人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第7回)
SIG-AM-07-02

文書情報を活用した連想支援システムの開発

Development of an “association” support system using document data

荒井　豊文 [1]

Toyofumi ARAI[1]

[1] 中京大学
[1]Chukyo University

概要： 蓄積した文書情報の中から，ユーザーが指定した情報要求に基づき抽出した情報を木構造(語木)表現で視覚化し，さらにユーザーの操作に応じてインタラクティブに変化させることで連想を支援するシステムを試作した．

文書情報の木構造表現による視覚化や，視覚化した文書情報を変化させる動作には，人のメンタルモデルに関する先人らの研究により得られた知見や経験則を反映させることを試みた．

人の情報認知に関するメンタルモデルに則した動作で情報の提示を行えるようにしたことで，連想支援に有効な効果が期待される．

Abstract: We have created a system that provides support for association by visualizing in a tree structure (word tree) information filtered based on a request for information specified by the user from information stored in documents and, further, based on user operation, changes this information interactively.

Through the visualization of document information in the form of a tree structure and the operation of changing the visualized document information, we have tried to reflect the knowledge and rules learned through experience obtained through the research of our predecessors into the human mental model.

It is expected that, by presenting information through actions that follow a mental model of human information recognition, that there will be useful benefits for association support.

１．はじめに

研究など創造的要素を含む知的活動においては，新たな気づきや発想を得るためのアプローチとして，連想を用いることがある．連想の情報源として論文などの文書情報を用い，これを熟読することが多い．しかしながら，文書テキストのままでは内容理解のための認知的負荷が高く，文書情報から連想を行う上での障壁となっていることが予想される．

そこで，文書情報理解のための認知的負荷を低減させ，連想を促し，気づきや発想を生み出しやすくすることを狙った支援システムを検討，試作した．

連想支援の方策としては，非定型情報である文書情報に対し一定のルールを適用し，形をもたせ視覚化し，視覚化した情報をユーザーの操作に応じて変化させることができるようにすることで，ユーザーの情報認知に刺激を与え連想を促す情報提示システムを考えた．

このようなシステムにおいて文書情報に形を持たせるためのルールは，人のメンタルモデルに則した方法を適用するのが認知的負荷の低減に有効と考える．またシステムとユーザー間のインタラクティブな情報のやり取りにおいては，行った操作に応じてシステムが提示する情報の変化を視覚的に認知できることも有効と考える．そこでこれら２点の実装に重点を置きシステムを試作した．

２．システムの検討

検討したシステムの構造を図１に示す．連想のもととなる文書情報を格納した情報源と，その情報をハンドリングするロジックからなる構造とした．ハンドリングロジックではユーザーの要求に対応した情報を情報源から抽出し，予めユーザーが指定した描画方法で視覚化し提示するとともに，一旦提示した情報に対し，ユーザーが操作を加えることにより

視覚情報を変化させることができるようにした.

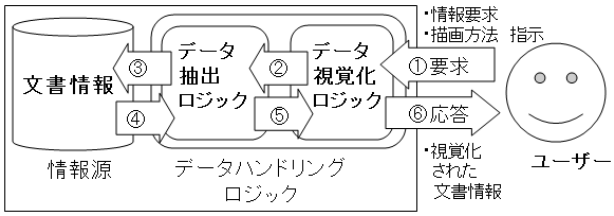

図１．システムの構造

2.1 情報源

情報源に用いる文書情報には特許情報を用い，これを関係型データベースに格納して用いた．特許情報は先人たちの知的活動の成果物であり，しかも電子化された情報を大量かつ容易に入手し利用できることから，提案システムの有効性を検討する際にも適切であると考えた．またデータベースを用いたのは，文書情報を解析に適した形に構造化してストックしておくことで，これを利用する様々な要求に対応できるようにするためである[1].

2.2 データハンドリングロジック

データハンドリングロジックは，データ抽出ロジックとデータ視覚化ロジックに分離し開発した.

2.2.1 データ抽出ロジック

データ抽出ロジックの中心は情報源である関係型データベースへのアクセス機能であり，データ視覚化ロジックでユーザーが指定した情報に基づき SQL 文を生成，実行し，必要な情報をデータベースから抽出する．さらに，抽出した情報をユーザーの指定した条件に応じて加工，編集する.

2.2.2 データ視覚化ロジック

システムとユーザーとのインタフェースであり，本システムの最重要部分である．メンタルモデルに則して情報を視覚的に表現するにあたっては，シンプルなルールにより意味づけした形で提示することとした．視覚情報とその意味との関係が複雑になると，新たな認知的負荷がユーザーに生じる恐れがあると考えたからである

また，ユーザーの思考を中断させることで新たな認知的負荷が発生させることの無いように，一連の操作が容易に繰返し実行できることも重要と考えた.

３．システム開発内容

3.1 機能実現の方策

情報の視覚化や，システムとユーザーのインタラクティブなやり取り，大量文書情報中からの情報抽出などに有効と思われる，人のメンタルモデルに関する先人の研究成果や経験則には，たとえば,

① 人は情報の集まりを見ると，そこに含まれる規則を見出そうとする．[2]
② 人は情報の並びを見ると，そこに含まれる規則を考え，それを元に次に現れる情報を先読み(予測)しようとする．[3]
③ 段落など，特定の部分を単位として検索し提供することにより，関係した情報を効率的に抜き出すレレバントな情報検索ができる．[4]
④ 複雑な理論により少量の情報を分析するよりも，単純な理論で大量の情報を分析した場合の方が有効な結果が得られる場合がある.

などがある.

①，②よれば，情報の提示方法を工夫することで連想が促進できると考えられる．また，③は，大量の情報の中から有効な情報を抽出すことに関するするものであり，人は文書中の纏まった箇所に特定の話題に関する内容を集中させる傾向があることを示すもので，これを利用すれば有効な情報をユーザーに提供し易くなることが考えられる.

さらに，これらのほかにも，ユーザーが使って楽しく感じるか否かということもシステムの有効性に影響することが知られている.

そこで，先人らの研究成果や知見や経験則を参考とし，検討した結果,「連想ゲーム」[5] 的動作を実装することとした．連想ゲームでは回答者が正解を答えるまで，ヒントとなる言葉が繰り返し提示され, 正解に至ることが必要であるのに対し，提案システムでは正解は求めない．またヒントに相当するものとして提示される情報は，抽出された文書情報に含まれているものに限定される．こうした違いはあるものの，関係を持つ情報を次々と提示し連想に結びつけるといった基本動作においては共通するものがあると考えた.

3.2 ユーザーへの情報提示

抽出した文書情報を視覚情報としてユーザーに提示し，さらに「連想ゲーム」的動作をさせるための単純化したイメージを図２に示す.

描画の形とその意味の関係の基本ルールとしては，ユーザーの情報要求に合致しているとして抽出された文書中に含まれる語の，出現頻度の多さを円の大きさで，またその文書中に含まれる文中での各語の共起関係の強さを円どうしを結ぶ線の長さで示すとした．これにより，根語を始点とし

て線で結ばれた各語をたどり終端となる語までの一連の語の並びが一つの文に相当し，円で示された語は文中に出現する語群を表すようにした．

根語に用いる語は，ユーザーの情報要求として入力された文中に含まれる語，もしくは情報要求に合致しているとして情報源から抽出された文書内の特徴語である．いずれを根語に用いるかは，ユーザーが指定できるようにした．

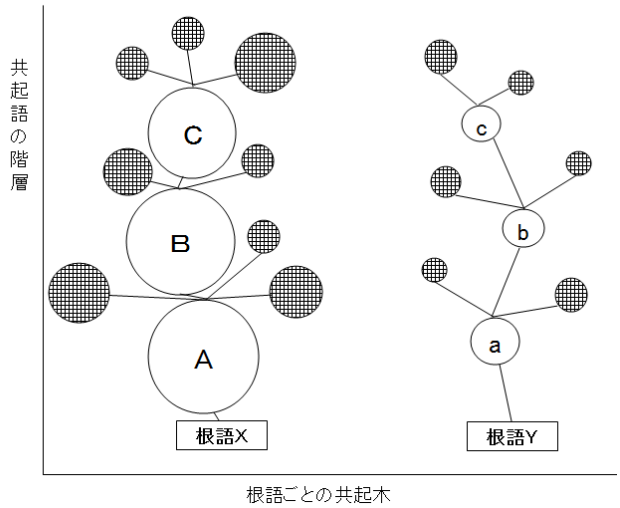

図２．語木表示イメージの例

情報要求に含まれる語を根語に用いた場合，語木はユーザーの視点を反映させたものとなる．一方，特徴語を根語に用いた場合はユーザーの視点とは別に，情報要求に合致した文書が持つ特徴をもとに形成した語木となる．前者のように視点をもって情報を見ることも重要ではあるが，後者の機能により，より自由な連想の元となる情報の提示が期待される．

このような表現ルールを用いた「連想ゲーム」的情報提示の主な動作は，

・ユーザーの操作に応じ，根語から枝葉語が展開するように段階的に表示する．
・語木を構成する個々の語系列単位で順番に強調表示したり，ユーザーが任意の一語を指定することで，その語が含まれる系列を強調表示する．
・語木で表現した語系列情報を元に再検索するなど，次の操作を連携して行うことができる．

などができるようにすることとした．

情報の関係を視覚的に表現することに関する先行研究では，ネットワーク図で表示するものが多い．提案システムで情報の視覚化に語木（木構造）を用いたのは，「大きさ」「長さ」「始点と終点」「方向性」「並び」「順序」等，人が容易に認知できる情報の尺度を対象に持たせることで，メンタルモデルを利用する効果をより有効にし，他の視覚化方法よりも情報認知において優位とすることを狙ったからである．さらに語木で表現することにより，人が「木」に対して持っているメタファーが連想の促進に生かされる効果も期待した．

このような表現方法によれば，たとえば文書中で強調されている内容については，同様の語群の語を用いて繰り返し文書中に記述されるであろうことから，それら語の出現頻度，共起頻度ともに高くなると予想され，図２の左に示した根語Ｘから始まる「X-A-B-C」の語系列のように語を囲む円が大きく表され，また語どうしを結ぶ線が短くなり互いの語が近くに描画されると推察される．逆に，述べられる頻度が少ない文を構成する語群は，図２で右に示した根語Ｙから始まる「Y-a-b-c」の語系列のように語を囲む円が小さく表され，また語どうしが離れて描画されると推察される．

さらに，複数の文書や段落の情報を一つの文書や段落とみなして描画することもできる機能も付加した．これにより，たとえば作成者の異なる複数の文書に含まれる情報を用いて描画した語木において，大きな円で囲まれた語の並びの語系列が出現した場合には，複数の人により同じ主張がなされている可能性があると推察され，より信頼性の高い情報を示すものになることが予想される．

マウス操作で根語から順に枝葉となる語を表現する「連想ゲーム」的動作では，各操作を実施する時点までに描画されていた語もしくは語群が，次に描画される語を推測するヒントの役目を果たす．動作イメージを図３に示す．

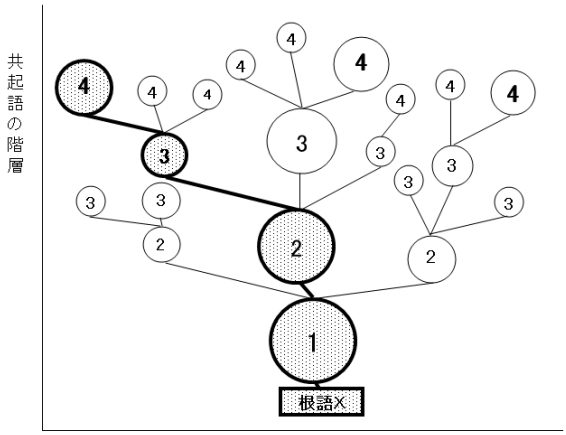

図３．段階的な語木表示及び特定語系列強調の例

図３の円の中に記した数字は段階的に描画される順番を表し，たとえば①の語をクリックすると②で表された語が表示される．

さらに，ユーザーの操作により，図３に太線で示したように特定の語系列を構成する語を他の語系列

と区別し強調表示することもできるようにした．さらに，強調表示した語系列に含まれる語を用いて文書データベースを再検索し，語木を再描画することや，該当する文書の原文を検索できる機能とも連携させた．これにより，ユーザーの思考を中断させずに必要な情報を提示できる効果も期待される．

なお，情報要求と，情報源から抽出する文書や段落との適合性の評価には一般的な Tf･idf 値を用いた．描画する円の直径の計算には前述したように相対出現頻度を，語間の関係の強さを示す語間の描画距離は共起頻度の相対値の逆数を用いて計算した．

情報要求に基づき文書データベースから情報を抽出する際に指定できる抽出対象情報の単位と，語木描画時に指定できる条件項目を表１に示す．

表１．情報抽出および描画に指定できる条件

■データベースからの描画情報抽出単位
・文書単位 ・段落単位
■抽出した情報の語木描画時に指定できる条件
・描画対象情報の単位 （特定文書／特定段落／複数文書／複数段落） ・共起頻度下限値 ・語木描画階層数 ・共起分析対象 （文内共起／段落内共起／文書内共起）

４．システム動作確認テストと考察

4.1 準備

4.1.1 テストに用いた文書情報

動作テストに用いた文書情報は，前記したとおり特定技術分野の公開特許公報(以下，「特許広報」)を特許庁特許電子図書館よりダウンロードして用いた．

用いた公開特許公報の数を表２に示す．

表２．テストに用いた特許情報

入手先	特許電子図書館（ＩＰＤＬ）
入手日	２０１０年５月３１日
入手件数	４３２件（公開特許公報）

4.1.2 テスト用文書データベースの作成

特許公報中のテキストを形態素解析し，名詞，動詞，形容詞のみについて出現形，基本形及びその語が出現する文書,段落,文等に関する情報を格納し，文書データベースとし，情報源に用いた．

文書データベースに格納した文の数，段落数，語数を表３に示す．432 件の公開特許公報から約 200 万語が抽出でき，これら全てを格納した．

表 3. テストに用いた特許情報の段落数，文数，語数

段落数	109,258
文数	139,570
語数	1,996,342

4.2 テスト結果と考察

4.2.1 基本動作

動作確認テストは情報要求を「地球温暖化防止のための二酸化炭素ガスの分離除去」とし，また，語木の根語は，情報要求に基づき抽出された文書中の特徴語とし，共起度下限値等描画条件を変え，意図したとおり語木が表されるか，また，語木を形成する語間の関係が描画できるかの動作を確認した．

抽出された特定の文書について，共起度下限値を 11 とし描画した結果を図４に示す．語を囲む円の大きさで相対出現頻度を，語間の距離で共起頻度が表現できてはいるものの，条件を変えて繰り返し実施した結果，共起度下限値を小さくした場合，描画される語が増えることにより語木が「混んで」しまい，情報が読み取れない状態となった．

そこで，描画した語系列を一覧リストとして表示する機能を追加した．これにより，語木中で任意の語系列を選択すると，それに対応しリスト中の文字列も強調表示される．また描画した語木中で任意の語を指定すると，その語が含まれる全ての語系列を強調表示する機能や，語木を段階的に表示する機能等，操作に応じて描画内容を変化させる一連の機能が意図した通り表示ができることを確認した．

4.2.2「連想ゲーム」的(段階的)表示機能

「連想ゲーム的」的機能の実装例として，語をクリックすることで共起関係にある語を段階的に表示させる機能を，図５に根語「フロン」についての動作例で示す．

根語「フロン」(①)から順番に，「分解」(②)，「光」(③)，「反応」(④)を選択してゆく経過を示す．順番に語をクリックするたびに共起語が次の階層の候補語として赤色で表示される．候補語の中の特定の語を選択すると，選択された語以外の語が青色に変わるとともに，次の階層の候補を赤色表示する．この一連の操作でのシステムとユーザーとのインタラクションを通じ，連想促進を期待している．

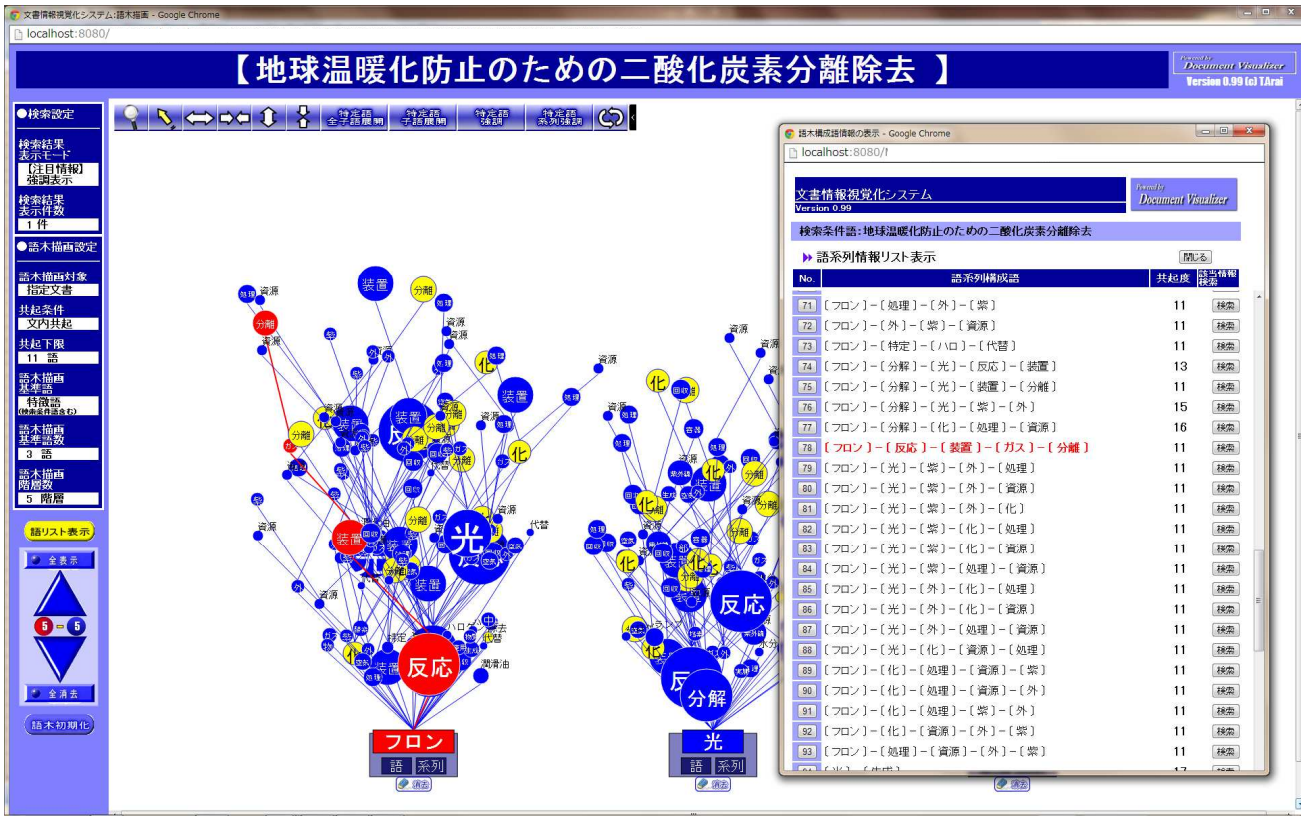

図４．語木の描画と語系列リスト表示例

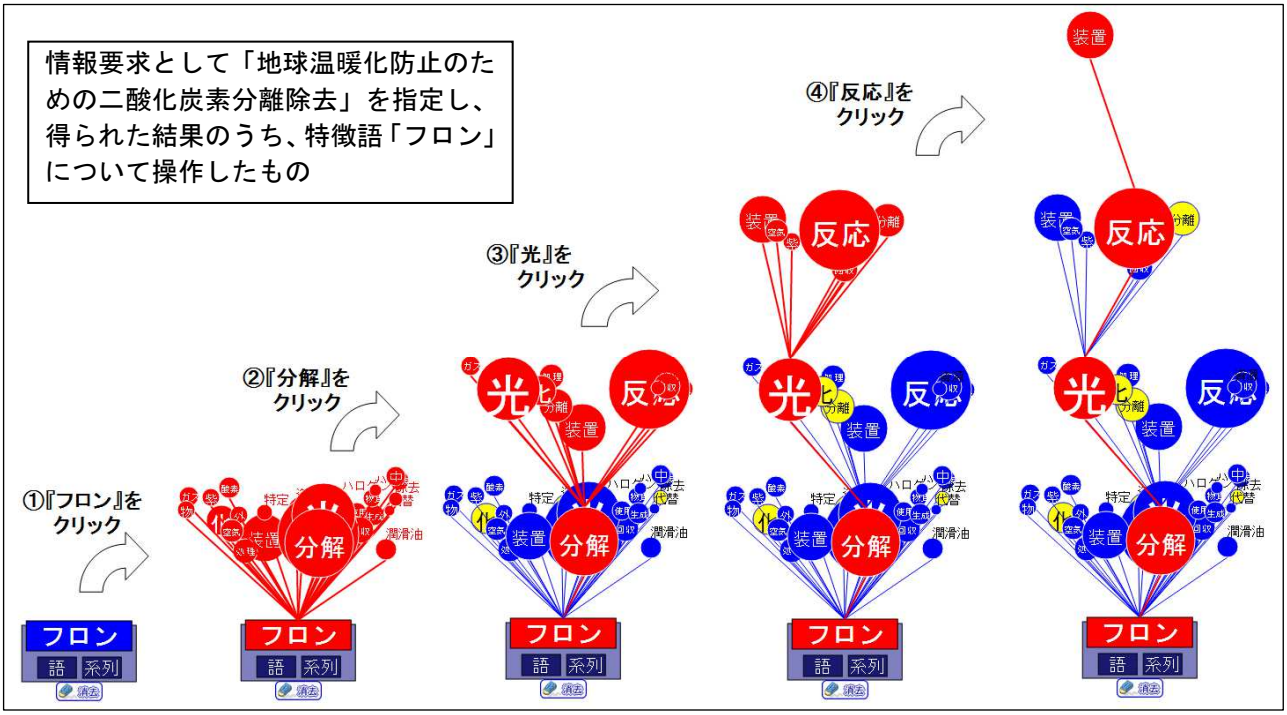

図５．「連想ゲーム」的語系列表示機能の例

4.2.3 語系列の選択と文書検索の連携機能

描画した語木中の指定した語系列に含まれる語群を検索条件として用いて文書データベースを再検索する動作例を図６に示す．図中赤で示した語系列を構成する語が自動的に検索条件語として用いられ(図中「指定語」欄)，該当する語を含む段落や文書が検索できている．この操作では，ユーザーが注目

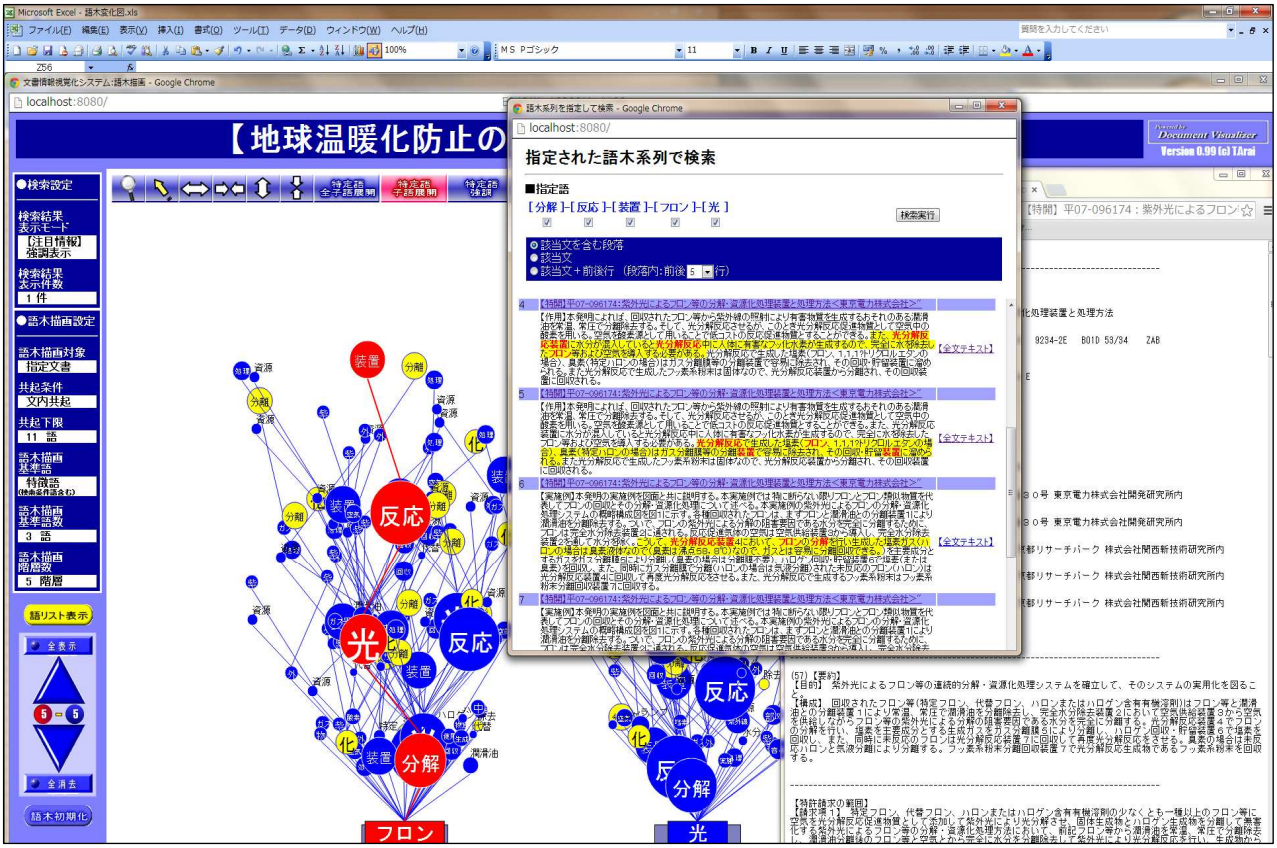

図６．語系列の選択と文書再検索の連携機能の例

した語系列構成語をそのまま検索条件として用いて再度文書データベースが検索できるので，思考の中断を極力抑えることが出来ると考える．

５．今後の展開

まずは提案システムの有効性の検証が必要である．しかしながら検証は，有効性の評価が検証に参加するユーザーの知識や経験に依存するため，検証方法を考案することは非常に困難なことが予想される．そこで，広範な分野の多くのユーザーに開放し，どのような場面でどのような経路の操作をしたときに有効な連想や気づきや発想が得られたかの情報をフィードバックしてもらうことにより，本システムが有効なケースを検証する方法での検証を考えている．

また機能にも改善を必要とする課題がある．まず，本格的な利用に向けては，パラフレーズや語のゆれへの対応が必要である． さらに，ユーザーにより有効な情報を提示するためには，情報源に用いる蓄積文書量を増やすことも必要である．一方，今回の試作および動作テストでは対象文書の分野を絞り，約 400 件程度の公開特許公報を用いたにもかかわらず，文書データベースに格納した語の数は約 200 万語となったことから，蓄積文書量を増やすと処理対象となる語の数が飛躍的に増大することが予想される．たとえば特許庁電子図書館所蔵の全特許情報を提案システムに格納し利用するとした場合，データベースに登録する語数を試算したところ約 550 億語となった．その場合，描画対象情報の抽出，分析時間がユーザーの要望に応えられなくなることが予想される．この問題への対処はデータベース処理を高速化することが必要と考えている．

参考文献

[１] 菰田文男，那須川哲哉，技術戦略としてのテキストマイニング，中央経済社，東京，2014.
[２] D.A.ノーマン，誰のためのデザイン，新曜社，東京，1990.
[３] 酒井邦嘉, 脳を創る読書, 実業之日本社，東京，2010.
[４] 野末道子，上田修一，”論文段落を対象とした日本語全文検索データベースの検索”，情報処理学会論文誌, Vol.1993，No.39，pp.9-16，1993.
[５] NHK(1969-1991) 「連想ゲーム」 NHK アーカイブス NHK Homepage http://cgi2.nhk.or.jp/archives/tv60bin/detail/index.cgi?das_id=D0009010143_00000 (2014，Feb，15)