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

松生　泰典† 　是津　耕司†,††† 　角谷　和俊††† 　田中　克己†††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 独立行政法人通信総合研究所　〒 184-8795 東京都小金井市貫井北町 4-2-1
††† 京都大学大学院情報学研究科社会情報学専攻　〒 606-8501 京都市左京区吉田本町
E-mail: †{matsuike,zettsu}@dl.kuis.kyoto-u.ac.jp, ††{sumiya,ktanaka}@i.kyoto-u.ac.jp

あらまし　Web 情報検索において，膨大な量の検索結果を全て閲覧することは困難である．このとき，選択したキーワードについての検索結果とそれを参照しているページ集合をわかりやすい形で提示することで，キーワードについての情報を簡潔に理解することができ，知識発見と質問修正につながると考えられる．本稿では，あるキーワードでの検索結果と，その検索結果のページ集合を参照しているページ集合の概要を表す近似的内包表現として，複数のキーワードからなるキーワード式を生成・提示することで，あるキーワードに関連している情報を簡潔に提示するシステムを提案する．

キーワード　情報検索，質問修正，知識発見，可視化

Supporting Web Information Retrieval by Approximate Intensional Representation of Search Results and Their Referential Contexts

Yasunori MATSUIKE†, Koji ZETTSU††,†††, Kazutosi SUMIYA†††, and Katsumi TANAKA†††

† School of Informatics, Kyoto University　Yosidahonmati, Sakyou-ku, Kyoto,606–8501 Japan
†† Communications Research Laboratory　4-2-1 Nukui-Kitamachi,Koganei,Tokyo,184-8795 Japan
††† Department of Social Informatics,Graduate School of Informatics,Kyoto University
Yosidahonmati,Sakyou-ku,Kyoto,606-8501 Japan
E-mail: †{matsuike,zettsu}@dl.kuis.kyoto-u.ac.jp, ††{sumiya,ktanaka}@i.kyoto-u.ac.jp

Abstract　In Web information retrieval, it is difficult to peruse a huge quantity of all retrieval results. In this case, it is thought that we can understand the information about a keyword and it leads to knowledge discovery and query modification by showing the retrieval result of a selected keyword and their referential page set intelligibly. In this paper, we propose the system which shows the information about a keyword by generating and showing a keyword formula which consists of two or more keywords as approximate intensional expression presenting the outline of retrieval result in a keyword and their referencial page set.

Key words　information retrieval , query modification , knowledge discovery , visualization

1. はじめに

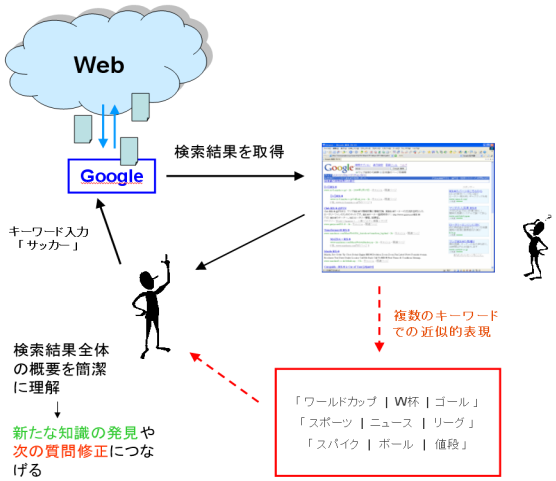
現在，検索エンジンはさまざまなものが開発され，実際にあらゆるユーザに使われている．中でも，Google [1] などが有名である．膨大な量の情報を含んでいる Web から，得たい情報を簡単に引き出してくるためには，検索エンジンは必要不可欠なものとなっている．

インターネットの普及により，さまざまなユーザが Web 情報検索を行うようになっている．そのため，情報検索の手法，得たい情報の種類やその提示の仕方などに対するユーザのニーズも多様である．ここで，既存の検索エンジンにおける検索は，情報探索には向いているが，情報探索には不向きである．例えば，Google などの検索エンジンでは，あるキーワードについての検索結果をそのページのタイトルと文章の一部分のみを箇条書きで表示しており，本当に欲しい情報がのっているページかどうかは，実際にそのページを開いて閲覧しなければ判断できないことが多い．

既存の検索エンジンに関して，次のような問題点がある．

- Web ページの増加による，検索件数が膨大な量である
- 検索のためのキーワードの決定が困難である

また，Web 探索において，さまざまなニーズが考えられる．例えば，「前に行ったレストランの名前を思い出したい」といっ



た場合や, 閲覧中の Web ページで, どのようなものなのかを知らないがその時に興味を持った言葉が出てくる場合などがある. このようなとき, そのレストランの場所や, レストランで見かけた物やメニューなど, 検索対象に関連する断片的なキーワードや, 閲覧中の Web ページ内の言葉を検索エンジンに入れてみるが, 思うように欲しい情報が出てこなかったり, どの情報が有用なのかがわからないときが多い.

このような場合, 絞込み検索のように詳細な情報のみを見ていくよりは, 選択したキーワードと関係のあるさまざまな情報を自動的に提示した上で, その中からよりユーザ自身の興味のある内容のものを選んだり, そこから質問をより詳細なキーワード集合に置き換えて再検索するほうが, 新しい知識を発見することができたり, 検索のために初めに選択したキーワードを, よりユーザの嗜好に沿った情報を取得するための質問に修正することができると考えられる.

そこで本論文では, 質問のために初めに選択したキーワードについての情報を取得して, その内容を近似的な形で表現したものをグラフ化し, 新たな検索のための質問を得るための情報を提供するシステムを提案する. システムで扱う情報は, 検索結果と検索結果外からの情報として,

- 検索結果のページ集合
- 検索結果のページを参照しているページ集合

の両方を用い, それらをグラフ表示する. ここで, 検索結果と参照ページ集合の中に含まれるキーワードそれぞれの違いを明確にするために, 参照ページに含まれるキーワードをアスペクトと呼ぶことにする. それぞれの情報を, 複数のキーワードの集合であるキーワード式で表現する.

本論文の２章では, 本提案手法に関係する基本的事項や関連研究について述べ, ３章では近似的内包表現について述べ, ４章ではシステムの概要と, その具体的手法としてページからのキーワードの抽出, キーワード式の作成, グラフの生成についてそれぞれ述べる. ５章ではプロトタイプの実装と評価について述べ, ６章でそのまとめと考察を述べる.

2. 関 連 研 究

2.1 Focus + Context View

Web 検索において, 一度に提示する情報の量はブラウザやモニタのサイズなどに制限があるのに加え, ユーザが一度に把握できる情報の量にも限度がある. そのため, 情報の量が膨大である場合は, その一部分だけを見せ, その上でユーザがさらに見たいと思った情報を選んで順に見ていくことで, 情報を効率よく取得することができると考えられる.

情報をキーワード集合のグラフなどの簡潔に表示したものからその中の一つを選択すると, その選択したものを中心とした表示に変えるようにする手法がある. このように, 選択したものを中心として, その周辺だけを見せる可視化のアプローチは, ”Focus + Context View ”と呼ばれている. この例としては, FishEye View [2] [3] などが有名である.

2.2 Webbrain

Webbrain [4] とは, とりあげたキーワードの関連分野をクラスタリングして視覚的に編集して表示することによって, 分野の中でのそのキーワードの位置づけを理解するのを支援することを目的としたシステムである. 取り上げたキーワードの上位概念, 下位概念, 類似概念にあたるキーワードを抽出してきて, 取り上げたキーワードを中心としてグラフ表示する. キーワード一つ一つではなく, ページ集合の近似的な概要を表すキーワード式をグラフ化して表示する点が, 本研究が異なっている点である.

2.3 Intensional Representation of a Data Set

Uesima ら [5] は, あるデータベース内の情報の中のある目的集合の概要を端的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くことで, キーワード式に含まれているキーワードの評価関数をあわせたものを求め, それを各キーワード式の評価値としている.

評価関数が高いキーワードを選んでいき, それらをキーワード式とする. 新しいキーワードをキーワード式に含める場合には, 新たなキーワード式の評価値が最も高いものを選ぶ. 評価値がそれ以上大きくならないようになった場合, それらのキーワード集合を, 目的集合の概要を最もよく表しているとみなすことができると考えられる. 評価関数には再現率と適合率の考え方を用い, 再現率と適合率をできる限り高くす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.

本研究では, 検索結果集合, およびそれらを参照する周辺ページの内容を概観する目的で, 同様の概念に基づく内包的表現の生成を行っている.

2.4 Web ページのアスペクトの発見

是津らは，Web ページについてのアスペクト [6] 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内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．コンテキストの抽出範囲の例を，図 1 に示す．

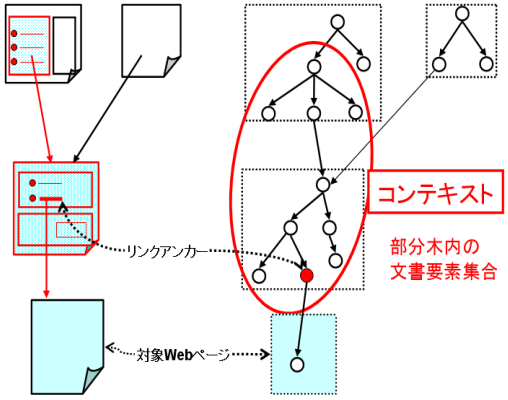


図 1　コンテキストの抽出範囲

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，あるキーワードについての検索結果集合全体についてのアスペクトを抽出し，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. Web ページ集合の近似的内包表現

検索エンジンから取得した検索結果のページの数は膨大であることが多く，その全てをみることは事実上不可能である．また Google などの検索エンジンでは，その膨大な量の検索結果を単にページごとにタイトルとその内容の一部分を文章で箇条書きとして表示するだけであり，どのような情報が検索結果に含まれているのかを全て理解するのは非常に困難である．しかし，検索結果の数ページだけを開いて閲覧するだけでは，検索結果のページ全体がどのような情報を含んでいるかを知ることはできない．そこで，検索結果の上位１００件だけを取り出してきて，その中に含まれているキーワードを用いて検索結果のページ集合を近似的に表すキーワード式を生成する．このとき，一つ又は複数のキーワードの集合を提示する情報とする．このキーワード集合は検索結果全体を完全に表しているとは限らないが，近似的にある程度検索結果を含むページ集合を表しているとして，検索結果集合の近似的内包表現となっているこれらのキーワード集合をキーワード式と呼ぶことにする．

例えば，あるキーワードでの検索結果を Q，Q の中に含まれるキーワードを k_i，キーワード k_i での検索結果を R(k_i) と表すことにする．このとき，Q を３つのキーワードによって近似的内包表現するとした場合，図２のように，３つのキーワードの検索結果それぞれによって，それぞれ Q の一部を表すことができる．このとき，近似的内包表現として多くのキーワードの検索結果集合全てを扱った場合，より大きな集合となるため初めに選択したキーワードの検索結果集合 Q をより確かに表現できるようになると考えられるが，その分 Q に含まれないページも多くなるため，精度は一概に大きくなるとは言えない．このとき検索結果集合 Q をキーワード式によって表現するキーワードの組み合わせは複数通り考えられる．今回，キーワード式の評価の基準として，再現率と適合率を用いる．

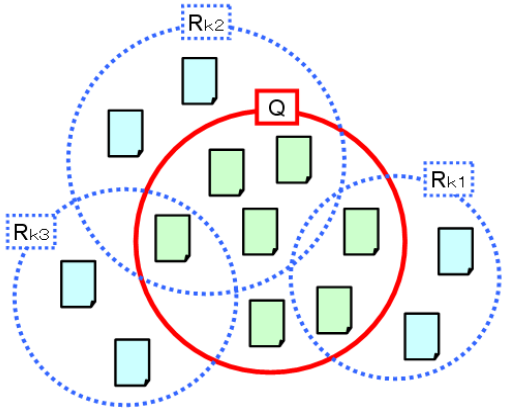


図 2　検索結果の関係

キーワード式の生成の例として，目的集合の中にキーワード k_1,k_2,k_3 が含まれていたとして，目的のページ集合の中に含まれるキーワードを用いてこの目的集合の近似的内包表現によって表すとき，下のように複数のキーワード式が挙げられる．

- 「k_1 ｜ k_2 ｜ k_3」
- 「k_1 ｜ k_2」
- 「k_1 ｜ k_3」
- 「k_2 ｜ k_3」

例えば,「k_1 ｜ k_2 ｜ k_3」のキーワード式は，Q = x — x in (R(k_1) U R(k_2) U R(k_3)) のように Q を近似するものとする．複数のキーワード式を生成し，これらの中から評価関数に基づいて最適なものを見つけ出す．上記の例の場合，例えばキーワード式「k1 ｜ k2 ｜ k3」を生成したとする．このキーワード式の評価関数は次のようなものとする．

$$f = (1 - z)\frac{|R \cap Q|}{|R|} + z\frac{|R \cap Q|}{|Q|} \quad (1)$$

$$R = |R(k\_1) \cup R(k\_2) \cup R(k\_3)| \quad (2)$$

ここで，z は再現率と適合率を計算するための重みであり，0¡z¡1 を満たすものである．評価関数が大きい値を示す複数のキーワードを，初めに選択したキーワードでの検索結果集合の

近似的内包表現としてその概要をよく表しているキーワード式とする．

4. 検索結果とその参照ページの近似的内包表現の生成と提示

4.1 提案手法の概要

提案手法の一連の流れは，以下に示すものである．

（1） ユーザが検索のためのキーワードを入力する
（2） ユーザが選択したキーワードを Google に入力し，検索結果を取得する
（3） 検索結果のページそれぞれからキーワードを抽出し，その検索結果内での重要度（後述）を求める
（4） 重要度の高いキーワードを様々に組合せ，各キーワード集合に対する評価関数の値をそれぞれ計算する．
（5） 評価関数の高いキーワードの集合を，検索結果の近時的な概要を表すキーワード式とする
（6） 検索結果のリンク元のページ集合それぞれから，コンテキスト (コンテキスト) を取得する
（7） 各コンテキストの中に含まれる代表的なキーワード集合（アスペクト）を抽出する（後述）
（8） リンク元のページから抽出してきたアスペクトについて，参照ページ集合についてのキーワード式を求める
（9） 検索結果についてのキーワード式と，参照ページ集合についてのキーワード式をグラフ化して表示する
（10） 表示したグラフについて，以下を繰り返す

- グラフの中のキーワード式に含まれるあるキーワード又はアスペクトを選択する
- 新しいキーワード又はアスペクトを選ぶたびに，その言葉を含むキーワード式を中心にそれぞれ表示する範囲を変える

4.2 ページ集合からのキーワードの抽出

4.2.1 検索結果からのキーワードの抽出

初めに選択したキーワードを Google に入力し，その検索結果の上位１００件を取り出してきて，そのページのテキストごとに文章全体を形態素解析する．このとき，形態素解析には茶筅 [7] を利用する．その解析結果から一般名詞と固有名詞を抽出し，キーワードとする．抽出してきたキーワードは，ページ集合の近似的な概要を表すキーワード式を生成する候補となる．あるページ集合の近似的な概要を表すために重要であると考えられるキーワードは，一つのページに多く含まれている言葉よりも，ページ集合の多くのページに含まれている言葉のほうが，重要であると考えられる．

ここで，検索結果内に含まれるキーワード k について，そのキーワードを含んでいるページの総数を document frequency と呼び，df(k) で表す．また，検索結果のページ数を N とする．キーワードそれぞれについての重要度を W(k) として，次のように定義する．

$$\mathrm{w(k)} = \frac{df(k)}{N} \tag{3}$$

検索結果の中から抽出したキーワードのそれぞれについてこの w(k) を求め，その値の大きなものから順にキーワード式に含めるキーワードの候補とする．

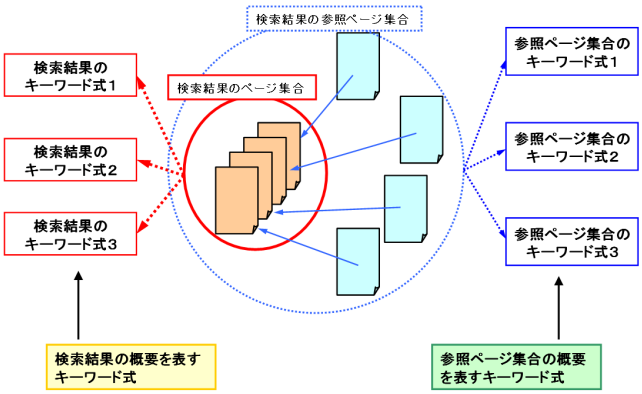


図３　キーワード式の位置づけ

4.2.2 検索結果の参照ページ集合からのアスペクトの抽出

アスペクトは，是津ら [6] によって提案されたもので，ある一つの Web ページが他の Web ページによってどのように参照されているのかを表す Web コンテンツのことを指す．本論文の提案手法では，幅広い情報を得るために，検索結果のページを参照しているページに含まれるキーワードをアスペクトとして扱うこととする．すなわち，一つの Web ページのアスペクトを抽出するのではなく，初めに選択したキーワードでの検索結果のページ集合に対してのアスペクトを抽出する．

（1） コンテキストの抽出

初めに選択したキーワードでの検索結果集合を参照しているページとして，検索結果のページそれぞれのリンク元のページを取り出してきて，その内容全体をコンテキストとして抽出してくる．参照ページ集合から得られる全てのコンテキスト集合をあわせたものを，検索結果のページ集合のコンテキスト集合とする．

（2） コンテキストからのアスペクトの抽出

コンテキストに含まれるキーワードは，初めに選択したキーワードでの検索結果集合に含まれないものも数多く存在する．コンテキストに含まれる名詞又は固有名詞を取り出してきてアスペクトとし，それを検索結果外の情報として扱うことで，初めに選択したキーワードに関する，検索結果からだけでは得られない情報として扱う．

コンテキストに含まれるアスペクト a について，検索結果におけるキーワードの重要度 w(k) と同じように df(a) を求め，それをもとに重要度 W(a) を計算し，その値の大きいものから順に，参照ページ集合の近似的な概要を表すキーワード式に含めるアスペクトの候補とする．

4.3 キーワード式の生成

抽出してきたキーワードとアスペクトについて，それぞれキーワード式を生成して表示することで，それらのページ集合の近似的な概要を表すことができると考えられる．検索結果と参照ページ集合の関係とその中でのキーワード式の位置づけを，次の図４に示す．

4.3.1 評価関数

検索結果から取り出してきたキーワード群の中から，W(k) の値の大きなキーワードを取り出し，それらを組み合わせて，初めに選択したキーワードの検索結果の近似的な概要を表すキーワード式を生成する．このとき，それぞれのキーワード式の評価関数 f の値を求め，その値の大きいものを優先してキーワード式に加えるようにする．初めに選択したキーワードの検索結果を Q とする．キーワード式の第一のキーワードを k_1 とし，k_1 から得られる検索結果を r(k_1) とする．n 個のキーワードによる検索結果集合を R(n) = r(k_1) U r(k_2) U ...U r(k_n) とする．

ここで，キーワード式の第一のキーワードになる可能性のあるキーワード k_1 の評価関数 f(1) は，(1) 式より次のようになる．

$$f(1) = (1-z)\frac{|Q \cap r(k\_1)|}{|Q|} + z\frac{|Q \cap r(k\_1)|}{|r(k\_1)|} \quad (4)$$

また，それ以降のキーワードをキーワード式に追加していく場合を考える．複数のキーワードについての評価関数となるため，キーワード式にｎ－１個のキーワードが含まれていて，新たにｎ番目のキーワードをキーワード式の候補とするときの R(n) は，次のようになる．

$$R(n) = R(n-1) \cup r(k\_n) \quad (5)$$

このとき，ｎ個目のキーワードを加えた時点での，ｎ個の要素を持つキーワード式の評価関数 f(n) は，(1) 式を応用して，次のようになる．

$$f(n) = (1-z)\frac{|Q \cap R(n)|}{|Q|} + z\frac{|Q \cap R(n)|}{|R(n)|} \quad (6)$$

4.3.2 検索結果のキーワード式

検索結果から抽出してきたキーワードは，それぞれ重要度 W(k) を計算する．重要度が大きなキーワードのうち上位３００件をキーワード式の候補とする．キーワード式の候補の i 番目のキーワードを，key[i] と表す．あるキーワード式を求めるとき，新たなキーワードをキーワード式に加えたときの評価関数ｆを計算し，その値の最も大きくなるときのキーワード key[i] を，もとのキーワード式に加える．ここで，もしどのキーワードを加えてもキーワード式の評価関数が元のキーワード式よりも大きくならなかった場合，新たなキーワードを付け加えないほうが，より検索結果の概要をよく表していると考えられる．

4.3.3 アスペクトのキーワード式

今回アスペクトとするものは，検索結果のページ集合のそれぞれを参照している Web ページの中のテキストに含まれるキーワードである．この参照ページ集合の概要をつかむことで，初めに選択したキーワードがどのようなものとしてとらえられているのかを，簡単に知ることができると考えられる．

検索結果の近似的な概要を表すキーワード式の生成と同じ手法により，検索結果を参照しているページ集合についてもキーワード式を生成する．

4.4 グラフの作成

4.4.1 提示するグラフ

グラフの表示の際に，初めに選択したキーワードの検索結果から得られたキーワード集合やアスペクト集合は，ただ箇条書きのように羅列するだけではどのようなことを表しているのかわかりづらい．そのため，次の２つの情報を一つのグラフとしてまとめて表示する．

- 検索結果集合の概要を表すのキーワード式
- 検索結果集合がどのようにとらえられているのかを表していると考えられている参照ページ集合のキーワード式

検索結果内の情報として検索結果集合の近似的な概要を表すキーワード式を，検索結果外からの情報として参照ページ集合の近似的な概要を表すキーワード式を提示することによって，検索の際に初めに選択したキーワードについての内外両方のさまざまな情報を得ることができ，それによって新たな知識の発見と，次の検索のための質問の修正につなげることを目的とする．

この提案手法によってユーザに提示するグラフは，図５のようなものである．

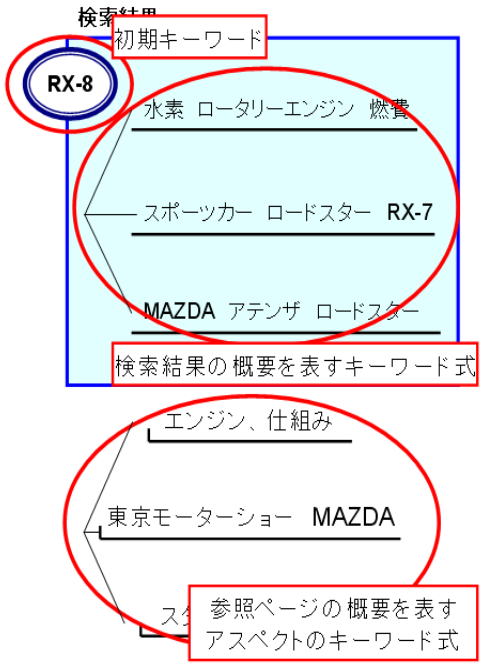


図４ 提示するグラフ

4.4.2 グラフ内の情報の強調

提示するキーワード式は，それぞれどの程度ページ集合の概要を近似的に表しているのかの度合いが異なり，度合いの違いは評価関数ｆの値によって表される．ここで，それぞれのキーワードにおける評価値をみなくてもどのキーワード式が検索結果の概要を表す上で重要であるかを判断しやすいように，評価値の高いものから順に提示する．ここで，提示したキーワード式の中で，ページ集合の概要を最もよく表していると考えられるキーワード式は，グラフの最も上に表示することで，わかりやすくする．

4.4.3 グラフの表示する範囲

次回の検索のためのキーワードを決定するために，初めに選択したキーワードに関係のある情報を得ようとする際，ユーザ

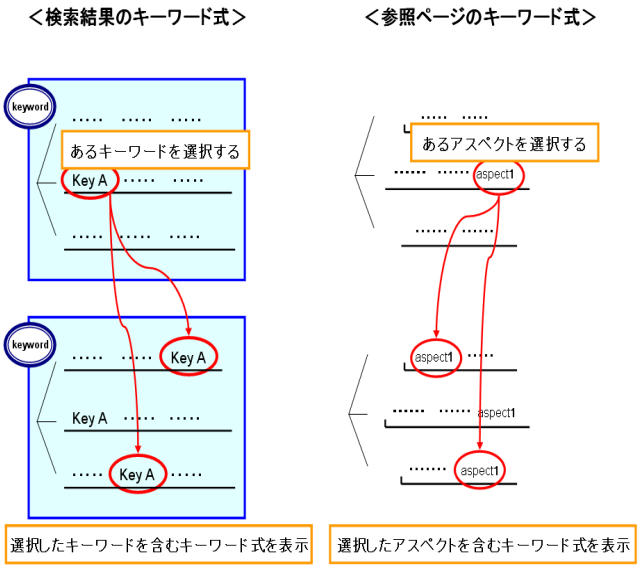


図５　グラフの表示範囲の変更

の興味対象がどういうものなのかはユーザ自身でなければわからないため，できる限りさまざまな情報をみせるほうが有効である．しかし，膨大な量の情報を一度に見せようとしても，本当に得たい対象がどこにあるのかがわかりにくく，またブラウザの表示可能範囲に限界があるなどの制限があるため，せっかくの情報の有用性がなくなってしまう．そこで，グラフとして一度にみせる情報の量を限定し，マウスやキーボードなどによる操作をできる限り少なくするようにする．

4.4.4　グラフの表示範囲の変更

グラフの表示範囲を限定した場合，その中にユーザの欲しい情報が必ずしも入っているとは限らない．取り出した情報はできる限り見せ，ユーザに選んでもらうために，Focus + Context View [2] [3] の概念を用いることにする．表示されている部分以外の情報が閲覧したい場合は，表示されているキーワードからユーザが新たに選択したものを中心として，グラフ表示する部分を変える．一度に提示する情報を少なく，かつ初めに選択したキーワードから得られた情報であることを把握しやすい形にするために，選択キーワードは常に表示する．また散策的にキーワード式を閲覧する場合のために，表示範囲を少しずつずらしていくためのボタンも用意する．最初に生成したグラフとして表示するものとしては，評価値の高い検索結果の概要を表すキーワード式を評価の高い順にある個数だけ提示する．参照ページのキーワード式も同様に，評価値の高いものから順にある個数だけ提示する．そして，他にどのような検索結果の概要の表され方があるかを知りたい場合は，表示する範囲を変えるようにし，一度に表示するキーワード式を制限する．表示するキーワード式を制御するボタンを用いて，ボタンを押すことで評価値の高いものや低いものを自由にみることができるようにする．

グラフからキーワードを選択してそのキーワードを中心としてグラフを再展開する例を，図６に示す．

提示してある検索結果集合と参照ページ集合のキーワード式それぞれについて，独立して変更することにする．キーワード式に含まれるあるキーワードに興味を持った場合，そのキーワードを選択することで，そのキーワードを含んでいるキーワード式を，評価値の高い順にある個数提示するように，グラフを変更するようにする．

5.　プロトタイプ

5.1　実 装 方 法

プロトタイプの具体的な流れは，次のようである．

（１）　初めに選択したキーワードを Google に入力して，その検索結果のうち上位１００件を抽出してくる．

- その検索結果の中のキーワードを抽出してきて，それぞれについて重要度 Wk(t) を求める．
- Wk(t) の値の大きいものから上位３００個のキーワードを，検索結果集合の概要を表すキーワード式の候補とする．
- Wk(t) の値の高いものから順にキーワード式に含まれるキーワードの候補とし，キーワード式に加えていく．
- 新たにできたキーワード式ごとに評価値 f を求め，さらにキーワードを付け加えていくことで，検索結果集合のキーワード式を生成する．

（２）　検索結果のページのそれぞれを参照しているページ集合を，Google のリンク検索により１０件ずつ取得してくる．

- 参照ページ集合に含まれるキーワードをアスペクトとして抽出してきて，それぞれについての重要度 Wa(t) を求める．
- Wa(t) の値の大きいものから上位３００個のアスペクトを，参照ページ集合のキーワード式に含まれるアスペクトの候補とする．
- Wa(t) の値の高いものから順にキーワード式に含まれるアスペクトの候補とし，キーワード式に加えていく．
- 新たにできたキーワード式ごとに評価値 f を求め，さらにアスペクトを付け加えていくことで，参照ページ集合のキーワード式を生成す る．

（３）　生成した検索結果集合のキーワード式と参照ページ集合のキーワード式をまとめてグラフ化する．

（４）　生成したグラフについて，その中のあるキーワード又はアスペクトを選択することで，キーワード式をそれぞれその選択した言葉を含んだキーワード式に表示を変更する．

5.2　実 装 画 面

プロトタイプの実装画面をつぎの図７と図８に示す．まずはデータ生成フォームにキーワードを入力すると，そのキーワードでの検索結果に含まれるキーワードと参照ページ集合に含まれるアスペクトを抽出，そしてキーワード式を生成する．データを取得した後，別フォームにグラフを表示する．

調べたいキーワードをキーワード入力部分に入れて動作ボタンを押すことで，抽出してきた検索結果の URL，参照ページ

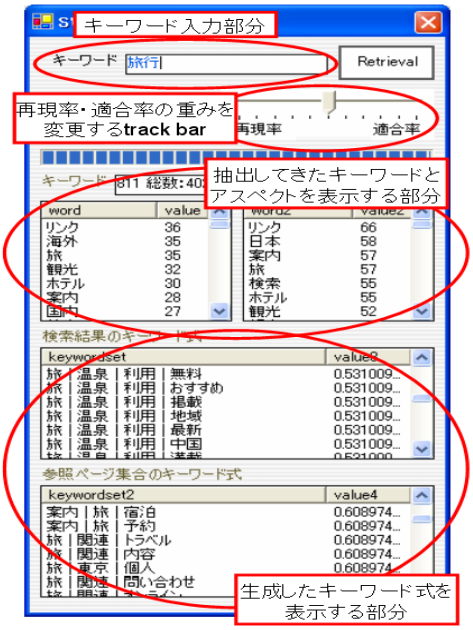


図 6　データ生成フォーム

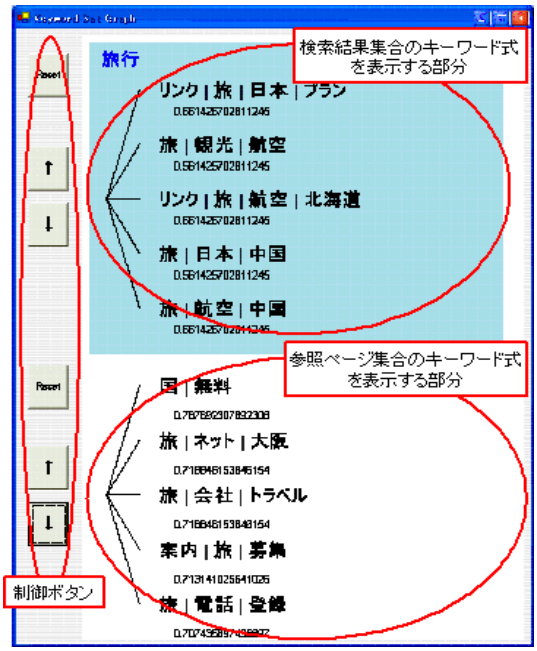


図 7　グラフ表示フォーム

の URL を取得し，それから得られる検索結果集合に含まれるキーワード，参照ページ集合に含まれるアスペクト，生成したキーワード式の一覧などをデータ部分に順に表示する．グラフ化については，全てのデータを抽出した段階で表示するようにする．また，再現率，適合率の重みｘは，ユーザの目的に応じて調整することのできるように，制御バーを用いて決定できるようにする．

5.3　提案手法の適用例

今回提案手法をどのようにユーザが用いて知識発見，質問修正に結びつくのかといったシナリオは，以下のようなものが考えられる．

- 閲覧ページの中で，意味を知らない言葉が出てきた場合

このような場合，その知らない言葉を選択キーワードとして入力し，グラフが出力される．言葉の意味は，その使われ方などによって変わる場合も多い．そのような場合は，初めに選択したキーワードがどのようなものとしてとらえられているのかというアスペクトや，そのキーワードのまわりのキーワードを見ることによって，どのような意味で使われているのかを知る手がかりを得ることができる．

- 言葉の意味は知っているがそれ以外の意味や使われ方を知りたい場合

このような場合，そのキーワードでの検索結果がどのようなものに参照されているのかというアスペクトを見ることで，キーワードに関する情報がどのようにとらえられているのかを知ることができる．検索結果の概要を表すキーワード式と，その言葉を含む参照ページのキーワード式を，その言葉の検索結果内外の情報として見ることで，その言葉のさまざまなとらえられ方を知ることができる．

- 適当に選択したキーワードとつながりのあるキーワードを発見したい場合

このような場合，生成されたグラフの中のキーワードを選択することで，そのキーワードまわりの他のキーワード式をみることができる．興味の対象となるキーワードを順に選択しグラフの表示範囲を変更していくことで，そのキーワードがどのような言葉と関係があるのかを知ることができる．

- 検索結果の概要を簡単につかみたい場合

あるキーワードを入力することで，グラフにその検索結果の概要を表すキーワード式が表示される．最もよく検索結果を表していると考えられるキーワード式を中心として表示するため，その検索結果のおおまかな概要をつかみたい場合は，中央に表示されているキーワード式を見ることで，より確かな概要をつかむことができる．しかし，自分の求めている情報がその最も適した概要によって表されていないと思われた場合は，そのグラフを随時新しいキーワードを選択して表示範囲を変更することで，得たい情報の概要を表しているキーワードを含むキーワード式を探すことができる．

5.4　評　　価

実装したプロトタイプを用いてわかった利点は，次のようなものである．

- 既存の検索エンジンなどではすぐにはわかりにくかった検索結果ページの近似的な概要を一目で確認できることができるため，入力キーワードがどのようなものとして Web で扱われているのかを理解しやすくなったと思われる．
- 検索結果内の情報だけではなく，検索結果外の情報もあわせて表示することで，普段の検索よりも幅広い知識を提供できるようになったと思われる．実際，参照ページ集合の中から，検索結果集合には含まれないキーワードも数多く抽出できている．

また，このプロトタイプの問題点としては，次のようなものが挙げられる．

- 実行時間

扱う情報として，検索結果の上位１００件，キーワードの評価

値の高い順で上位３００個，参照ページは各検索結果のページごとに２件，と限定してキーワード式を求めているが，実行時間が非常に長いものとなっており，実用は困難なものとなってしまっている．この原因としては，Google からページのテキストを抽出してくる部分で，他の演算よりも非常に時間がかかってしまっているからである．

- 表示するキーワード式について

検索結果集合からキーワードを抽出してくる際，非常に一般的な語もあわせてとってしまっているため，キーワード式を見てもその概要がわかりにくい場合も出てくるときがある．

- キーワード式の評価関数について

キーワード式の評価関数ｆについて，再現率と適合率を用いてどの程度検索結果集合を表せているのかを求めようとしているが，検索結果全てを扱うことは物理的に不可能であるため，近似的に検索結果の上位１００件としている．しかし，Web から得る検索結果の数からみれば非常に少ないと考えられる１００件だけを選んで抽出してきているため，実際にはある程度再現率と適合率の値があるはずであるのに，取り出してきた１００件ではその共通集合が扱われずに評価関数の値が実際よりも低くなってしまい，正しい評価関数が求められていない場合が出てくる．

5.5 今後の課題

- アルゴリズムの効率化による実行時間の短縮

このシステムを実際に使う上で，実行時間は非常に重要なものである．Google を呼び出すときにかかる時間は Google の仕様によるものであるので，変えることはできない．そのため，他の演算部分のアルゴリズムを効率よく行う形に変えることで，実行時間の短縮をはかることを検討している．

- データの抽出方法の検討

ページ集合からキーワードを抽出する際の重みとして，あるキーワードを含んでいるページの数としている．しかし，今回のプロトタイプでは，抜き出してきたキーワードを重みの大きい順に３００個しかキーワード式の候補としていない．幅広い情報を扱うためには，さらに多くのキーワードをキーワード式の候補とすべきである．そのため，さらに効率のよい重みを用いることで，表示できるキーワード式もより幅広い概要を表しているものにすることができると考えられる．

- キーワード式の評価関数

本提案手法では，キーワードの評価関数として再現率と適合率を用いているが，Web 空間は非常に膨大な量のページを含んでおり，効率よく再現率と適合率を用いるのは非常に困難である．そのため，ページ集合全体をページの類似度などによりクラスタリングして，各クラスタごとに均等に検索結果集合として抜き出してくることで，より概要の精度を上げることなどを今後の課題として検討している．

- 表示する情報について

表示する情報について，キーワード式の表示法も一意に決まるものではない．特に，検索結果内の情報と検索結果外の情報のつながりというものも見せることができると，さらに概要の理解の促進と，質問修正のためのキーワードの発見につなげることができると考えられる．表示するデータも，さまざまな形でわかりやすくユーザに提示できるように検討することが今後の課題である．

6. ま と め

本論文では，Web 情報検索のユーザ支援を目的とした，検索結果とその参照ページ集合の近似的な概要を表すキーワード式の生成とそのグラフ化についての提案と実装を示した．概要を一瞥できるという点に加え，検索結果内だけではなく検索結果外の情報をあわせて提示することにおける知識発見と検索の質問修正の支援が行えたと思われる．

しかし実装にあたり，本論文で提案した方法であると，Web の中の全てのページを取得してきて行うのは，実行時間などの関係によりほぼ不可能に近い．そのため近似的に検索結果の上位１００ページをとってくるようにしているが，それでは検索件数が莫大な場合は非常に評価の精度が低くなってしまう．Web 空間の広さを効率的に近似することによって，ページ集合の概要を表す度合いをより効果的にするように，また実行時間も短くするようにしながら改良を検討することを今後の課題として考えている．

グラフ表示についても，キーワード式のさまざまな提示の仕方があると考えられるのに加え，検索結果と参照ページ集合の近似的な概要を表すキーワード式同士の関連性もグラフに表すことができれば，さらなる Web 情報検索支援につながると考えられるので，今後の課題として考えている．

7. 謝　　辞

本研究は，一部平成 15 年度科研費特定領域研究「Web の意味構造に基づく新しい Web 検索サービス方式に関する研究」（課題番号：15017249　代表：田中克己）による．ここに記し謝意を表します．また，本研究は，一部，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．ここに記し謝意を表します．また，本研究は，一部独立行政法人通信総合研究所と京都大学の共同研究「インターネット・コンテンツの意味構造発見に基づく新しいコンテンツ検索・配信方式の研究」による．ここに記し謝意を表します．

文　　献

[1] http://www.google.co.jp/

[2] Sarkar, Manojit and Marc H. Brown, “graphical fisheye views,” Comm. of the ACM, Vol.37, No.12, pp.73-83, 1994.

[3] Furnas , G. W., “Generalized Fisheye Views,” Proc. ACM SIGCHI ’86 Conference on Human Factors in Computing Systems, pp.16-32, 1986.

[4] http://www.webbrain.com/

[5] Shinichi Ueshima, Kazuhiro Ohtsuki, Jun-ya Morishita, Qing Qian, Hiroaki Oiso, Katsumi Tanaka: Incremental Data Organization for Ancient Document Databases. DASFAA 1995: 457-466

[6] 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

[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