2004－DBS－134　(I)　(26)
2004／7／13

Web 検索結果とその周辺情報の近似的内包表現とその視覚化

松生　泰典[†]　　是津　耕司[††,†]
小山　聡[†]　　田　中　克　己[†,††]

ユーザが Web 情報を使っていかに目的を達成できるかを重要視する，search experience という考え方が広まってきている．質問に対し適合性の高い検索結果をリストするだけでなく，検索結果から様々な情報を発見し，ユーザ自身が検索目的を確認しながら漸進的に検索を進められるようにすることが重要である．質問キーワードはユーザの興味を反映しているものの，必ずしも欲しい情報を正確に記述しているとは限らない．そこで本研究では，検索結果と関連性を持つページ集合，例えば，よく共起するキーワードや同じ主題，あるいは同じ見られ方（アスペクト）を共有するページ集合を，”質問の周辺情報”として調べ，ユーザが欲しいと思っている情報を特徴付ける他のキーワードを発見しながら質問を修正し，漸進的に検索を進めていくアプローチを提案する．本論文では，質問の周辺情報を近似的内包表現によって表し，質問修正の候補となるキーワード式を生成する方法について述べる．また，質問と生成されたキーワード式との関連性を可視化によって直感的に示すことにより，ユーザが検索の途中で迷子にならないようにする方法についても述べる．

Approximate Intensional Representation showing the Outline and Surrounding of Web Search Results and its Visualization

Yasunori MATSUIKE,[†] Kouji ZETTSU,[?] Satoshi OYAMA[†]
and Katsumi TANAKA[†,††]

The idea of “Search Experience” that focuses on users to achive their goals effectively by Web information becomes popular in these days. It is important not only to list search results highly adaptive to query, also to help users to search step by step without losing their goals. Query keywords reflect users’ interest, however they do not always reflect information that users want to get. In our works, we propose an approach to inquire page sets associated to search results, such as page sets that share the same frequently cooccurring keywords, subjects, and aspects, as “circumference information of queries”, to discover other keywords characterizing information users want to get, modify the query, and search step by step. In this paper, we propose a method to present circumference information by approximate intensional representation and to generate keyword formulas which can be candidates of query modification. We alose propose to help users not to lose their goals in the middle of information retrieval by visualizing relation between queries and keyword formulas.

1. はじめに

昨今，Web ページの爆発的な増加に伴い，その膨大な量の Web ページの中から得たい情報を求めて検索する際，ユーザが Web 情報を使っていかに目的を達成できるかを重要視する，search experience という考え方が広まってきている．この考え方を実現するためには，既存の検索エンジンのようにユーザの質問に対して適合性の高い検索結果をリストするだけでなく，検索結果を含めた Web ページ集合から様々な情報を発見し，ユーザ自身がその検索目的を確認しながら漸進的に検索を進め，ユーザの本来の目的であると考えられるページ集合を得るための質問を生成できるようにすることが重要である．

Web には多種多様なページが含まれ，検索エンジンでの検索結果の中には，ユーザが初めて見るページが圧倒的に多いと考えられる．求める情報を正確に特徴付けるキーワードを，ユーザ自身が判断することは困難である．質問キーワードはユーザの興味を反映してはいるものの，必ずしもユーザの求める情報を正確に記述しているとは限らない．そのため，ユーザの質

† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
†† 独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology

問キーワードから得られる検索結果の中の情報だけを見るのではなく，その周辺情報も合わせてユーザにさまざまな形で提示することで，ユーザ本来の目的により近い情報を得るための質問を生成することができると考える．

そこで本研究では，検索結果と関連性を持つページ集合，例えば，よく共起するキーワードや同じ話題構造であるページ集合，同じ見られ方（アスペクト）を共有するページ集合を，”質問の周辺情報”として調べ，ユーザが欲しいと思っている情報を特徴付ける他のキーワードを発見しながら質問を修正し，漸進的に検索を進めていくアプローチを提案する．本論文では，質問の周辺情報を近似的内包表現によって表し，質問修正の候補となるキーワード式を生成する方法について述べる．また，質問と生成されたキーワード式との関連性を視覚化によって直感的に示すことにより，ユーザが検索の途中でその目的を見失わない方法についても述べる．

本論文の２章では，本提案手法の関連研究について，３章では本提案手法の概要を述べ，４章では Web ページ集合の近似的内包表現について述べる．５章では Web 検索結果とその周辺情報からのキーワード式の生成について述べ，６章では，キーワード式の視覚化について述べる．７章でまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 KeyGraph

大澤ら[1] は，ある文章がどのような内容であるかを調べるために，その文章の中から重要であると考えられる単語を複数抽出してきて，それぞれの単語同士の関係を線でつないでグラフ表示するシステムを構築している．KeyGraph では一つの文章が対象であるので，Web ページであれば最大１つの Web ページという単位でグラフを生成するが，検索結果のページ集合に対してキーワードを抽出してそこから関連する周辺情報のキーワードも抽出してくる点が，本研究が異なっている点である．

2.2 TEOMA

Teoma[2] は米 AskJeeves が開発したサーチエンジンであり，データベースの中にある Web サイトをその内容に応じてどれだけ多くの“ 権威あるサイト ”からリンクされているかという情報に基づいて仕分けする．同じ内容を扱っているサイトの「コミュニティー」を構築し，その中からその分野のエキスパートを発見するという方法をとっているものである..　結果として表示結果のページに「Results」「Refine」「Resources」の 3 つの欄が表示される.「Results」の表示結果は一般的なサーチエンジンと同じように表示されるが，これは分別したコミュニティーのなかでサイトがどこに位置するのかをランキングしたデータに基づいたものである．

2.3 KartOO

KartOO[3] は，ビジュアルなインターフェースを持ったサーチエンジンの１つであり，キーワード検索を行う際に，クラスタリングを行ってクラスタを作成し，それを地図のように配置して表示するものである．各検索結果から得られたクラスタはアイコンで表示され，そのアイコンにはクラスタを代表するキーワードがつけられている．

2.4 Web ページのアスペクトの発見

是津ら[4] は，Web ページについてのアスペクトを提案している．このアスペクトとは，ある一つの Web ページが外部からどのように参照されているかという評判や役割を表し，Web を一つの社会と見なした際の，Web ページの ”社会的評価 ”ととらえることができる．Web ページがどのような容から参照されているかを表す部分として，そのページのリンク元のページのリンクアンカー周辺をコンテキストとして抽出してクラスタリングする．このとき，各コンテキストについて文脈貢献度と呼ばれる評価値を計算し，コンテキストの抽出範囲を決定する．

クラスタリングされたコンテキスト集合それぞれについて，その中に含まれる各 Web コンテンツの典型性という評価値を計算し，その典型性の値の大きな Web コンテンツを，そのクラスタを代表する Web コンテンツとしてその Web ページのアスペクトとしている．

Web ページ一つについてのアスペクトを求めるのではなく，検索結果の周辺情報としてリンク元のページ集合を取得し，検索結果の近似的内包表現であるキーワード式として表現する点が，本研究が異なっている点である．

3. 本提案手法の概要

3.1 検索結果とその参照ページ集合の近似的内包表現による Web 情報検索支援

筆者ら[5] はこれまでに，検索結果とその参照ページ集合それぞれの概要を表すために，検索結果とその参照ページ集合の中にそれぞれ含まれるキーワードを複数組み合わせ，それらのキーワードから得られる検索結果のページ集合によりはじめの検索結果を近似的に表すシステムを提案した．複数のキーワードの OR を

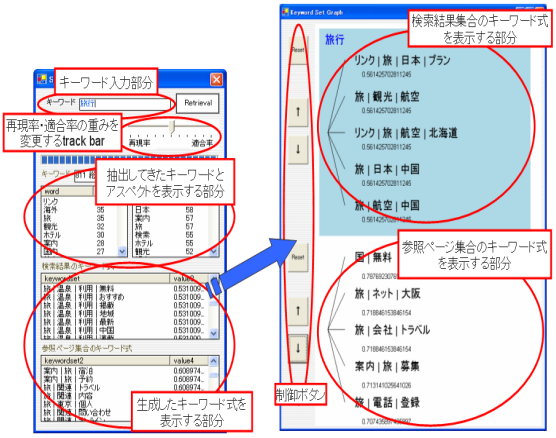

図 1 プロトタイプ

とったものをキーワード式とし，その評価関数として再現率と適合率を足し合わせたものを用いている．生成したキーワード式は，評価関数の高いものから順に提示し，そのキーワード式に含まれるキーワードを選ぶことで，そのキーワードを含んだキーワード式を複数提示するようにする．また，次の図１のようなプロトタイプを作成し，キーワード式の提示を行った．

キーワード式の提示方法が，図１のように複数同時に表示するようにしている．

本論文では，この研究を踏まえて，検索結果の概要を表すだけでなく，検索結果から得られる周辺情報も含めたページ集合をさまざまなつながりからとらえ，それぞれについて近似的内包表現で表し，ユーザの目的とするページ集合により近いページ集合を得る質問を生成するための情報としてユーザにわかりやすい形で視覚化して提示する．

3.2 近似的内包表現の生成のその視覚化の手順

本提案手法では，まず，最初の質問から得られた検索結果についての周辺情報として，検索結果内のキーワードについての共起つながりのページ集合，検索結果のページ集合と話題構造つながりのページ集合，検索結果のページにリンクをはっているアスペクトつながりのページ集合の３つのつながりを取り上げる．それぞれの周辺情報について，そのページ集合を近似的内包表現で表し，その内包表現の満たしている条件であるキーワード式を生成する．生成したキーワード式は，「共起つながり」，「話題構造つながり」，「アスペクトつながり」の情報として，最初の質問との関連性を直感的に理解するために，そのつながりを表す語とともに視覚化する．それぞれのつながりの仕方について，視野を変えることで，ユーザにさまざまな視点から検索結果とその周辺情報から知識を得ることを目的とする．

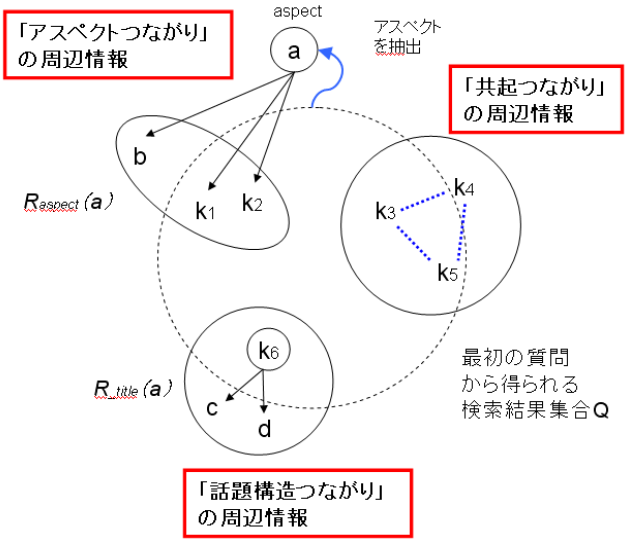

図 2 提案手法の概要

本提案手法の手順を下に示す．

（1） ユーザの最初の質問を Google に入力し，検索結果を取得する．

（2） 次の３つのつながりという観点から，キーワード式を生成する．

- 得られた検索結果集合について，共起関係つながりとして，次のようにキーワード式を生成する．
  - 最初の質問から得られた検索結果のページ集合から語を抽出し，キーワードとする．
  - 抽出してきたキーワード同士の共起度の高いものを組み合わせ，キーワード式を生成する．
- 得られた検索結果集合について，話題構造つながりとして，次のようにキーワード式を生成する．
  - 最初の質問から得られた検索結果の中から抽出されたキーワードのページ内での出現場所を調べる．
  - 同じ話題構造を持つページ集合を取得し，その中のキーワードを用いてキーワード式を生成する．

- 得られた検索結果集合について，アスペクトつながりとして，次のようにキーワード式を生成する．
  - 検索結果のページそれぞれのリンク元のページを取得する．
  - リンク元のページから，リンク先のページの「見られ方」を表すアスペクトを抽出する．
  - 同じアスペクトを持つページ集合を取得し，その中のキーワードを用いて，キーワード式を生成する．

（3）以上の3つのつながりでのキーワード式を，ユーザが直感的に理解できるように視覚化する．

4. Web ページ集合の近似的内包表現

4.1 Intensional Representation of a Data Set

Uesima ら[6] は，ある定型データベースの部分集合を簡潔に表すキーワード集合をキーワード式として抽出する手法を提案している．このとき，キーワード式は目的集合全体を補完している度合いの差異によって，複数存在する．従ってこのキーワード式は近似的である．そのため，キーワード単位で目的集合の評価関数をそれぞれ計算し，キーワードの関係を木構造で表現し，階層ごとに新しいキーワードをつけたしていきながらキーワード式の評価値を求め，最も評価値の高くなるキーワード式を生成し，そのキーワード式でデータベースの部分集合の概要を表す手法を提案している．

キーワード式の評価関数には再現率と適合率を用い，その両方が大きくなるようなキーワード式を，目的集合の概要をより高い精度で表現しているキーワード式としている．キーワード式は，評価値が高いキーワードから順に組み合わせていくことで生成する．新しいキーワードをキーワード式に追加する際，評価値が最も高いものを選ぶ．

本研究で述べる近似的内包表現は，このキーワード式生成の手法に基づいている．

4.2 Web 検索結果の近似的内包表現

検索結果などのページ集合を扱う際，既存の検索エンジンのようにユーザの質問に対して適合性の高い検索結果をリストするだけでは，どのような情報が得られているのかを理解することは，その検索結果の膨大さから，困難である場合が多い．そこで，そのページ集合がどのような情報を含んでいるのかを簡単に表すために，複数のキーワードで表すこととする．

本研究では，複数のキーワードの組み合わせをキーワード式とする．このキーワード式の位置づけとして，キーワード式に含まれるキーワードそれぞれについての検索結果のページ集合を合わせたものによって，検索結果を近似的に内包表現で表す．

あるページ集合 Q に含まれるキーワード k_1，k_2，$\cdots$，k_n を組み合わせたキーワードの集合として，キーワード式「 k_1 ｜ k_2 ｜ $\cdots$ ｜ k_n 」を考える．キーワード k_n での検索結果を $\mathrm{R}(k_n)$ とし，$\mathrm{R}(k_1)$，$\mathrm{R}(k_2)$, $\cdots$，$\mathrm{R}(k_n)$ の OR 集合を R とする．このとき，Q の近似的内包表現とは，

$$\mathrm{Q} = \{\ x \mid x\ in\ R\ \} \quad (1)$$

$$\mathrm{R} = R(k_1) \cup R(k_2) \cup \ldots \cup R(k_n) \quad (2)$$

のように Q を表すものとする．ここで，x は1つのページを表すものとする．

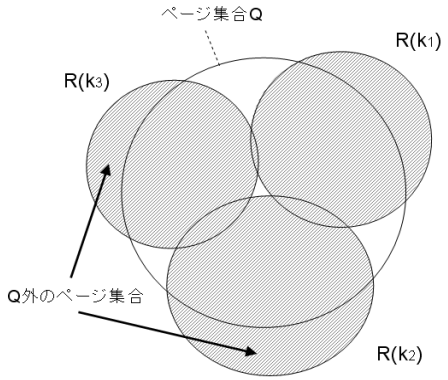

図 3 近似的内包表現

近似的内包表現で検索結果を表した場合，図3のように，R に含まれるページの中には，最初の質問を含まないページ，つまり Q に含まれないページも含まれる．そのため，検索結果内のページだけでなく，検索結果外のページも扱うことになる．このことから，検索結果だけでなくその周辺情報を含んでおり，ユーザにとって新たな知識を発見するてがかりになると考えられる．

4.3 キーワード式の評価

あるページ集合を近似的内包表現の満たす条件としてのキーワード式で表すとき，ページ集合内のページに含まれるキーワードの組み合わせとしてのキーワード式は，複数通り生成できる．しかしその際，元のページ集合をより的確に表しているキーワード式のほうが，ユーザに提示するものとしては有用であると考える．そこで，キーワード式の評価として，次の評価

関数 Fpr を求め，その値を評価値とする．

$$\mathrm{Fpr} = (1-w) * \frac{|R \cap Q|}{|R|} + w * \frac{|R \cap Q|}{|Q|} \quad (3)$$

ここで，Q は元のページ集合，R は近似的内包表現に含まれるページ集合とする．R が Q に含まれているページを多く含んでおり，なおかつ Q に含まれないようなページをあまり含んでいないものであれば，R は Q をより正しく表現できていると考えられる．そこで，R の Q に対する再現率と適合率を足し合わせるものを使用している．この値が高いものほど，より的確に元のページ集合を表している近似的内包表現が表すページ集合であるといえる．

5. Web 検索結果とその周辺情報からのキーワード式の生成

ユーザが Web 検索を行う際，ユーザの質問により得られた検索結果だけでなく，その検索結果をいろいろな角度からとらえた周辺情報を得ることができれば，検索結果には含まれていない知識も得ることができ，そこからユーザの本来の目的により近い情報を含むページ集合を得られる知識を次の質問とすることができると考えられる．その際，周辺情報はさまざまなものが考えられるが，最初の質問はユーザの興味を表しているものであると考えられるため，検索結果の周辺情報として，その最初の質問と何らかのつながりのあるものを抽出する．そのつながりとして，次のようなものを考える．

- 共起つながりの周辺情報
- 話題構造つながりの周辺情報
- アスペクトつながりの周辺情報

この３つのつながりそれぞれについてキーワード式を生成し，それぞれの周辺情報がどのような情報を含んでいるのかを知ることを目的とする．それぞれのつながりについて，キーワード式の生成の仕方を，つながりの特性をよく反映したものとすることによって，それぞれが検索結果のページ集合に対して異なる位置づけであることから，ユーザに提示する情報を周辺情報を含めた検索結果をさまざまな角度からとらえた，マルチモーダルな視点から検索結果をとらえ，そこから知識を得ることができると考える．

5.1 共起つながりの周辺情報

ユーザの最初の質問から得られた検索結果を近似的内包表現で表すと，その内包表現の満たす条件であるキーワード式は，その中に含まれるキーワードそれぞれで検索した結果のページ集合を端的に表すものであると言える．キーワード式の表しているページ集合は，最初の質問の検索結果外のページも含んでいる．そこで，その検索結果外のページを，検索結果の周辺情報としてとらえ，検索結果とその周辺情報を的確にとらえるキーワードを組み合わせてキーワード式を生成する．ここで，検索結果を含めたページ集合を簡単に表すキーワード式として，キーワード式に含まれるキーワード同士の，ページ集合内での共起関係に注目する．多くのページで共起する２つの語は，深い関係があると考えられる．つまり，共起関係の強いキーワード同士を組み合わせて生成したキーワード式は，検索結果の一部分をより詳細な形で表していると考えられる．そのため，共起関係でのキーワード式を複数見せることで，検索結果とその周辺情報をクラスタリングした結果を見ているように，さまざまな分野を詳細化して見ることができると考えられる．例として，検索結果 Q を共起度の高い３つのキーワードからなるキーワード式「k_1 ｜ k_2 ｜ k_3」で表現した場合，図４のようになる．

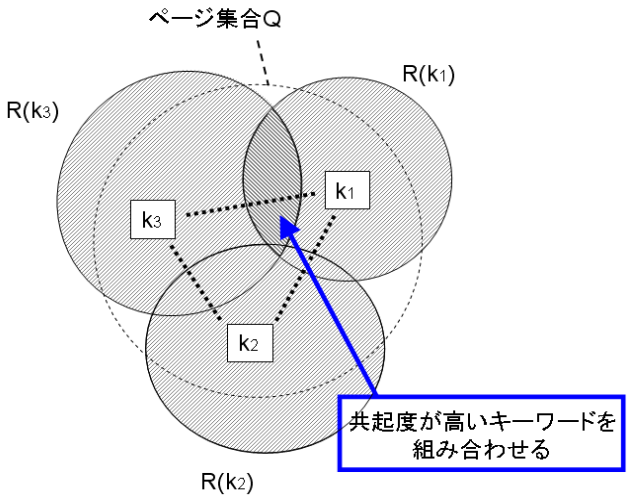

図 4 共起つながり

まず，最初の質問での検索結果の上位１００件を抽出してきて，そのテキストを形態素解析する．形態素解析には，茶筅[7] を用いる．その中から名詞と固有名詞を取り出し，キーワード式を生成するためのキーワードの候補とする．その中で，より多くのページに含まれている語を優先的にキーワード式に含めるキーワードとする．その中で，語と語の共起度の高いものを組み合わせて，複数のキーワード式を生成する．キーワード式の評価としては，（３）の評価関数を用

いる．

5.2 話題構造つながりの周辺情報

検索結果のページそれぞれから共起つながりの場合と同様にキーワードを抽出してきた際，同じキーワードであっても，ページ内での位置によりその使われ方や意味が違ってくると考えられる．そこで，話題構造としてページ内でタイトルに含まれる主題語である場合とテキストに含まれる内容語である場合を，話題構造からのつながりととらえてキーワード式を生成する．ここで，例えば主題語として現れる頻度の高いキーワード k をキーワード式に含む場合，k を主題語に含むページの内容語は，話題構造つながりで重要であると考えられる．

例えば，図５のように，検索結果内のキーワード k に着目した場合を考える．k がページのタイトルによく含まれている場合，そのページにおいて k は主題語であると言える．k を主題語として含んでいるページ集合において，そのテキスト内によく含まれているキーワードは，k についての詳細な内容を表しているものと考えられる．また，k がページのテキストによく含まれている場合，そのページにおいて k は内容語であると言える．k を内容語として含んでいるページ集合において，そのタイトルによく現れるキーワードは，k という語の上位概念を表していると考えられる．

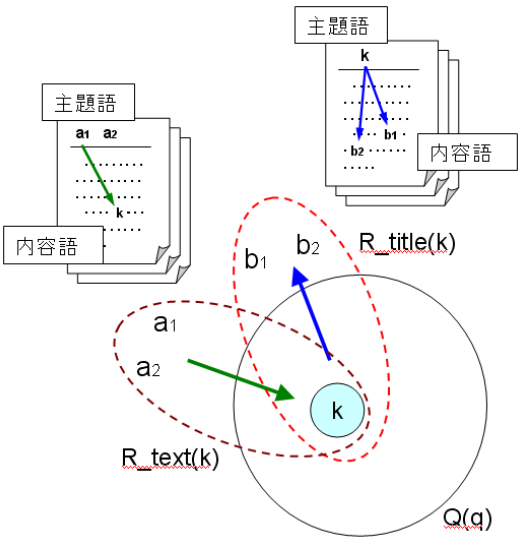

図 5 話題構造つながり

検索結果のページ集合に含まれるキーワードに対して，そのキーワードが主題語として現れているページ集合 R_title(k) と，そのキーワードが内容語として現れているページ集合 R_text(k) をそれぞれ取得してくる．R_title(k) から内容語として出現頻度の高いキーワードを組み合わせて，キーワード式を生成する．また，R_text(k) から主題語として出現頻度の高いキーワードを組み合わせて，キーワード式を生成する．

5.3 アスペクトつながりの周辺情報

アスペクトは，Web ページのさまざまな「見られ方」を表しているものである[4]．検索結果のページ集合がどのような見られ方をしているのか，というものをアスペクトつながりとして見ることができれば，検索結果の位置づけとその評価を知ることができると考えられる．

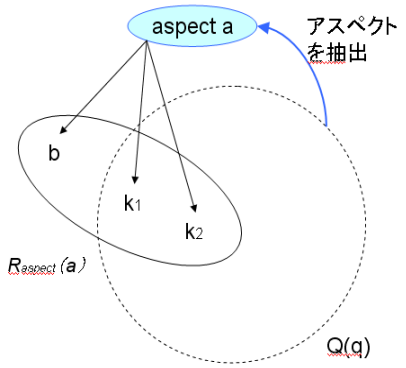

図 6 アスペクトつながり

検索結果のページ集合について，１つ１つのページにリンクをはっているページを取得してくる．そのリンク元のページ集合を特徴づける語として，アスペクトを抽出してくる．このアスペクトによって，検索結果のページ集合がさまざまなとらえられ方をしていると考えられる．アスペクトは複数取得できるため，検索結果のページ集合には複数のとらえられ方があると言える．そこでそのとらえられ方ごとに，つまり抽出してきたアスペクトごとに，そのアスペクトが表している検索結果内のページ集合と，さらに検索結果に含まれていないが，同じように参照されているものとして，同じアスペクトを持つページ集合を取得してきて，１つのページ集合とし，そのページ集合内の情報を，検索結果と「アスペクトつながり」であるものとする．１つのアスペクトに対して，１つのページ集合が対応する．そこで，アスペクトごとに対応しているページ集合を近似的内包表現で表し，その中に含まれるキーワードからキーワード式を作成し，「アスペクトつながり」のキーワード式とする．

6. キーワード式の視覚化

これまでの章では，検索結果とその周辺情報のつながりとして３つの手法を用いて近似的内包表現でページ集合を表すことについて述べた．この章では，検索結果とその周辺情報から生成されるキーワード式を，ユーザがその関係なども含めてとらえやすくするために視覚化することについて述べる．

6.1 つながりごとの視覚化

6.1.1 共起つながりのキーワード式の視覚化

共起つながりから得たキーワード式は，それぞれその検索結果とその周辺情報の概要をよく表していると考えられる．そのため，それぞれのキーワード式を図４のようにキーワード同士の共起度も表せるようにユーザに提示する．

6.1.2 話題構造つながりのキーワード式の視覚化

話題構造では，そのつながりとして，例えば「主題語つながり」,「内容語つながり」といったように，検索結果から抽出してきたキーワードをそのつながりを示すキーワードとすることができる．そのため，図７のように，最初の質問と生成したキーワード式を，そのつながりを示すキーワードでつないで表示する．

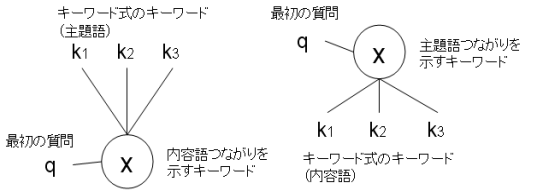

図 7 共起つながり

6.1.3 アスペクトつながりのキーワード式の視覚化

検索結果のアスペクトとは，そのページ集合の Web 上での「見られ方」を表すものである．抽出したアスペクトが検索結果とその周辺情報を参照しているということであるので，図８のように提示する．

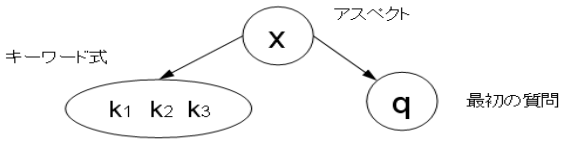

図 8 共起つながり

6.2 視野の変更

生成したそれぞれのつながりごとのキーワード式は，検索結果の周辺情報として扱っている情報も異なり，また検索結果に対してどのような性質をもつのかも異なる．そのため，３つのつながりでのキーワード式をユーザの興味に合わせて表示を変えることのできるようにする．また，提示する情報が多くなりすぎないように，一部分のみを表示し，その視野を変えていけるようにする．このような考え方は，Focus and Context View[8)9)] と呼ばれている．

7. まとめと今後の課題

本論文では，検索結果とその周辺情報をさまざまな視点からとらえて近似的内包表現を用いて表し，その内包表現のもつ条件であるキーワード式を生成し，そのつながりごとにキーワード式を直感的に理解するための視覚化の手法を提案した．これにより，ユーザの検索目的を確認しながら，検索結果とその周辺情報からユーザの目的を正しく特徴付ける語を発見しながら，質問を修正し，漸進的に検索を進めることができると考えられる．

今後の課題として，システムのプロトタイプの作成と，それぞれのつながりごとのキーワード式がどのようなものとして利用できるかを調査していくことを考えている．

謝　　辞

本研究の一部は，平成１６年度科研費特定領域研究（２）「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：16016247, 代表：田中克己），および２１世紀ＣＯＥプログラム「知識社会基盤構築のための情報学拠点形成」による．ここに記して謝意を表します．

参　考　文　献

1) 大澤幸生, Nels E. Benson, 谷内田正彦： KeyGraph:単語共起グラフの分割統合によるキーワード抽出，電子情報通信学会論文誌 J82-D1, No.2, pp. 391–400, 1999.
2) http://www.teoma.com/
3) http://www.kartoo.com/
4) 是津 耕司　木俵 豊　田中 克己　”Web ページのアスペクトの発見 ”データベースと Web 情報システムに関するシンポジウム (DBWeb2003) 論文集, 情報処理学会シンポジウムシリーズ, Vol.2003, No.18, pp.93-100, 2003 年 11 月
5) 松生 泰典, 是津 耕司, 角谷 和俊, 田中克己 : 検索結果とその参照文脈の近似的内包表現による Web 情報検索支援，第 15 回データ工学ワークショップ (DEWS2004), 2004 年 3 月

6) Shinichi Ueshima, Kazuhiro Ohtsuki, Jun-ya Morishita, Qing Qian, Hiroaki Oiso, Katsumi Tanaka: Incremental Data Organization for Ancient Document Databases. DASFAA 1995: 457-466
7) 奈良先端科学技術大学松本研究室　茶筅ホームページ：
http://chasen.aist-nara.ac.jp/index.html
8) Sarkar, Manojit and Marc H. Brown, “graphical fisheye views,” Comm. of the ACM, Vol.37, No.12, pp.73-83, 1994.
9) Furnas , G. W., “Generalized Fisheye Views,” Proc. ACM SIGCHI ’86 Conference on Human Factors in Computing Systems, pp.16-32, 1986.