, No.5, pp.16-23 (1993).
[14] Lee, H. and Ong H., “Visualization Support for Data Mining”, IEEE Expert, Vol.11, No.5, pp.69-75 (1996).
[15] 那須川哲哉、諸橋正幸、長野徹、”テキストマイニング－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).
[16] Sarkar M. and Brown M. H., “Graphical Fisheye Views”, Communications of the ACM, Vol.37, No.12, pp.73-83 (1994).
[17] 佐藤允一、問題構造学入門－知恵の方法を考える－、ダイヤモンド社、pp.223 (1984).
[18] Shen, W., “Discovering Regularities from Knowledge Base”, International Journal of Intelligent Systems, Vol.7, pp.623-635 (1992).
[19] 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).
[20] 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).
[21] 舘村順一、”Docspace：文献空間のインタラクティブ視覚化”、田中二郎編「インタラクティブシステムとソフトウエアIV」日本ソフトウエア科学会WISS '96、近代科学社、pp.11-20 (1996).
[22] 三末和男、渡部勇、“テキストマイニングのための連想関係の可視化技術”、情報処理学会研究報告99-FI-55,99-DD-19、情処研報、Vol.99, No.57, pp.65-72 (1999).
[23] 渡部勇、三末和男、“単語の連想関係によるテキストマイニング”、情報処理学会研究報告99-FI-55,99-DD-19、情処研報、Vol.99, No.57, pp.57-64 (1999).

＜発 表 資 料＞

題　　名	掲載誌・学会名等	発表年月
WWWとテキストマイニングの統合による問題構造可視化支援	信学技法 Vol.99, No.447, pp.51-58	1999年11月
テキストマイニングによる問題構造の生成と可視化支援	2000年情報学シンポジウム論文集 情報処理学会 pp.161-168	2000年1月
情報視覚化と問題発見支援 －問題構造の可視化による仮設生成－	あるむ pp.302	200年2月