キーワード平面を用いたインタラクティブ検索に関する研究

林　大策†　佐藤　哲司††

† 筑波大学図書館情報専門学群　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学図書館情報メディア研究科　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †s0813181@u.tsukuba.ac.jp, ††satoh@slis.tsukuba.ac.jp

あらまし　Web 検索エンジンが提示する一次元リスト構造の検索結果では，類似したページが多くある場合や多様な内容が混在する場合に欲しい情報を見つけることが難しい．本研究では，検索結果をアイコン化して二次元平面に配置し，多くの検索結果を俯瞰的に表示することで，ユーザがインタラクティブに視点を変えられる検索結果の閲覧手法を提案する．ページの分布や配置を見渡しつつ，個々のページにアクセスできる検索結果の提示は，多様な内容が混在する検索結果の中から欲しい情報にアクセスするための手段を提供する．検索エンジンのフロントエンドとなるシステムを実装し，提案法の実現性と二次元平面を用いた可視化の有効性を確認した．

キーワード　情報検索, インタラクティブ, 可視化，二次元平面

Interactive Information Retrieve System with Two-dimensional Keyword Plane

Daisaku HAYASHI† and Tetsuji SATOH††

† School of Library and Information Science,University of Tsukuba
1–2,Kasuga,Tsukuba,Ibaraki,305–0855 Japan
†† Graduate School of Library Information and Media Studies,University of Tsukuba
1–2,Kasuga,Tsukuba,Ibaraki,305–0855 Japan
E-mail: †s0813181@u.tsukuba.ac.jp, ††satoh@slis.tsukuba.ac.jp

Abstract　When the case that there are many similar page or various contents were intermingled, it is difficult to find demand information by the search results of the one-dimensional list structure that Web search engine shows. This paper proposes method of watching search results that make search results an icon and arrange it on a two-dimensional plane and a user can change a viewpoint interactively, by displaying many search results panoramically. While looking around distribution and the placement of the page, the presentation of the search results that can access the individual page offer means to access demand information from the search results that various contents were intermingled. We implemented the system which became the front end of the search engine and confirmed the efficacy that used feasibility and panel of the suggestion method for.

Key words　Infomation Retrieval, Interactive, Visualization, Two - Dimensional Plane

1. あらまし

近年は情報爆発と言われるように，インターネット情報空間は爆発的にページ数が増え続けている．一語や二語程度のわずかな検索語で検索を行った場合，様々な内容のページが大量に検索結果として返され，ユーザが求めるページを簡単には見つけ出すことができない．一般的な検索方法として，ユーザは検索結果をブラウズしながら検索語の追加と削除を繰り返しているが，このような試行錯誤にも限界がある．追加の検索語を想起することや，余計な検索語を除外し適切な検索語だけで検索質問を的確に構成するのは，ユーザにとっては負担が大きいからである．

このような検索語の想起・選択に関わる問題を解決するための手法として，例えば Google suggest(注1)は，検索語を入力すると，検索語との組み合わせ頻度が高い検索語の候補を検索結果件数とともに提示する．ユーザは，提示された検索語の候補

(注1)：http://www.google.co.jp/

の中から適切な検索語を選んで検索質問とすることができる．しかし，Google suggest では選択した検索質問で検索してみるまで得られる結果が分からず，試行錯誤を繰り返さなければならない．

現在，広く使われているインターネット検索エンジンでは，検索質問との類似度順に一次元のリスト構造で検索結果が出力されており，様々な内容が混在していることも検索を難しくしている．例えば，検索質問「アップル」で検索を行うと，コンピュータ会社や自動車販売会社，ホテルチェーンなど様々なページが混在して検索される．そのため，どのような検索語を追加すれば，どのように検索結果が変わるのかをあらかじめ知ることは，繰り返し実行する検索操作を円滑にするために非常に重要であると考えられる．

本論文では，このような検索語の選択に関わる問題を解決することを目的として，検索結果を俯瞰しながら簡便に検索語を変更できる，新しい検索インターフェースを提案する．検索結果を二次元平面上に配置し，個々のページ間の関係や全体の分布を可視化する．平面を構成する軸は検索結果を特徴づけるキーワードとし，提示されるキーワード候補からユーザが自由に選択し与えることができる．検索結果のページは，選択されたキーワードとの関連度に基づいて座標を計算し二次元平面上に配置する．キーワードを取り換えるユーザのインタラクションで平面上のアイコン化されたページの配置を変え，全体の特徴をつかむことができる．以上により，キーワードと関連の強いページの発見やページ間の関係を把握することができる，インタラクティブな検索を実現する．提案手法のプロトタイプとなるシステムを実装し，課題と応用について考察を行う．

以下，本論文では 2 章で関連研究について述べ，本研究の位置付けを明らかにする．3 章では提案手法の概要について述べる．4 章では提案システムでの実装について述べ，5 章では実験結果について示す．6 章で評価と考察を行い，7 章でまとめる．

2. 関 連 研 究

本論文では，検索結果のページをそれぞれアイコン化して二次元平面に表現することを検討するが，検索結果を可視化する研究は多数存在している．

林ら [1] は，文書情報を可視化することで絞り込みを行う検索支援法を提案している．検索結果が含む単語を列，各文書を行としたマトリクス形式で検索結果を表現する．ユーザがある文書を正解文書として指定することで，それに類似する単語を持つ文書を並べ替えて提示する例示型の検索支援を実現している．また，ある単語を追加することで得られる検索件数を求め，絞り込み率を計算する．検索結果中での単語の出現頻度と，絞り込み率の二つの軸で二次元平面を構成し，単語を配置する．これにより，絞り込みのための単語を視覚的な情報で与えている．

関ら [2] は，任意のフィールドを縦軸と横軸に設定してマトリクスを作成し，検索結果を各セルに配置する手法を提案している．縦軸，横軸を任意の数に分割し，縦軸と横軸でそれぞれ分割数だけクラスタリングした結果から二次元のマトリクスを構成する．行と列の項目には，クラスタリングされたページの特徴語が複数提示され，セルに含まれるページの特徴を俯瞰的に見ることができる．

本論文の第 2 の特徴は，軸キーワードを取り換えるというインタラクティブな操作であるが，インタラクティブな操作で結果を再ランキングする手法も多く提案されている．

中谷ら [3] は，検索結果から複数の特徴語を抽出し，それらを項目とするレーダーチャートを作成する手法を提案している．レーダーチャートに表示された語の重要度をユーザが調整することで，結果を再ランキングする．吉田ら [4] は，検索結果を特徴語と共起語のグラフとして表現している．作成されたグラフにある語と語の距離を調整することで，検索結果の再ランキングを行う手法である．

本論文は，検索結果の可視化と操作のインタラクティブ性を向上することを目的としているが，林らや関らが提案する検索結果を集約して特徴を簡潔に提示する手法とは方向性が異なる．むしろ，検索結果はページ一つ一つを単位として扱うことでページ間の関係を浮き彫りにすることを目指す．検索結果件数などの数字として検索結果の分布を表現するのではなく，二次元平面に検索結果をダイレクトに表示することで，検索結果の可視化とインタラクティブな操作をシームレスに接続することができると考える．

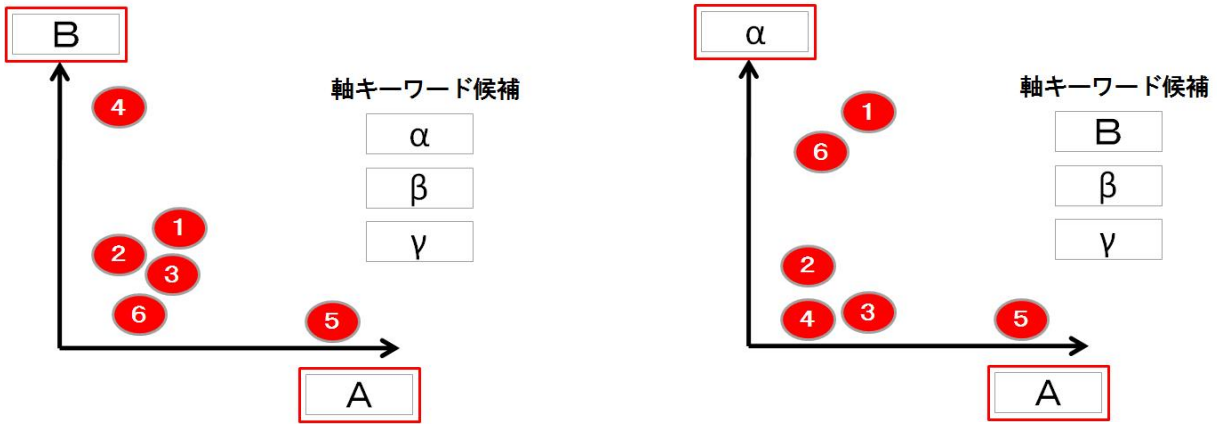
また，中谷ら，吉田らの手法は，ユーザのインタラクションによってダイナミックに検索結果を変化させるという点では類似した研究となっているが，ページ間の関係をより直接的に表現しようとする点が本論文の特徴として位置づけることができる．ページをアイコン化して表示する手法は，ページ間の関係性を直感的に把握でき，インタラクティブな操作との親和性が高いと考える．

3. キーワード平面の作成方法

3.1 キーワード平面を利用した検索の概要

本論文で提案するインタラクティブな検索支援手法では，検索結果の可視化とユーザによる操作のシームレスな連続性を重視する．これまでの一般的な Web 検索では，検索結果を一次元のリスト構造で表示しているが，検索結果を上位から順番に見ていくのでは混在している様々な情報に埋もれてしまう情報もある．また，下位の情報を閲覧するにはページを繰るなどの操作が必要となる．そこで，検索結果を多次元空間に表現することで，空間的に検索結果を表現し，必要とする情報を分析，選択しやすくすることを検討する．また，多次元空間に表示された検索結果を見ることで，次元空間を構成する軸の再設定や，新たな検索質問を想起するインタラクティブな操作が行えるようにする．このように，多次元空間で可視化された検索結果を見て，そこから次の操作を円滑に誘発できる支援システムとすることを本論文の目的とする．

可視化する検索結果の次元数を決めることは，重要な課題である．三次元以上の任意の次元数での可視化は，検索結果を多様な視点から高度に分析するのには極めて有効な手法であるといえるが，一方で，どのような軸を与えるか，あるいは，ど



(a) 縦軸キーワード「A」　　(b) 縦軸キーワード「α」

図 1　検索操作の概念

の軸を選んで相関を見るべきかなどの自由度が高まることで，ユーザの負担が増大する可能性も否定できない．

検索結果の可視化とユーザ操作のシームレスな連続性を重視する本論文の目的から，最も単純な多次元空間である二次元平面を用いて検索結果を可視化し，二次元の各軸に検索結果を配置するための軸キーワードを設定する方法を提案する．検索結果のページを特徴付けるキーワードを軸キーワードの候補としてシステムが提示し，ユーザは提示されたから候補から興味のあるキーワードを選んで軸キーワードとする．また，検索結果を更に絞り込みたい場合は，提示されたキーワード候補を検索質問に (AND 条件として) 追加することができる．これらのインタラクションを実現することで，ユーザの情報探索行動を支援できると考える．

二次元に限定したことで検索結果の分析は限定されるが，軸キーワードを簡便に入れ替え可能とすることで，任意の二軸間で検索結果を配置することができ，その配置のパターンから軸と検索結果の相関や軸相互の依存関係などを様々に分析することはできると考えられる．

提案手法におけるインタラクティブな検索操作の概念を図 1 で示す．図 1(a) は横軸にキーワード「A」を，縦軸にキーワード「B」を設定した場合のキーワード平面であり，縦軸のキーワード「B」を，軸キーワード候補として提示された「α」と置き換えた場合を示したものが図 1(b) である．図の (a)，(b) いずれにおいても，平面上に配置したページ ① ～ ⑥ は，ある検索質問 Q によって得られた 6 件の検索結果ページのアイコンであり，軸キーワードを入れ替えることで同一のページ集合が再配置されることを表している．いずれの図においても，右側にあるページほど横軸キーワードとの関連度を高く，上側にあるページほど縦軸キーワードとの関連度が高いとする．ユーザが Q を与えて図 1(a) の結果を得た時，ページ④は，二次元平面の上端にあることからキーワード B と関連が強いと判断できる．また，ページ⑤が A というキーワードと関連が強いことも分かる．A,B を含む軸キーワード候補 α，β，γ は Q によって得られた検索結果から抽出したものである．ここでユーザが，縦軸キーワードを A から α に置き換えることで，図 1(b) の結果が得られる．新たに設定された縦軸キーワードとページ間の関連度が再計算され，各ページの配置が変化する．このとき計算するのは縦軸との関連度のみで，横軸については更新しない．そのため全てのページの x 座標は変化しない．新たに得られた結果を見ると，キーワード α と関連の強いページは①と⑥であることが図 1（b）の結果から分かる．このように，ユーザが軸キーワードを取り換えることで，ダイナミックに検索結果が再配置され，ページの分布形状から軸キーワードとの関係の強さや，結果ページ間の関係性などの特徴を知ることができ，目的の情報を見つけやすくなると期待される．

3.2　軸キーワード候補の算出法

キーワード平面を構成するために，軸キーワードの設定が重要である．前節で述べたように，ページ発見のためにキーワード平面を活用する場合，ページを特定しやすい特徴的なキーワードを与えることが不可欠である．

3.2.1　キーワードの抽出

ユーザが選択するキーワードの候補を抽出するにあたり，検索結果のページからタイトルとスニペットを対象として形態素解析を行った．検索エンジンは Yahoo!検索 Web API(注2) を利用し，形態素解析ソフトには MeCab(注3) を使用して，「一般名詞」「固有名詞」「形容詞」を抽出した．「代名詞」「記号」「数詞」「助数詞」「感動詞」「動詞」などは用いていない．これらの品詞・記号はページを特徴づけるキーワードとしては不適切な可能性が高いからである．形容詞を抽出するのは，「軽い」「赤い」「安い」携帯電話など，ページのに記述された対象の特徴を表現するときに形容詞は有効であると考えたからである．また，複合名詞を扱うために，一般名詞が連続するとき，一定の文字数を超えない範囲で結合した語を作成した．これは，ページの特定性を高める特徴的な語を抽出できるようにするためである．また，結合前の語も結合語とは別に抽出候補とする．結合語は

(注2)：http://developer.yahoo.co.jp/webapi/search/
(注3)：http://mecab.sourceforge.net/

字数が多く，検索結果中でもあまり出現しないため，あるページを特徴付けるのに役立つ．文字数制限を超えた場合，最後に結合された語を先頭として，再び結合条件を満たせば結合する．文字数に制限がない場合，結合の過程で語を生成し過ぎてしまう問題がある．この問題については，後に例を交えて説明する．本研究では結合語の文字数制限は 5 文字とし，結合した後の文字が 5 文字以上であった場合はそれ以上結合しないとした．

以上の方法で，あるページ(注4)から抽出したキーワードの例を表 1 に示す．

タイトル	筑波大学 春日キャンパス
スニペット	図書館情報専門学群の案内等． ... 図書館情報専門学群. 情報学群 知識情報・図書館学類. 情報学群 情報メディア創成学類 ... 図書館情報メディア研究科. 図書館情報学図書館. 知的コミュニティ基盤研究センター. 学術情報メディア ...
キーワード	筑波大学, 春日, キャンパス, 春日キャンパス, 図書館, 情報, 図書館情報, 専門, 情報専門, 学, 情報専門学, 群, 学群, 案内, 等, 案内等, 情報学, 情報学群, 知識, 知識情報, 図書館学, 類, 図書館学類, メディア, 情報メディア, 創成, メディア創成, 創成学, 創成学類, 研究, メディア研究, 科, 研究科, 情報学図書館, 知的, コミュニティ, 知的コミュニティ, 基盤, コミュニティ基盤, 基盤研究, センター, 基盤研究センター, 学術, 学術情報, 学術情報メディア

表 1　キーワード抽出例

「春日」「キャンパス」はそれぞれ形態素解析で別の単語となるが，本手法では「春日キャンパス」と結合される．また「図書館情報専門学群」という文字列は，形態素解析では「図書館」「情報」「専門」「学」「群」という単語に分けられる．本手法では「図書館」と「情報」は名詞が連続しているため結合され「図書館情報」となる．この時点で 5 文字以上となるため，これ以上「図書館情報」に結合はされない．最後に結合された「情報」は次の複合名詞の先頭の文字列となる候補として残す．次が「専門」という名詞であるため「情報専門」と結合し，次の「学」も結合する．この時点で 5 文字以上となり，「情報専門学」以上は結合しない．また「学」は頭として残り，次の「群」と結合して「学群」となる．次は名詞ではないため結合はせず「学群」で確定されて，頭には何も残らない．文字制限がなければ「図書館情報専門学群」が結合可能であるが，そこに至るまでに「図書館」「図書館情報」「図書館情報専門」「図書館情報専門学」「図書館情報専門学群」と五つの単語を生成する．このようにして，一つの文字列から複数のキーワード候補となる文字列を生成する．

3.2.2　キーワードの得点付け

上述した方法で生成されたキーワードに得点を付けることで，軸キーワードとして有効な候補を抽出する．得点付けの手法として $tf-idf$ 法を変形して用いた．一般的な $tf-idf$ 法は，ある文書における単語の得点を計算するが，軸キーワード候補を選ぶ用途では個別の文書毎に $tf-idf$ 値を求める必要はない．そこで，検索結果全体を一つの文書として考え，検索結果全体におけるそれぞれの単語の $tf-idf$ 値を算出する．

$$
\begin{aligned}
tf(t) &= \text{検索結果の全ページ中に単語 } t \text{ が登場する回数} \\
idf(t) &= \log \frac{\text{検索結果数}}{\text{単語 } t \text{ を持つページ数}} \\
tf-idf(t) &= tf(t) \times idf(t)
\end{aligned}
$$

以上の式を用いて，検索結果中のキーワードの点数を求める．全文書に登場する語は IDF 値が 0 となり，$tf-idf$ 値の得点が 0 となる．これは，全文書に出現する語を軸キーワード候補として提示しても有用でないためである．この手法により全てのキーワードの得点を計算し，上位 30 語を軸キーワード候補としてユーザに提示することとした．また，軸キーワードの候補のうち，上位二つを初期状態とし，1 位を横軸，2 位を縦軸のキーワードとしてキーワード平面を構成する．

3.3　関連度計算による配置方法

関連度の計算は各軸で独立して計算する．あるページの横軸キーワードとの関連度を計算し，もう一方で縦軸キーワードとの関連度を計算する．最後に関連度を正規化し，座標として用いる．本手法では軸キーワードを含むページを，その軸に対して関連度が高いものと判断する．軸キーワードとページの関連度の計算は，単純に軸キーワードを含むページに得点を与える手法を用いる．ただし，一つのキーワードでは，そのキーワードを含むか含まないかの二値でしか評価できない．ページを平面上に分散させるためには更に手順が必要である．そこで，軸キーワードを含むページが持つキーワードは軸キーワードと関連があると仮定して，関連語として扱う．軸キーワードを含むページに点数を与え，次に関連語を含むページにも点数を与えていく．点数付与の流れを図 2 に示す．

具体的には，軸キーワードか関連語を持つページには，該当するキーワードの $tf-idf$ 値を累計的に与える．つまり，関連語の中でも特徴的な語を持つものが，高い関連度を得るように

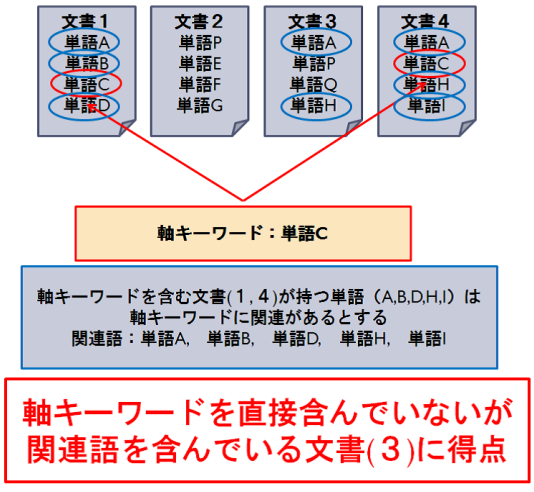


図 2　軸キーワードと関連語を用いた関連度計算

(注4)：http://www.slis.tsukuba.ac.jp/

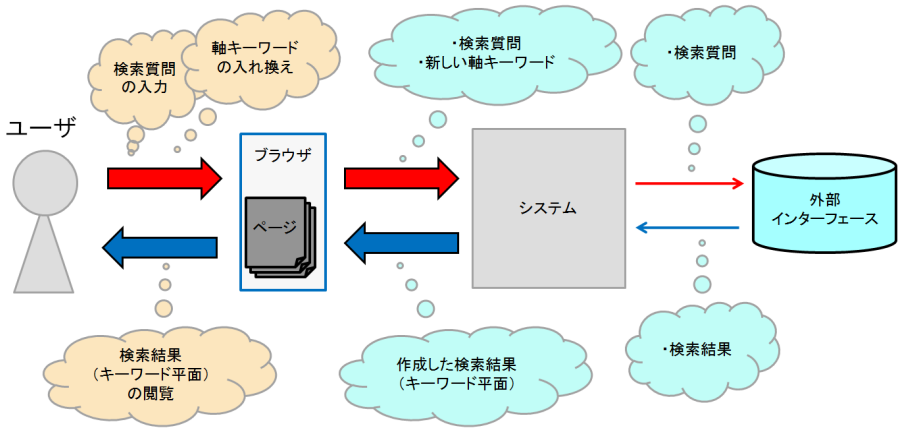


図 3　システム概要図

設計する．こうすることで多くのページに差異のある点数付けを行い，分布が広がるように設計した．これは単にキーワードの有無による二値で点数を与えた場合よりも結果が分散する効果を期待したからである．ページごとに横軸キーワードとの関連度，縦軸キーワードとの関連度を計算して正規化し，それを座標とする．正規化は，各軸における最大の関連度で各ページの関連度を除算した．これにより全てのページの関連度を 0〜1 の範囲にする．また，軸キーワードは関連語よりも重要なキーワードであることは明らかなので，軸キーワードの $tf-idf$ 値には補正得点を加算する．この補正得点は，もっとも $tf-idf$ 値の高いキーワードの 0.5 倍とした．

以上の方法で，キーワード抽出から関連度計算までを行う．

4. 提案システムの実装

実際に作成したシステムの概要を図 2 に示す．ユーザからの入力や結果の出力を行うプログラムがブラウザ上でインターフェースとして提供され，処理を行うシステムがサーバ上にある．まず，ユーザから入力された検索質問を，Javascript を介してシステムに渡す．システムは検索質問を受け取ると，Yahoo!検索 Web API に検索質問を入力する．システムは Yahoo!検索 Web API から検索結果を受け取り，3.2 節で述べた手法でキーワードの抽出と点数付けを行う．そして 3.3 節の手法で関連度の計算を行い，各軸の座標を計算する．その後にシステムは検索結果ページ全ての「タイトル」「スニペット」「アドレス」「x 座標」「y 座標」をブラウザ側に送る．それを受け取ったブラウザ上のプログラムが実際にページの配置と描画を行う．以上によって，ユーザは，入力した検索質問に対する検索結果をキーワード平面上に取得する．同時に，軸キーワード候補も提示される．

サーバ上のシステムは Java で書かれたプログラムである．Java と Javascript との通信は dwr(注5) というフレームワークを用いて，Ajax による非同期通信で行う．よりインタラクティブ性の高い検索を行うためには，Ajax による高速な反応が望ましいと考えた．Yahoo!検索 Web API の利用には Slothlib(注6) を用いて実装した．

次に軸キーワードを取り換えたときの動作を説明する．軸キーワードが取り換えられたとき，ブラウザ側の Javascript プログラムがシステムに新しい軸キーワードを渡し，配置を再計算する．Yahoo!検索 Web API に新しく検索質問を投げて検索結果を取得することはしない．計算した結果の座標をブラウザ側に返し，再び描画する．ユーザは軸キーワードを取り換えて，手早く結果を再配置することができる．実際にシステムを使って検索した結果を図 4 に示す．使用した検索質問は「図書館情報専門学群」で，初期状態設定された横軸キーワードは「研究」，縦軸キーワードは「類」である．この結果が図 4(a) で，横軸キーワードを「春日」に変更した結果が図 4(b) である．検索結果件数は 100 件である．

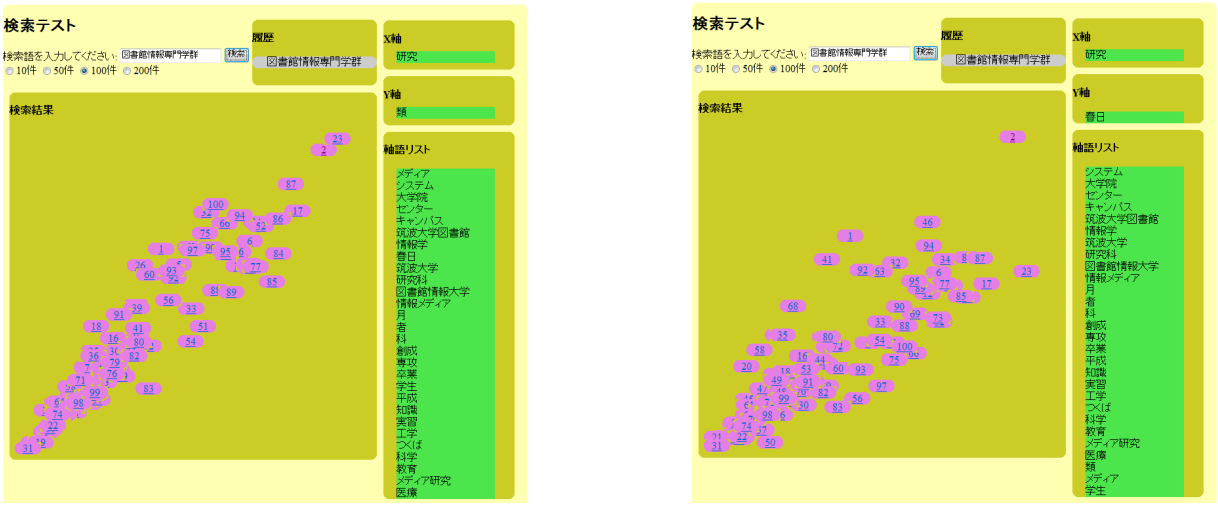
左上がユーザが検索質問の入力と検索結果件数を指定する部分である．その下の検索結果と書かれた領域は，上で入力された検索質問の結果ページをアイコン化して表示する領域である．右側ほど横軸キーワードとの関連度が高く，上側ほど縦軸キーワードとの関連度が高くなる．アイコン上に表示されている番号は，従来の Yahoo 検索エンジンが返す一次元リスト構造の検索結果の順位である．右側の横軸，縦軸と書かれた領域は現在選択されている軸キーワードを表示している．その下の軸語リストという領域は軸キーワードの候補を提示している領域である．

本システムでは，50 件の検索結果を取得するのに 5〜10 秒，100 件の検索結果を取得するのに 10 秒〜20 秒，200 件の検索結果を取得するのに 25 秒〜40 秒程度であった．軸キーワードの取り換えによる再配置は，200 件で 1 秒〜2 秒程度であった．Yahoo!検索 Web API に問い合わせて結果を取得する時間を除くと，キーワードの抽出と得点付けに時間がかかっており，関連度の計算そのものにはあまり時間がかかっていないと分かる．処理時間短縮のためには，不要なキーワードを減らすことなどが考えられる．

動作環境

本システムが動作する環境について表 2 に記しておく．

(注5)：http://directwebremoting.org/dwr/index.html
(注6)：http://www.dl.kuis.kyoto-u.ac.jp/slothlib/



(a) 横軸：「研究」，縦軸：「類」　　(b) 横軸：「研究」，縦軸：「春日」

図 4　検索質問「図書館情報専門学群」の検索結果

OS	Windows Vista Enterprise 64bit
CPU	AMD Athlon(tm) 64 X2 Dual Core Processor 5200+ 2.70GHz
メモリ	3966MB
サーバ	JBoss Server 5.1

表 2　動 作 環 境

5.　提案手法の実験結果

本章では提案手法を実装したプロトタイプシステムによる実験結果ついて述べる．5.1 節では実際にどのような軸キーワードが選ばれるのかについての結果を示す．5.2 節ではシステムが提示する検索結果を示し，考察する．

5.1　軸キーワードの提示

本節では具体的な検索例に基づいて提示されるキーワード候補を調べる．軸キーワードの選び方は 3.2 節で説明した通りである．$tf-idf$ の値が高い 30 語を軸キーワードの候補として提示する．検索質問「図書館情報専門学群」における軸キーワードの候補の結果を表 3 に示す．検索結果の件数によって提示されるキーワードが変化するため，本節での検索結果件数は全て 100 件とする．

表 3 の項目は左から，キーワード，検索結果全体でのキーワードの出現回数 (TF)，検索結果 100 件中での出現件数 (IDF)，計算された $tf-idf$ 値となっている．$tf-idf$ 値の降順でソートされており，上部の方が $tf-idf$ 値の高いものとなる．これらのキーワードは検索結果のタイトル・スニペットを使って作成されたものであり，この候補リストを見て検索結果の概要を知ることができる．ユーザはこの中から興味のあるキーワードを選択し軸とすることで，そのキーワードおよび，キーワードの関連語が使われているページの傾向を確認できる．

本手法で用いた複合語生成により「図書館学類」や「情報学群」など，検索質問に対して特徴的な複合語が取得されている．一方で「図書館学」「情報学」は結合途中の複合語である．本来の文脈に沿ったものではないので，軸キーワードとして活用しづらい．本手法では結合語の文字数が 5 文字以上であればそれ以上は結合しないルールとしたが，これが無制限であると，上位三十件中に更に結合途中の複合語が増加することを確認した．結合途中の複合語を使わず，最後まで結合したもののみを重視すると一般語が上手く取れないケースが増える．複合語と一般語をバランス良く選ぶことが今後の課題となる．

キーワード	出現回数 (TF)	出現文書数 (IDF)	$tf-idf$
類	116	31	135.86
メディア	66	25	91.50
研究	55	23	80.83
センター	32	8	80.82
情報学	51	31	59.73
システム	31	15	58.81
創成	24	9	57.79
つくば	22	11	48.56
科学	25	15	47.43
筑波大学	111	68	42.81
情報メディア	28	22	42.40
筑波大学図書館	33	28	42.01
キャンパス	14	5	41.94
知識	21	14	41.29
春日	17	9	40.94
者	23	17	40.76
月	22	16	40.32
学生	23	18	39.44
資料	14	6	39.39
専攻	14	6	39.39
平成	19	13	38.76
試験	12	4	38.63
知識情報	18	13	36.72
医学	19	15	36.05
図書館学	18	14	35.39
入試	14	8	35.36
マップ	10	3	35.07
情報学群	17	13	34.68
図書館学類	17	13	34.68
編入	15	10	34.54

表 3　検索質問「図書館情報専門学群」に対する軸キーワード候補

5.2 提案システムによる実験結果

実際にシステムで検索を行った結果を示す．検索質問「図書館情報専門学群」，横軸キーワード「図書館学類」，縦軸キーワード「学生」，検索結果数 100 件のキーワード平面を図 5 に示す．

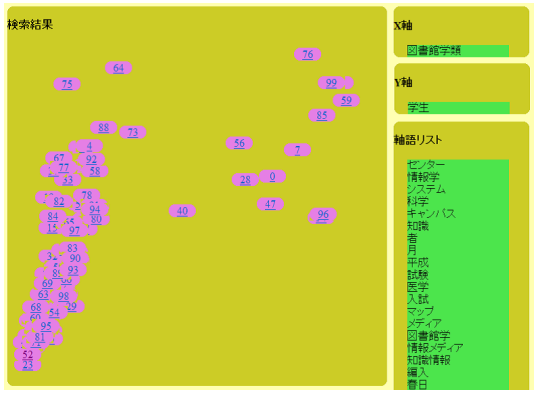


図 5　検索質問「図書館情報専門学群」，横軸：「図書館学類」，縦軸：「学生」，検索結果数 100 件

この結果は，大きく分けて右半分と左半分にページが分散した．まず横軸から結果を見る．関連度計算アルゴリズムより，「図書館学類」と関連度が高いページほど右に配置されている．表 3 より，検索質問「図書館情報専門学群」の検索結果 100 件で「図書館学類」は DF（出現文書数）が 13 となっている．図 5 の右半分にあるページは全部で 13 あり，全てが「図書館学類」をタイトル・スニペット中に含んでいることを確認した．つまり，軸キーワードを含むページは漏れなく関連度が高いと計算できていることになる．

次に縦軸について結果を見る．下端から上端までページが分散しており，ページが幅広い関連度点数を持っていることが分かる．表 3 より,「学生」の DF は 18 となっている．DF が高いキーワードほど，軸キーワードになったとき関連語の数が大きくなる．関連語の数が多くなると多くのページに点数付けができるようになり，得点が分散する．よって，ページも分散して配置される．また「図書館学類」よりも「学生」の方が一般的なキーワードであるため，作られた関連語も様々なページが含みやすいものだったと考えられる．その結果，横軸よりも縦軸の方がページが分散した可能性が高い．

次に検索質問「2009 年 十大ニュース」，横軸キーワード「AP 通信」，縦軸キーワード「米経済」，検索結果件数 200 件のキーワード平面を図 6 に示す．　本手法での関連度計算方法では横軸と縦軸が同じキーワードであった場合 Y=X の直線上に全てのページが並ぶ．また，同じキーワードでなくても，横軸キーワードを含むページが全て縦軸キーワードも含んでいるという状態であれば，Y=X 直線上に全てのページが並ぶ．両方のキーワードから作成される関連語が同じであるからである．したがって「2009 年 十大ニュース」という検索質問で得られる結果の中で「AP 通信」を含んでいるページはほとんどが「米経済」を含んでいるということが図 6 から明らかとなる．また

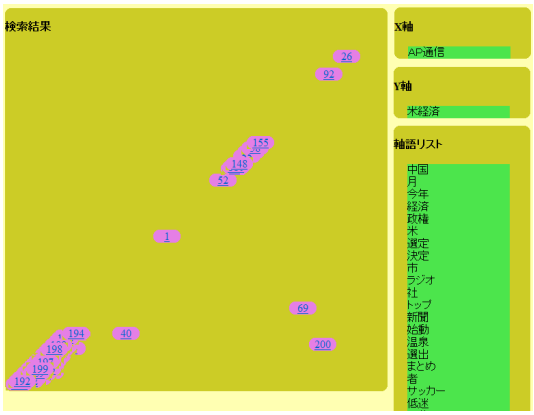


図 6　検索質問「2009 年 十大ニュース」，横軸：「AP 通信」，縦軸キーワード「米経済」，検索結果数 200 件

右下に離れて配置された 200 番のページは，AP 通信の芸能の十大ニュースであった．これは AP 通信に関するページの中で，ただ一つだけ米経済について述べていないページである．他の直線上にないページは「AP 通信」を含んでいるものの,「米経済」ではなく「経済」とされているものであった．このことから「米経済」が出現するページは全て「AP 通信」が出現するページであることも分かる．このように，二つの軸を組み合わせることでキーワードの関係の分析や，関連度の高さ・低さを利用して少数派のページを発見することが可能であることが分かった．また 200 番のページは従来の検索結果では 200 位として提示されるものである．一般的な検索ではほぼ閲覧されないページと言ってもよい．こういったページもキーワード平面では発見することができる．

検索結果件数を変更して実験を行ったところ，検索結果件数が小さくなると，関連語数が少なくなり，上手く得点付けができないという問題があった．提案手法は，ある程度は検索結果が多い方が有効であると考えれる．ただし，検索結果件数が増えると処理に時間がかかるため，より優れたアルゴリズムが必要となる．

6. 評価と考察

本手法で用いた関連度計算方法により，キーワード平面にページが配置されることが確認できた．軸キーワードを取り換えることにより，再配置ができることも分かった．従来の検索では，検索結果を閲覧しながら目的のページを探し，次の検索のための検索語を探していく．本手法では，検索質問を修正するよりも手軽に，軸キーワードを取り換えることで配置を変更することができる．あくまで同一の検索結果内での再配置であるが，軸キーワードごとに特徴を変えるので，欲しい情報が探しやすくなると考えられる．また，軸キーワードを含む文書がどれくらいあるのか，軸キーワードが検索結果中でどのような位置付けなのかなど俯瞰的な閲覧が可能となることも分かった．

本手法で見込めるキーワード平面の活用は主に二つあると考えられる．一つ目は，ページをキーワードとの関連度順に並べることで，関連度の高いページや低いページを見つけ出す方法

である． 図 5 では「図書館学類」のようなキーワードを軸にすると，それをタイトル・スニペットに含むページを取り出すことができる．これを利用して，特徴的なキーワードを軸にすることで特定のページを探し出すことが可能と分かった．二つ目は，ページの散らばりによって検索結果を分析する方法である．軸キーワードを含むページの多寡や傾向，次の検索に使うキーワードの有効性などを調べることが主な目的になる．「学生」のようなキーワードを軸としても，それに関するページを特定して探すことは難しい．しかし，「学生」というキーワードは「図書館情報専門学群」という検索質問に対して普遍性のあるキーワードだと分析することもできる．また図 6 のように二つのキーワードを使うことで，分布状態からページの傾向をつかむことも可能である．

関連度計算方法は，現在はタイトル・スニペットに軸キーワードを含むものが高得点となっているが，ページ個々の関連は考えられていない．$tf - idf$ 値の高いキーワードを含むページほど関連度は高くなり，ページ間での座標の差は現れるが，実際にページの関連度順に並んでいるとは判断できない．関連度が高いページの一群としての閲覧は可能である．提案法における関連度の計算法は，二次元平面に配置することを目的としており，算出された関連度が，そのまま軸キーワードとページ，あるいはページ間の類似度として有効であるかどうかは評価されていない．これを行うためには，軸キーワードとその関連語について詳細に検討する必要がある．たとえば，軸キーワードの類義語には高得点を与えるなど，共起する語の中からも関係を探っていく方法が考えられる．その上でページの配置と実際の関連度について調査することが有効と思われる．

キーワード平面を使ってスムーズに検索を行うためには，軸キーワード候補の提示が重要である．本手法で提示される 30 語中には，検索結果中でよく使われている語や，特徴的な語も現れた．候補リストを見れば，検索結果中でどんなページがあるのかをある程度把握できる．しかし，候補リスト内にあるものは手早く使われるが，ユーザの興味に合うものが提示できなければシステムの有益な効果は望めないという問題もある．どのような軸キーワードがどんな検索に適しているのか，提示方法はどのようなものが良いかを調査していく必要がある．

インターフェースとしての機能の充実化も重要な課題である．現在は軸キーワードの変更や検索結果の表示はできた．しかし，有効な配置やデザイン，必要な機能については検討の余地がある．軸キーワードの選択方法や，軸キーワードの追加方法，新しい検索質問を入力する方法など，ユーザのインタラクションに対する機能が必要となる．

本論文ではキーワード平面を利用した検索インターフェースの基礎的な部分しか研究ができなかった．実際にこのシステムを使った検索と，従来の検索とどういった違いが出るのかを利用者実験などで明らかにすることが必要である．そこから有利な点と不利な点を分析し，応用可能な領域を探していく．

7. おわりに

本論文は，インターネット検索における検索語の変更や追加と削除に関する問題を解決するために，検索結果を俯瞰しながら簡便に検索語を変更できる新しい検索インターフェースを提案した．検索結果のページをアイコン化しして二次元平面上に配置し，ユーザのインタラクションによって再配置を行い，検索結果を閲覧しやすくする．二次元平面は二つの軸となるキーワードで構成され，それぞれの軸キーワードとの関連度で平面上でのページの座標を決定する．ユーザが軸キーワードを取り換えることによりページの配置を変化させ，インタラクティブな検索が可能である．

提案するインターフェースのプロトタイプシステムを実装し，評価を行った．検索結果がキーワード平面上にどのように配置されるかを実験している．その結果，軸となるキーワードを含むページが，その他のページと分離されるように配置されることを確認した．また，軸キーワードを取り換えることにより，検索結果から該当するページをシームレスに取り出すことができた．これにより，検索結果の構成を俯瞰的に分析することが可能であると分かり，様々な応用への展開ができる基盤技術となりうるとの知見を得た．

今後の課題は，より使いやすいインターフェースとするための改良・拡張，ならびに，利用者実験による評価と応用の開拓である．

文 献

[1] 林一成, 岩佐英彦, 竹村治雄, 横矢直和. 文書情報の可視化による検索絞り込み支援. 電子情報通信学会技術研究報告, NLC99-80, 99(708), pp.15-20 (2000).

[2] 関隆宏, 和多太樹, 山田泰寛, 廣川佐千男, 検索支援と分析のための多面的検索システム. 電子情報通信学会第 18 回データ工学ワークショップ (DEWS2007), E1-2 (2007).

[3] 中谷文彦, 河合由起子, 熊本忠彦. 柔軟な Web コンテンツ検索のためのレーダーチャート検索システムの提案. 電子情報通信学会第 19 回データ工学ワークショップ (DEWS2008), B5-4 (2008).

[4] 吉田大我, 小山聡, 中村聡史, 田中克己, Web 検索結果におけるキーワード出現相関の可視化と対話的な質問変換. 電子情報通信学会第 18 回データ工学ワークショップ (DEWS2007), C7-2 (2007).