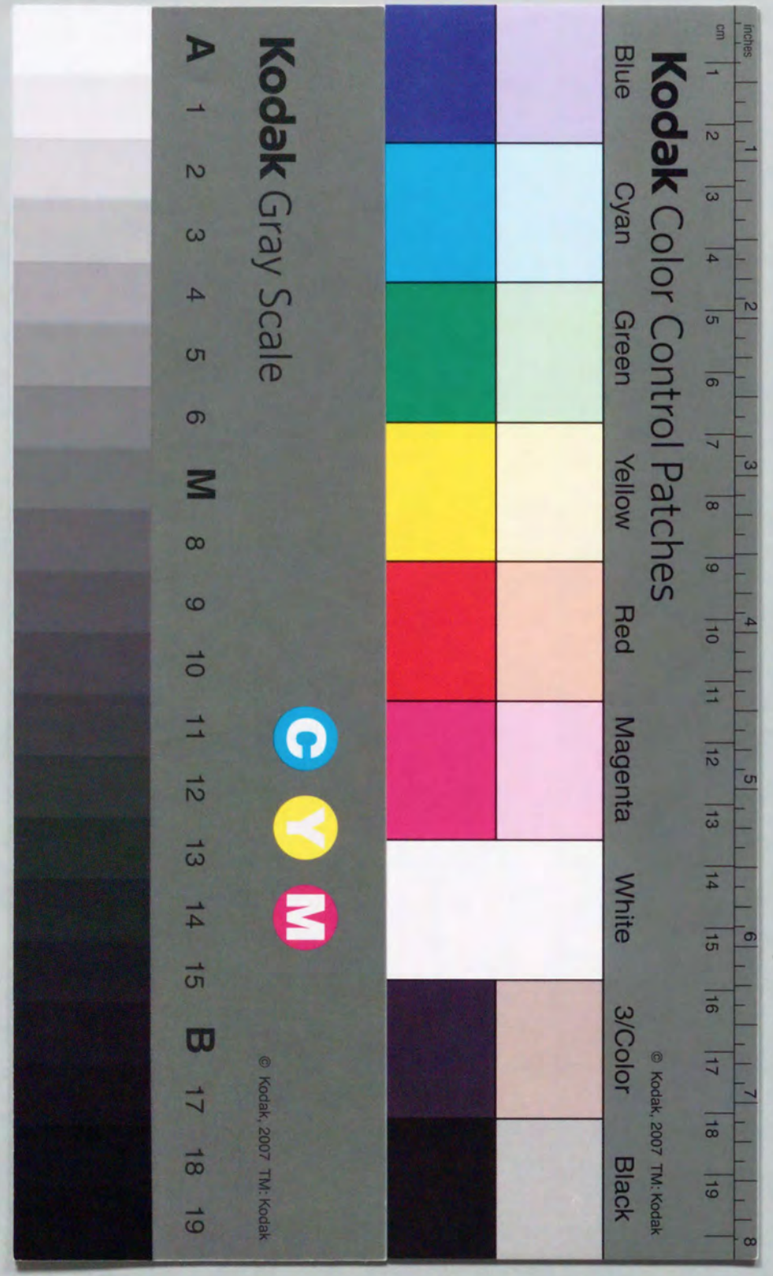


# 検索における使用単語の想起と配置の支援

## ―興味表現支援システム―

砂山　渡

大阪大学大学院基礎工学研究科

博士学位論文

２０００年４月

# 検索における使用単語の想起と配置の支援

## ―興味表現支援システム―

砂山　渡

大阪大学大学院基礎工学研究科

博士学位論文

２０００年４月

# 要約

近年における情報化産業の発展には目を見張るものがあり，インターネットや電子メールなどによる情報の伝達が盛んに行なわれている．ある話題に関する情報を得たいと思えば，それを入手するための手段としてインターネットを用いることが，ごくあたりまえの時代となってきた．しかし，あふれんばかりの情報の中から，真に欲しい情報を探し出すことは情報量の拡大にともなって徐々に困難になりつつある．

本論文では，インターネット上で情報を探すために検索システムを用いるユーザが，単語の組合せを入力として与える作業すなわちユーザの興味表現を支援し，ユーザの素早い情報獲得を実現するシステムを提案する．本システムの狙いは，ユーザが，検索される側の Web ページの情報を知り，検索要求の元となっているユーザ自身の興味をより具体的な単語として表せるように導くことで，ユーザの検索に用いた単語すなわちユーザの知識と，存在する情報との間のギャップを埋めることである．

そこで，ユーザが検索に用いた単語の関連語を，実在する Web ページから抽出してユーザに提供する．これらの単語は検索された Web ページ集合を端的に表すものであり，ユーザは存在する情報の特性を確認しながら，自身の興味をより具体的かつ的確に表す単語を選び用いることで効率の良い検索を行なえる．

また，関連語を二次元平面上に配置したインターフェイスをユーザに提供する．このインターフェイスは，提供する関連語を，検索に用いられたユーザの興味を表す単語との関わりに応じて分類し，存在する情報とユーザの興味との関係や相違を明示する．これら，ユーザが検索に用いるべき単語を想起すること，および検索のための適切な検索条件となる単語の配置を支援するシステムを本論文で実現する．

# 目 次

# 1　序論

「表現する」とは人間の思考，気持ち，感情，感覚，意図，興味などの本人のみが理解できる形のないものを，言葉や文章という多くの人が意味を理解できる形あるものへと変化させることである．

人がこの表現を行なう際には，頭の中に存在する不明確な感情を具体化せねばならない．しかしこの作業において，自らの考えが自分で摑めない，表現として表すための適切な言葉を即座に思い起こせない，あるいは初めから頭の中や実世界に該当する言葉が存在しないなどの理由によって，自らの思考そのものをちょうど表す表現を生成することは極めて困難である．それでも日常生活においては何げない会話が行なわれ，電話，手紙，電子メールなどを用いるさまざまな状況において表現が行なわれている．すなわち人は実世界で生活する間に，自らの思考に最も近い表現を自然と選ぶ能力を身に付けることで，情報の伝達を行なっている．

しかし時に，情報を伝達するお互いの間に誤解や問題を招くことがある．自らの生成した表現が自らの意図とは異なる方向に解釈され，情報が正しく伝達されないという状況もしばしばである．これは，人がみな異なっているために人それぞれの思考や考え方が異なり，たとえ同じ伝達手段としての言葉を用いていたとしても，その受け止め方はさまざまとなることに起因する．すなわち情報伝達というプロセスは,「表現」した後の「解釈」まで達して初めて完了するものであるため，この解釈が正しく行なわれる表現を選ぶことが必然となる．そのためには，客観的に広く用いられる意味において言葉を使う必要があり，この表現において適切な言葉を表現者に示唆することは表現の支援となり得る．それゆえ，表現を支援する側と表現する側とが言葉の示唆と表現の生成という２つの作業を互いに繰り返すことによって思考と言語の間のギャップを埋め，表現者の意図を十分に表しかつ誤解のない解釈を生じる表現が作成可能になると考える．

しかし，表現の支援としてはこれだけでは十分でない．適切と考えられる言葉の示唆を受け表現に必要な言葉を知っていたとしても，それらを適切に使う術を知ら

なくては表現として成り立たない．正しい言葉の語順や組合せによって言葉を使えてこその表現なのである．

そこで，具体的な表現の支援として次の２つを挙げることができる．

- 単語の想起の支援：表現に必要，適切な単語を表現者に提示すること
- 単語の配置の支援：表現に適切な単語の配置を表現者に示唆すること

本論文においては，インターネット上の情報検索を行なうユーザのための，検索語入力による興味表現を支援するシステムについて述べる．

本システムは検索を行なうユーザに対して，上に挙げた表現の支援方法に基づく検索支援を行ない，ユーザの素早い情報獲得を支援する．すなわち，ユーザが入力した検索語に関連する検索に用いられる単語をユーザに提示することで単語の想起を支援し，提示された単語を検索条件として追加する方法を，検索インターフェイス上で明示して単語の配置を支援することによってユーザの検索を支援する．そのため本システムのユーザは，検索語の入力とシステムからの検索語の提供という２つの操作を繰り返すことで，自らの曖昧な興味を適切な検索語の組合せとして具体的に表現することができ，後戻りなく目的の情報にのみ向かう素早い情報獲得を実現できる．

以下本論文においては，第２章で従来の検索および検索の支援に関する研究について述べた上で，本研究の位置付けを明確にする．続く第３章で大衆の興味に一致するユーザの興味を発見し，ユーザの興味に関連する単語をユーザに提供する興味表現支援システムについて述べ，第４章で検索式をユーザの興味の生成される過程に基づいて分割して，ユーザの興味を構成するすべての要素に関連する単語をユーザに提供する興味表現支援システムについて述べる．第５章では，存在するWebページ情報を，提供する単語によって視覚化して検索を支援する二次元平面インターフェイスについて述べ，第６章で検索実験およびアンケート調査に基づく本システムの評価を行なう．最後に第７章で結論を述べて本論文を締めくくる．



# 2　検索および検索の支援に関する研究の背景

近年，インターネットが急速に普及するにつれ，さまざまな情報を容易に獲得できる時代になってきた．電子メールをはじめ，Webページ，ネットニュース，電子図書館などさまざまな情報が電子化された状態で存在し，それら電子的なデータベースの中から情報を獲得する機会が増えている．しかしその一方で情報が氾濫を始め，欲しい情報の所在を突き止められない，情報の有無を確認できないという事態も起こっている．この情報量の膨大さと人間における時間および心身の制約によって，全ての情報に目を通すことは不可能であり，この膨大な情報の中からいかにして情報を収集するかについての研究 [Fayyad 96] が，現在盛んに行なわれている．

本章では，情報検索を含む情報収集の方法について述べ，情報検索の現状が抱える問題点とそれらへの解決策とを対比した上で，本研究の位置付けを明確にする．

## 2.1　情報収集の方法

WWWのように大きなデータベースの中から情報収集を行なう方法は次の３つに大別される．それらは，必要とする情報の分野やキーワードなどをあらかじめ設定しておくことで入ってくる情報を制御する「情報フィルタリング」，それから，あるWebサイトを起点として自由に情報を探し回る「情報散策」，また，検索システムを用いて情報を探す「情報検索」の３つである．まず以下で，このそれぞれに関する概略と特徴について見ていく．

### 2.1.1　情報フィルタリング

情報フィルタリングとは，予め，自分の興味のある分野やキーワードなどの情報をプロファイルとして設定しておくことで，ネットワークから流れてくる膨大な情報の中からプロファイルに見合った情報だけを自動的に選別して受けとる技術である [森田 96, Mori 99]．この技術は，情報を自ら探しにいくこととは逆の立場に立っ

ており，メールやニュースのように毎日自ずと入ってくる情報の選別に用いられる．情報フィルタリングにおいては，ユーザ独自の情報であるプロファイルの設定が重要であり，プロファイルの初期設定および更新方法が研究の対象となっている．プロファイルの設定および更新には，ユーザが介在する場合と自動学習による場合とがあり，ユーザが提供された情報を評価する関連フィードバック技術は情報検索においても用いられている．

### 2.1.2 情報散策

WWW のネットワーク上を Web ページに含まれているハイパーリンクを逐次辿ることで，目的（明確な目的とは限らない）の情報を目指すことが情報散策である．WWW 上の情報散策は巨大な WWW ネットワークの空間内を漂う様子を喩えて，ネットサーフィンとも呼ばれる．ネットサーフィンの方法として，階層（ディレクトリ）型データベース Yahoo[Yahoo]，AltaVista[AltaVista]，NTTdirectory[NTT Directroy] などの中にあるハイパーリンク（リンク）を逐次辿りながら目的の情報を探して漂う場合と，検索システムによる検索結果の Web ページを起点としてリンクを辿る場合などがある．また，検索システムのデータベースを構築するために用いられている Web ロボットも，情報散策によって情報を集めている．

### 2.1.3 情報検索

情報検索とは，WWW 上に存在する Web ページを収集して構築されたデータベースに対して，ユーザが検索システムを通して質問 (Query) を与え，その質問に合った Web ページ[1] をデータベースから抽出することである．WWW 上で最もよく用いられている検索システムがサーチエンジン [原田 97]（検索エンジン）であり，WWW ユーザの 9 割以上に用いられている [インターネット白書 98]．本論文においては，このサーチエンジンを用いるユーザを支援するシステムを構築して提案し

[1]実際には存在する Web ページへのポインタとなる URL



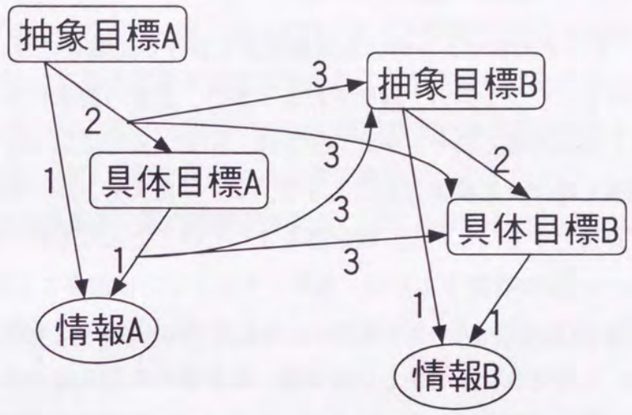


図 1: 情報検索による効果

ている．情報検索を行なうことの効果は，目的の情報を獲得するのみに留まらない．検索ユーザは抽象的もしくは具体的な目標（図 1の抽象目標 A や具体目標 A）を抱いて検索を行なう．検索（図 1内の矢印）によって目的の「情報を獲得」（矢印１）できる以外にも，検索開始時には漠然とした検索目標しか有していなかったユーザの興味が，検索過程において絞り込まれる「目標の具体化」（矢印２）や，ユーザが検索過程において新たな知識を発見し検索目標の翻意を行なう「新たな検索目標の生成」（矢印３）にも役立っている．本論文で述べるシステムでは，この検索の効果のいずれをも支援する．情報検索による効果をまとめると次のようになる．

**[情報検索による効果]**

1. 目的の情報を獲得する
2. 漠然とした検索目標が具体化される
3. 新たな検索目標が生成される

次節では，この情報検索における従来研究についてより詳しく述べる．

## 2.2 情報検索の現状:問題点と解決策

本節では，サーチエンジンを用いる情報検索における問題点について言及する．ユーザはサーチエンジンを用いて検索を行なう際に，自身の興味を表す検索語を入力する．すると現在の多くのサーチエンジンは，NGRAM[Cohen 95] などの手法を用いた全文検索と呼ばれる検索方式によって，ユーザに入力された検索語そのものを含むページを出力する．しかし，検索語を１つだけ入力してもその検索語を含む数多くの Web ページが存在するため，各サーチエンジンは独自の方法[2]に基づいて各 Web ページに点数を与え[3]，その点数による上位の Web ページを順番に一定数出力する．だが，入力された１つあるいは少数の検索語のみからユーザの具体的な興味を正確に推し量り，正に適切な Web ページにのみ高い評価を与えて出力することは不可能である．そこで，正確ではないにしろユーザの興味を近似して推量し，その近似に基づいて Web ページを出力する方法が必要となる．まず，ユーザの興味を推定する上で問題となる点を次に挙げる．

1. ユーザの入力する検索語が不足する
2. 語彙不一致の問題 [Furnas 87] が生じる
3. ユーザ自身の興味が不明確である

１番目の点はサーチエンジンへの入力としての検索語が不足することである．検索語の数はもともと少数であるが，それでも１つよりは２つ，２つよりは３つの検索語が入力された方が，より具体的なユーザの興味を推定できる．入力される検索語が少なくなる原因としては，ユーザが検索対象として存在する Web ページの内容を知らない，すなわちどのような検索語が目的のページに用いられていているのかをユーザが知り得ないこと．また，度忘れなどによってユーザが検索語をすぐに思

[2]検索語の Web ページ内での出現頻度と使われ方（タイトル，太字，ハイパーリンクへの使用）によることが多い

[3]順位付け操作，もしくはランキング操作と呼ばれる



いつけないことや，検索語として適当な単語をもとから知らないこと．さらに，少しサーチエンジンに慣れたユーザにとっては，ヒット件数[4]に応じて情報を過度に絞り込み過ぎないように徐々に検索語を追加することなどが挙げられる．これらのユーザに共通する望みは，実際の Web ページに使われている単語の中から，ユーザ自身の興味を具体的に表す単語を獲得することである．

２番目の点は語彙不一致の問題が生じることである．この問題には次の２種類が存在する．

- ユーザの与えた検索語が，Web ページ中でユーザの意図とは異なる別の意味で用いられている．
- ユーザの望む情報が，Web ページ中でユーザが与えた検索語以外の語で表現されている．

前者には多義語などが相当し，たとえば同じ「氷」という単語であったとしても，アイスコーヒーのグラスに入っている氷を連想する人もあれば，南極大陸の氷山をイメージする人もあるだろう．このように，同じ単語でも時と場合で大きく意味合いが異なるため，関係のない多くの Web ページが存在する場合には，必要な情報のみを検索することが叶わなくなる．

後者は一つのものに複数の名称，呼び方がある場合に相当する．たとえば同じ「かたつむり」の名称として他に，でんでんむし，まいまいつぶり，蝸牛（かぎゅう）などがある．このような場合，複数の呼び方のうちの一つを用いて検索を行なっても，他の単語で表現されているユーザの興味に関連する情報が検索されない．

すなわち，上記 1. の検索語不足の問題が解決され，ユーザの具体的な興味を表す単語が数多く与えられたとしても，それらをうまく組み合わせて欲しい情報をより多く，不要な情報をできるだけ除く検索条件によって検索が行なわれない限り，ユーザの満足がいく検索を実現することはできない．

[4]検索条件に当てはまる Web ページの総数

問題点の3番目として，ユーザの興味が具体化されないまま検索が行なわれている場合がある．2.1.3項でも触れたように，漠然とした興味の下で検索が始められ，検索を繰り返すうちに徐々にユーザの興味が具体化されていく場合があり，ユーザは検索結果として出力されるWebページを見ていくうちに，ユーザの望む情報に関する単語を思いつくことで検索語を追加していく．たとえば,「最近の流星群」について調べたいユーザは初め「流星群」という検索語のみで検索を行なうが，検索を行なう中で「しし座」という単語を発見しこれを検索語に加えて「しし座流星群」という，より具体的な興味を表す表現を用いて検索を行なうようになる．このユーザの興味の具体化を支援するためには，検索に必要な単語をユーザに提供すること以外に，存在するWebページの主な内容をユーザに知らせる必要がある．

これら本節でここまで述べてきた問題点の解決策として，現在以下のものが考えられている．

- 知的情報検索による解決
- キーワード提供による解決
- ユーザの興味を学習することによる解決
- 検索結果の可視化による解決

以下の各項でこれらについて一つずつ見ていきたい．

### 2.2.1 知的情報検索による解決

知的情報検索と呼ばれる検索を行なうサーチエンジンがある．これは，ユーザが入力した検索語をもとにして，ユーザの検索語に関連する検索語を自動的に補なって検索を行なう．本項では，関連語による検索語の補完および，キーワードベクトルによる検索と関連フィードバック技術について述べる．

まず曖昧検索 [増井 91] においては，一つの言葉が言い換え可能である場合に，その換言可能な単語を共に検索条件として用いることによって情報の漏れを少なくすることを目指している．たとえば，虎について調べたいユーザが「虎」という検索



語を入力した場合，システムは「とら」「タイガー」という同義の語を検索語に加え，そのいずれかにあてはまるWebページを検索して獲得する．この方法を用いると，異なる表現から生じる語彙不一致の問題を避けることができる．しかし，ユーザが動物の虎を調べたいにも関わらず，プロゴルファーの「タイガーウッズ」が検索にマッチするなど，かえって不要な情報までも検索されることがある．また，曖昧検索の中には同義語だけでなく類義語も自動的に追加しているものもあり，さらに不要なページが数多く検索される結果も引き起こしている．これはすなわち，ユーザの意図を無視して勝手に情報を補完しても，必要な情報が得られるという必然性がないために，かえって検索の効率を下げていることが多い．

検索語の関連語を補完する従来研究として，検索対象となる文書集合を特定し，その文書集合内で用いられている単語間の関連度を学習することによって，検索語と関連の強い語句を用いて検索を行なう方法 [Chen 96, 吉川 98] がある．しかし，検索対象となる文書の数が増せば，用いられる単語の数や単語の組合せの数が増加するため，実存の膨大なWebページ情報から関連度を学習する場合，情報が更新されるごとに単語間の関連度を学習し尽くすことは困難となる．

次に，ベクトル空間法によるキーワードベクトル [Salton 83] を用いた検索がある．この検索ではまず各データベース中のWebページを，そのページが含む単語あるいはキーワードを各次元とする空間上のベクトルとして表す．その上で，ユーザが入力した検索語をキーワードとして表した検索ベクトルと各Webページのベクトルとを比較し，検索ベクトルに近いベクトルをもつWebページを順に出力する．この検索方法によると，全ての検索語を含んでいなくても，一部の検索語を含むページは類似性が高いとして出力される．それゆえ，一部の検索語の表記が異なっていても出力に含まれる可能性がある反面，曖昧検索と同様に不要なページまでも出力されることが多くなる．すなわち，必要な情報を逃さないための検索になっているが，情報を十分に絞り込む操作に欠けている．

キーワードベクトルを用いた別の検索手法に，関連フィードバック [Rocchio 71] を用いた検索がある．自動的に関連フィードバックを行なうシステム [Robertson 94,

新田 99] においては，システム側で予備検索を行ない，その結果による上位の Web ページのベクトルを，ユーザの興味に関連があるページのベクトルとして検索ベクトルに加算し，新たな検索ベクトルを用いて再び検索を行なった上で，検索結果をユーザに提示する．この検索においても，システムの予備検索で得られた上位の Web ページに含まれるキーワードが新たに検索語として補われる効果があり，自動的に関連があるページが集められる反面，ユーザの意図とは異なる結果を出力する可能性が残される．そこで，細かなユーザの興味はユーザにしかわからないため，ユーザ自身に興味があるページをフィードバック [Salton 88] してもらった上で，検索条件を更新するという方法 [Eguchi 99, 帆足 99] がとられている．

ユーザは，出力された上位の Web ページの中から自身の興味のあるページに印をつけ，逆に興味とは関係のないページにも印をつける．この操作の後，印をつけられたユーザの興味に関連がある Web ページのベクトルを加算，逆に関連のない Web ページのベクトルを減算して検索ベクトルを更新し，より正確にユーザの興味を表す検索ベクトルを構成している．ここで加減算に用いられるベクトルのための係数パラメータの設定も研究対象となっている [Buckley 95, Salton 90]．このユーザの介在する関連フィードバックを用いると，より具体的なユーザの意図が反映された検索を行なえるとともに，関連があるとユーザが示した Web ページに含まれている新たな検索語が自動的に追加されるため，検索語を補う効果も得られている．

しかし，この手法にもいくつかの問題点がある．まず，各 Web ページの取捨を判断をするためのユーザの負担が軽くないことが挙げられる．有効な関連フィードバックの結果を得るためには，およそ 20 くらいの Web ページをそれぞれの内容に基づいて関連の有無を判断する必要がある．サーチエンジンの出力に含まれる各 Web ページの要約は，現在のところ各 Web ページの冒頭文に過ぎないために，各ページを正確に評価するためには，一つずつ Web ページへのリンクを辿って内容を実際に確認する必要がある．仮に冒頭文だけで判断する際には，一つの Web ページに複数の情報が含まれている場合などにおいて，各 Web ページの中から具体的にどのようなキーワードが抽出されるかが不明であり，ユーザの気づかないところで誤った検



索ベクトルが生成されることがある．

また，関連フィードバックを用いる検索においてはユーザの興味が具体化されず，興味を具体化するための労力が従来のサーチエンジンと変わらない．なぜなら，各 Web ページを 1 単位として情報をフィードバックしているために，ユーザが具体的な単語として自らの興味を確認することによる興味の具体化が行なえないからである．また，ユーザの興味という点に関して，自身の興味への関連の有無を判断するためには，ユーザの興味がある程度具体的である必要があり，それ以前の曖昧な興味しかもたない段階における検索では，関連フィードバックを用いることができない．

そこでこれらの知的情報検索による検索の新たな問題点を踏まえると，具体的なユーザの興味はユーザ自身にしかわからないという立場がとられる．すなわち，Web ページではなく，より具体的なキーワードを用いたユーザとのインタラクションによって，ユーザ自身が適切なキーワードを用いて検索を行なうことを支援するアプローチが現れる．

### 2.2.2 キーワード提供による解決

ユーザが一度目の検索を行なった後に，ユーザの具体的な興味を表す単語をユーザに提供することで再検索を支援するアプローチが存在する．ユーザに提供される単語には，ユーザが入力した検索語によって検索された Web ページに含まれる**絞り込みキーワード**と，ユーザが入力した検索語に関連する**関連キーワード**，そして，シソーラス（類義語辞書）から得られる検索語の類義語がある．以下これらの単語を提供する従来システムについて述べていく．

まず，多くの情報の中からユーザの興味に焦点を当て，情報を絞り込むためのキーワードとして，**絞り込みキーワード**をユーザに提供するシステム InfoNavigator[InfoNavigator], Mondou[Mondou], ditto[ditto] がある．これらのシステムは，ユーザが入力した検索語全体にマッチする Web ページから取り出される単

語をキーワードとしてユーザに提供している．すなわち，検索されたWebページ集合の中の一部に含まれる単語であるから，検索語と提供された単語とを同時に含むWebページの存在が保証されており，もとの検索結果を確実に絞り込むことができる．しかし，ユーザが初めに与えた検索語がユーザの興味を表すのに適切でない場合には，ユーザの望むページが検索されず，有効な絞り込みキーワードが得られないことがある．また検索にヒットするWebページが一件も存在しなかった場合には絞り込みキーワードが得られないという問題もある．

情報を絞り込むための検索条件として，入力した検索語のすべてを含むページを出力するAND条件による検索がある．この絞り込みによる検索を繰り返すことで目的の情報を得るためには，絞り込みの途中で目的の情報を外さないように検索語を選んで用いる必要があり，もし目的の情報を外してしまい検索結果のWebページ件数が少なくなった場合には，いくつかの検索語を削除して検索条件を緩めた上で異なる検索語を用いる必要がある．この，検索語を削除するという操作は「後戻り」であり，検索の効率の改善とは，この後戻りをいかにして少なくするかということである[臼澤 99]．後戻りをなくすためには必要な情報を逃さないように，徐々にではあっても確実に目標に到達できる検索条件の作成が必要となる．そのためにも，絞り込みのみによる検索の欠点である，一つの情報に異なる単語表現が存在することによる情報の漏れをなくさねばならない．

情報の漏れをなくすための検索条件として，入力した検索語のうちのいずれか1つを含むページを出力するOR条件による検索がある．ユーザが入力した検索語に関連する単語をOR条件として追加することで，関連する多くの情報を得ることが可能となる．そこで，情報の範囲を広げるために検索語に関連するキーワードとして，**関連キーワード**を提供するシステムもある．関連キーワードは，ユーザの入力した各検索語を含むページから抽出される単語や，Webページ内の単語の共起頻度から学習された各検索語と関連度の高い単語となる．そのため関連キーワードは，情報を絞り込むよりもむしろ情報の範囲の拡張に有効となる．関連キーワードを提供するシステムにはサーチエンジンExcite[Excite]，AltaVista[AltaVista]，Infoseek[Infoseek]



がある．しかし，Exciteなどの検索語がOR条件として扱われるサーチエンジンにおいては，情報の絞り込みが全てシステム任せとなり，サーチエンジンによるランキング（Webページの評価）が絞り込みを代替するために，欲しい情報が上位に現れず目的のページが得られない場合があるとともに，似通ったページばかりが検索される結果となる．また，ユーザが複数の検索語を入力した場合においては，全ての検索語が同等に扱われてしまい，提供できる関連語の数の制約からユーザの意図と異なる検索語の関連キーワードが多く出力されることもある．したがって，OR条件のみによる検索では関連語が提供されていても，十分な検索を行なうことができない．

これら以外のキーワード提供を行なうシステムとして，シソーラスをユーザに提供することで検索を支援するシステムThe thesaurus-step2[Thesaurus]がある．すなわち，辞書に含まれる類義語をユーザが望む検索語の関連語として得ることができ，ユーザの知識不足を補うと共に，新たな検索語として追加することができる．しかし，類義語辞書には多くの類義語が含まれており，ユーザの知識を補う効果は得られるが，検索に用いた類義語が必ずしも，実在するWebページ中に用いられているとは限らない．

これらのユーザの興味に関連するキーワードを提供する代わりに，ユーザの検索履歴からユーザの興味そのものを学習することでユーザの興味に関連するキーワードを補って検索を行なうアプローチについて次項で述べる．

### 2.2.3 ユーザの興味を学習することによる解決

ユーザの検索時における興味の推定に関して，あらかじめユーザの興味を学習する手法が存在する．ユーザの検索履歴をもとにユーザの興味を推定し，ユーザの興味に関係する単語をユーザに教示するシステム[Ohsawa 97, Bruza 97]や，検索履歴からユーザの興味を帰納学習や，ベイジアンネットワーク[Charniak 91]，遺伝的アルゴリズム[Goldberg 89]などの手法よって学習することでユーザモデルを作成し，自

動的にユーザの興味に関する検索語を補完するシステム [Santos 99, Balabanovic 95, Armstrong 95, 渥美 97, Krulwich 95] などがある．

しかし，サーチエンジンを使うユーザの興味を学習するには時間がかかるために，サーチエンジンの使用頻度とユーザの興味の変化の速さ，さらにはWeb情報の変化の速さを考えると極めて短い時間のユーザの検索履歴からユーザの一時的な興味を学習しなければならない．そのため，ふいに思い立った検索要求に対しては学習が効果を発揮できず，その場限りの検索に対応することができない．また，学習された結果を用いて検索を行なうためには，ユーザが普段使っている端末もしくはその端末にアクセスできる環境からでないと用いることができない．

ユーザの検索環境を整えるという点においては，検索におけるインターフェイスが重要となる．次節では，検索の結果としてどのようなWebページが出力されているかを素早く理解するための検索結果の出力の方法について，研究の背景を述べる．

#### 2.2.4 検索結果の可視化による解決

サーチエンジンの検索の結果得られる多くのWebページに，どのような情報が多く含まれているかをいち早く理解することは，必要なキーワードを探し出してユーザの興味を具体化することの助けとなる．また，欲しい情報が検索結果に含まれていないことを素早く理解することは，次に再検索を行なうまでの時間の短縮にもつながる．検索結果の出力手法として，検索結果として得られたWebページをクラスタリングしてユーザに提供する研究 [Hearst 95] や，予め用意されたカテゴリに分類する研究 [仲川 99]，またサイトごとに分類を行なって結果を表示するシステム [Lycos] などがある．これはユーザがクラスタリングされたWebページを眺めることによって，ユーザの興味を含むクラスタを順次選んでいくことで，目的の情報へ導くことを目指している．また，関連の強いオブジェクトを近くに配置することによる発想支援 [角 94] をもとに情報の視覚化を行なう研究 [村田 99] もある．

しかし，雑多なWebページ集合をもとにしてユーザの多様な興味に合致するクラスタを生成することは難しい．そのため，生成された各クラスタの意味づけが難しく，どのクラスタがユーザの興味に対応しているかの判断に手間がかかることや，興味を表すクラスタが複数にまたがるという状況に陥ることがあり，検索を続けるのが困難になることもある．あらかじめカテゴリを用意しておく場合においても，ユーザの興味にふさわしいカテゴリが存在するとは限らず，ユーザの観点に基づいたカテゴリを用いて分類を行なうのは容易ではない．



また従来の情報検索インターフェイスにおいては，データベースとなっているWebページ空間の表現を中心としている [Lamping 95, 塩沢 97]．しかし，急激に増加し続ける膨大な量のWebページの視覚化には限界があり，局所的な視覚化にならざるを得ない．文書集合を単語によって可視化する研究 [渡部 99] もあるが，すべての単語間の関連度を学習しているために，WWW上の膨大なWebページに対しての適用が困難である．

次節では，ここまでで述べてきた従来研究における検索の問題点の解決策および新たに残された問題点を踏まえた上で，本論文で提案する興味表現支援システムの位置付けを明確にする．

## 2.3 興味表現支援システムの位置付け

本興味表現支援システムが含むサーチエンジンは，ユーザが入力した検索式に厳密にマッチするページのみを出力する．このことは，曖昧検索やキーワードベクトルによる検索の特徴である，検索条件に完全に当てはまらなくてもWebページが検索されるという特性を外すことになる．しかしユーザの意図と無関係に多くのWebページを提供するよりむしろ，ユーザ自らが多くのキーワードを用いて情報を集める検索の方が効率良く情報を収集できると考え，情報収集の材料として実際のWebページからのキーワードを提供する．提供するキーワードは絞り込みキーワードと関連キーワード[5]の2種類であり，ユーザはAND条件とOR条件の両方を組み合わ

[5]本システムにおいては興味キーワードと呼ぶ

せた検索式を作成し、システムはユーザ自身の興味を具体的かつ的確に表す検索式を作成するように導く．

2.2.2項で述べたAND条件による情報の絞り込みとOR条件による情報の範囲拡張は，検索の目的に応じても行なわれる必要がある．絞り込み検索は情報が１つだけ欲しい場合（天気予報，円相場，歴史の年号など）には有効であるが，より多くの情報が欲しい場合（関連文献の検索，画像収集など）にはOR条件の方が有効となる場合が多い．またいずれの場合にも，欲しい情報の検索からの漏れを防ぐために，欲しい情報の全体を包み込むように，情報を徐々に絞り込む必要がある．すなわち，絞り込みキーワードで情報を絞り込む際に，関連キーワードで必要な情報の範囲を広げて過度の絞り込みを避けるという，AND条件とOR条件の適切な組合せによって検索を行なうことで有効な検索が可能となる．

本システムにおいてはユーザ自身が検索式を作成し，徐々に興味を具体化して目的の情報に近付くことを支援する．すなわち，毎回の検索に対してシステムはユーザにキーワードを提供し，ユーザはそれらのキーワードを適宜用いた検索条件を作成するという一連の作業が繰り返される．これはユーザとシステムとの間のインタラクションであり，関連フィードバック技術と対比される．関連フィードバックにおいてユーザがシステムに返す情報は個々のWebページであり，ユーザの興味が明確な言葉（単語）として次の検索に反映されないばかりか，ユーザの興味の具体化の助けにも効果が発揮されない．本システムにおいては，キーワードを用いてユーザとのインタラクションを行なうために，検索されたWebページ情報を明確なキーワードによってユーザに提供するとともに，ユーザはそれらのキーワードの中から，自身の興味と強く関連するキーワードを選びシステムに返すことで，ユーザの意図が確実に反映される．たとえば出力として20のWebページが得られた時に，それらのページを別個に要・不要のラベルをつけてフィードバックすることに比べ，その20ページ全体を特徴付けるキーワードを有効に活用して情報を探す方が，労力や効率の点において有利である．

ユーザの興味を学習するアプローチは，ユーザの興味が学習されれば検索に有



効とも考えられるが，学習に速度が要求される上に，学習が利かない突発的な検索要求に対しては効果が発揮できなかった．これに対し本研究では，一回の検索ごとに，ユーザの興味と存在する情報との間のギャップを埋める検索語をユーザに提供し，ユーザの時々に応じた興味に対応できる検索支援システムを目指している．

ユーザの興味の具体化という点において，検索結果がどのようなWebページから構成されているかを素早く摑むことは，検索効率の向上のためには非常に重要となる．そのため，検索結果のWebページをクラスタリングする試みは有効である．それでもなお，クラスタリングが十分な精度で行なわれないことを想定して，類似するページを集めてグループを作るクラスタリングは行なわない．本システムはユーザの興味に主眼をおき，ユーザの一つの興味を構成する各検索語もしくは検索語の組合せに関わる関連キーワードを得てそれらを分類することによって，ユーザの興味が具体的に表されていない検索語を示唆する．それに加え，ユーザの興味との関わりにおいて，どのようなWebページがWWW上に存在するかをキーワードによって明示する．すなわち，Webページ空間の標本となる部分空間をキーワードを用いて表現することで，存在する情報の視覚化を目指す．

キーワードを明示してユーザに提供するために，二次元平面インターフェイスを本システムは備えている．従来のサーチエンジンの絞り込みおよび関連キーワードは，一次元に並べられてユーザに提供されている．しかし，並列に並べられたキーワードにはそれらを分類し区別するための特徴がないため，多くのキーワードを提供することができない．また，三次元のインターフェイスについていえば，三次元空間上へキーワードを配置することを考える場合，まず三次元空間をどの方向から眺めるかという視点の問題が生じる．じっくり，腰を据えた検索を行なって必ず情報を獲得する意気込みがユーザにあればよいのだが，多くのユーザは検索をできる限り素早く行ないたいと考えるため，視点を色々と変えながらキーワードを探し出す操作は手間がかかる．また，ある視点から三次元空間を眺めている一時点において，その空間はディスプレイ上では二次元平面となっている．加えて，キーワードベクトルなどで表されるキーワードの空間は多次元であるため，それを二次元で表

す場合と三次元で表す場合との間で，表せる情報量に大きな差が現れにくい．特に三次元で表すことの有効性が確認されず，三次元上に見やすくキーワードを配置することで，存在する情報を容易に理解できる画期的な技法がない現時点では，普段から Windows や Macintosh の計算機上のファイル整理で使いなれている二次元平面インターフェイスの方が優れる．そのため，GUI（Graphical User Interface）を用いた検索語入力のためのインターフェイスに関する研究 [Beaulieu 97] もあるが，検索結果を視覚化してそのまま再検索を行なえるという枠組にはなっていない．

ここまで本章で述べてきた検索の現状と問題点およびその解決策とを踏まえ，さらなる検索効率改善のために本論文で提案するシステムは，ユーザが入力する検索語に関連する 2 種類の絞り込みキーワードと興味キーワードとをユーザに提供し，ユーザの興味との関わりにおいて，存在する Web ページ情報をキーワードを用いて二次元平面インターフェイス上に表すことでユーザの検索を支援する．この，ユーザ自身の興味を具体的かつ的確に表す検索式を作成するための興味表現支援システムについて次章以降で述べる．



# 3　新仮説生成による興味表現支援システム

本章では，ユーザの検索式更新による再検索を支援するために，ユーザに有効なキーワードを提供する興味表現支援システムについて述べる．本システムにおいてはユーザが検索式を作成し，システムはユーザによって入力された検索式（検索語の組合せによる検索条件）をもとに，ユーザが入力した全ての検索語で表される興味に関連する**絞り込みキーワード**と，一部の検索語で表されるユーザの興味の一部に関連する**興味キーワード**とを検索支援キーワードとしてユーザに提供する．新仮説生成による興味表現支援システムの全体構成を図 2に示す．

検索を行なうユーザは，初めに Web ブラウザ上の検索インターフェイスから，自身の興味を表す検索式（ユーザが入力する検索語全体の組合せ，詳細は 3.1節）を入力として与える．入力された検索式は，絞り込みキーワードを得るための処理を行なうモジュール（図 2枠内左側）とユーザの興味に関連する興味キーワードを得るための処理を行なうモジュール（図 2枠内右側）とにそれぞれ与えられる．

左の絞り込みキーワードを得るための処理の流れのうちで，サーチエンジンからの出力として Web ページをユーザに返す部分までが最も単純な検索システムのプロセスである．絞り込みキーワードは，サーチエンジンの検索結果としてユーザに提供される Web ページ集合の中から抽出される．それらのキーワードは検索結果となった Web ページ集合の中の一部の Web ページに含まれるため，情報を絞り込むために用いることができる．

また，右の興味キーワードを抽出するための処理の中においては，検索式の一部にのみ焦点を当てる．すなわち，多くの Web ページ作成者が一つの Web ページ中で同時に用いやすい単語の間には相関があると仮定し，その相関の原因となっている一つの話題を大衆（多くの Web ページ作成者）の興味とみなす．この大衆の興味に一致するユーザの興味を検索式中から発見し，ユーザの興味に関連する Web ページを新たに獲得した後に，獲得された Web ページから抽出されるキーワードを興味キーワードする．

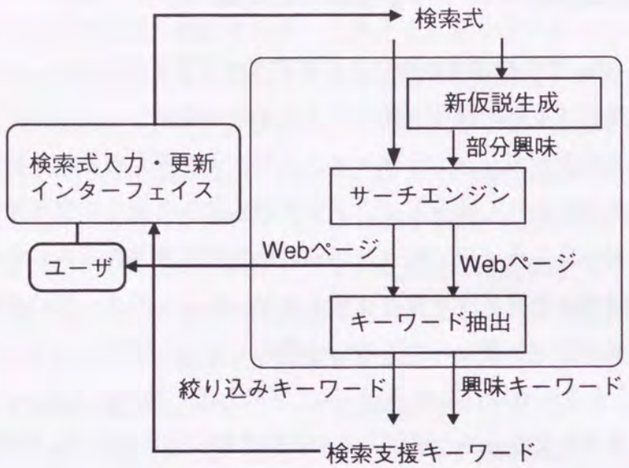


図 2: 興味表現支援システム（新仮説生成）

最終的に，これらの絞り込みキーワードと興味キーワードがシステムの出力（検索支援キーワード）となりユーザに提供される．ユーザは提供されたキーワードを用いて，自身の抱いている興味と実際に存在する Web ページとを結び付けるキーワードを選び採って検索式を更新する．すなわち，ユーザ自身にしか知り得ないユーザの興味を，実際の Web ページに用いられている単語を用いて，正確かつ具体的な検索式としてユーザが表現する．ユーザは検索結果の中から目的の Web ページを発見するまで，検索式の更新とシステムからのキーワードの享受を繰り返す．ユーザは本システムとの間のインタラクションによって，自身の興味を具体化して目的の情報を得ることができるとともに，新たな関連情報を取得することも可能である．

以下では興味キーワードを抽出する処理の手順（図 2枠内右側）にしたがって説明を行なう．3.3節のキーワード抽出の節においては，興味キーワードの抽出法と合わせて絞り込みキーワード抽出の説明を行なう．



## 3.1 サーチエンジンへの入力：検索式

サーチエンジンへの入力は検索式で与える．**検索式**とは複数の検索語を AND や OR の論理演算子を用いてブール代数式（**Boolean**）で表現し，検索条件としたものである．たとえば,「関西でだんごを食べながら観光したい」ユーザがいたとすると,「（京都 OR 奈良 OR 神戸）AND（観光 OR 名所）AND だんご」なる検索式が作成される．

現在，サーチエンジンの多くは同時に入力された検索語をすべて含む（AND 条件）Web ページを検索するもの [goo, Yahoo] が多数であり，入力された検索語のいずれかを含む（OR 条件）Web ページを検索するサーチエンジン [AltaVista, Excite] も存在する．

Boolean による検索を行なうためには，AND や OR の論理条件をユーザが理解して用いる必要があるため，初心者には不向きな面もある．しかし，Boolean による検索は多くのサーチエンジンでサポートされており，今後の WWW 情報の増大にともなって，ユーザの多様な興味をより具体的，的確に表現する必要が増すと考えている．また，5章で述べる検索支援インターフェイスにおいては，AND や OR の論理条件をユーザが入力せずに Boolean の検索式を表すための二次元インターフェイスを構築する．

AND や OR 以外の検索方法として，指定した検索語を含むページを出力から除くための NOT 条件や，AND 条件を厳しくして入力した複数の検索語がその入力順に出現するページを検索する方法，またベクトル空間法に基づいて，各 Web ページを表すベクトルとユーザが入力した検索条件を表す検索ベクトルとの間でマッチングをとる方法[6]などさまざまな検索条件が存在する．しかし，検索要求の高まりに応じた複雑な検索条件の追加は，ごく一部のコンピュータに慣れている人にのみ用いられるだけであり，情報の増加に対応できる本質的な解決とはなっていないために，一時しのぎの解決策と言わざるを得ない．

[6]検索結果の上位は AND 条件による結果，下位には OR 条件による結果が現れる

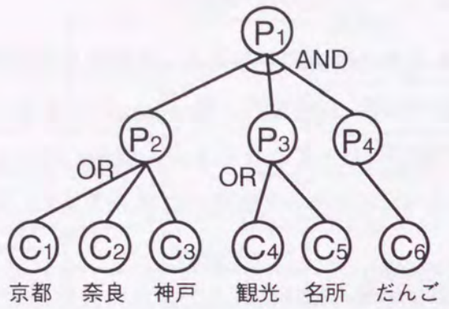


図 3: 検索式のネットワーク

ユーザに入力された Boolean による検索式は，3.2節で述べる新仮説を生成するために，Directed Acyclic Graph, DAG[Pearl 93] と呼ばれる有向非循環の木構造のグラフ（ネットワーク）に変換される．たとえば図3は「(京都 OR 奈良 OR 神戸) AND（観光 OR 名所）AND だんご」なる検索式のネットワークを表す．すなわち，ユーザが入力した各検索語を葉ノード，また AND や OR の論理演算子によって結びつけられた部分検索式を中間ノードとして検索語間の関係を表したネットワークを作成する．この木構造のネットワークが次節で述べる新仮説生成部への入力となる．

## 3.2 ユーザの興味に関連する新仮説生成

本節では 3.1節で作成された木構造のネットワークをもとに，各葉ノードに当たる検索語間の相関関係を調べることで互いに依存関係にある検索語を特定し，ユーザの興味と Web 上の情報との関わりを示す新仮説を生成する方法について述べる．

### 3.2.1 ユーザの興味と新仮説の関係

ユーザが心に検索要求となる興味を思い立った後に，どのような思考過程を経て検索式を入力するかについて説明する．たとえばユーザが,「関西でだんごを食べながら観光したい」と考えたとする．このときユーザはまず，自身の興味を構成する要素として,「関西」「だんご」「観光」という 3 つの部分があることに気付く．するとユーザは，単純に「関西 AND 観光 AND だんご」という検索式を作成してサーチエンジンへの入力とすることもできるが，より的確にユーザの興味を表す検索条件を入力するために，各興味の部分（一部の検索語）に関して興味を具体化する．すなわち,「関西」の具体的な表現として「京都」「奈良」「神戸」が，また「観光」の類義語として「名所」という検索語が思い起こされる．その後再び，ユーザの「関西」という興味の一部が「京都 OR 奈良 OR 神戸」と,「観光」が「観光 OR 名所」という検索語の組合せと置き換えられ，具体的かつ情報の漏れをなくした検索式として，最終的に「(京都 OR 奈良 OR 神戸) AND (観光 OR 名所) AND だんご」という検索式がユーザの頭の中に構成される．



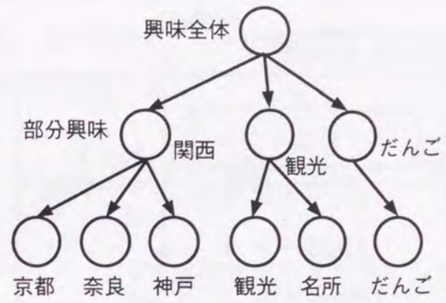


図 4: 検索式作成過程

ユーザが検索式を作成する過程（図 4）を換言すると，ユーザは自身のもつ 1 つの興味をまず構成要素に分解し，各要素に関する検索語を用意した後に，検索語を組み合わせて検索式を用意することになる．本システムにおいては，このユーザの興味を構成する各部分（**部分興味**）に着目する．すなわち，実際の WWW 上で同時に用いられやすい（1 つの Web ページ内で同時に現れやすい）検索語間に相関を認め，その相関の原因として存在する 1 つの概念をユーザの興味の一部として，それ

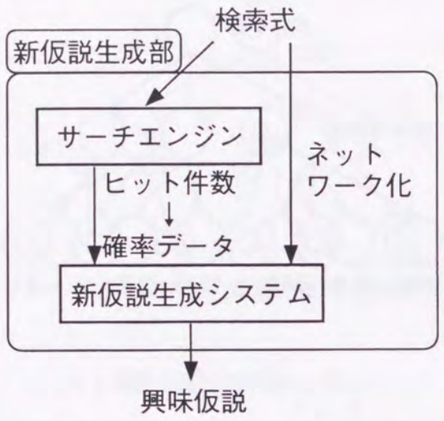


図 5: 興味仮説生成手順

に相当する新仮説を生成する．ここで生成される新仮説は，すべての部分興味を網羅しないが，大衆（Web ページ作成者）の興味（単語間の相関）に一致するユーザの興味に着目することで，ユーザと WWW 上の大衆とをつなぐ興味の仮説として捉えることができる．そこで，この生成する新仮説を**興味仮説**と呼ぶ．

興味仮説は図 5の手順にしたがって生成される．すなわち，ユーザが入力した検索式をもとにして，まず前節で述べたネットワーク化された検索式が 3.2.3項の新仮説生成システムに与えられる．それに加えて，作成されたネットワーク上の各ノードに対応する検索語（検索式）をサーチエンジンに与えた時のヒット件数を確率データに直し，合わせて新仮説生成システムへの入力とする．次項でこの各ノードに与える確率値と，検索語間の依存関係を測る指標となる依存度を定義する．

### 3.2.2 ノードの確率値と依存度の定義

検索時点における Web ページ中において，1つのページに同時に現れやすい検索語を調べるための道具として確率を用いる．ユーザが入力した検索式を表すネット



ワーク上の検索語や部分検索式を表す各ノードの確率値 $P$は，検索対象となる Web ページの総数に対する，各検索語または部分検索式 $E$にマッチする Web ページ数の割合として式 (1) のように定義する．

$$P(E) = \frac{E\text{にマッチする } Web \text{ ページ数}}{\text{全 } Web \text{ ページ数}} \tag{1}$$

次に，検索語の Web ページ中の共起度を表す指標として，**依存度**を定義する．単語間の相関を計る方法としては，相関係数などの統計的指標が存在するが，より少ないデータとして各単語毎の出現頻度を表す確率データのみから相関を計ることはできないため，以下の定義に基づく依存度を用いる．まず，式 (2) から式 (7) がそれぞれ，$A$を親ノード，$X_i$を子ノードとした時の親ノードの確率値の，上限，独立，下限となる値である [Bonissone 87]．

**AND ルール：** $A \leftarrow X_1 and X_2 and ... and X_n$

$$\overline{P(A)} = \min\{P(X_1), P(X_2), \cdots, P(X_n)\} \tag{2}$$

$$P(A) = \prod_{i=1}^{n} P(X_i) \tag{3}$$

$$\underline{P(A)} = \max\{0, \sum_{i=1}^{n} P(X_i) - n + 1\} \tag{4}$$

**OR ルール：** $A \leftarrow X_1 or X_2 or ... or X_n$

$$\overline{P(A)} = \min\{1, \sum_{i=1}^{n} P(X_i)\} \tag{5}$$

$$P(A) = 1 - \prod_{i=1}^{n}(1 - P(X_i)) \tag{6}$$

$$\underline{P(A)} = \max\{P(X_1), P(X_2), \cdots, P(X_n)\} \tag{7}$$

そこで依存度をノード $P_i$の子ノードが互いに独立と仮定した時のノード $P_i$のとる確率と，実際にノード $P_i$がもつ確率値との差を，最大で 1 となるように正規化した値として定義する．すなわち，ルール $r$における依存度 $depend(r)$ を，式 (8) と式

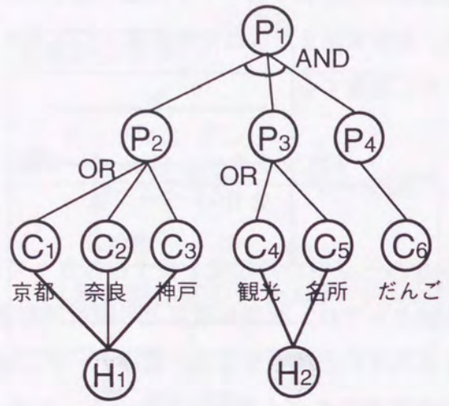


図 6: 興味仮説発見システムの動作例

(9) のように与える．ただし，$a$ はルール $r$の条件事象が互いに独立な時の結論 $A$ の確率である．

**AND ルール：** $A \leftarrow X_1 and X_2 and ... and X_n$

$$depend(r) = \frac{P(A) - a}{\overline{P(A)} - a} \tag{8}$$

**OR ルール：** $A \leftarrow X_1 or X_2 or ... or X_n$

$$depend(r) = \frac{P(A) - a}{\underline{P(A)} - a} \tag{9}$$

たとえば，図6の各ノードの表す検索式と確率値が表1のように与えられる[7]．

すると,「京都」「奈良」「神戸」の各確率値から，ノード $P_2$「京都 OR 奈良 OR 神戸」の確率値の範囲が式 (5) から式 (7) をもとに，式 (10) から式 (12) のように定められる．

$$\overline{P(A)} = 1.639 + 0.558 + 0.755 = 2.952 \tag{10}$$

[7]検索対象は 1999 年 12 月当時，サーチエンジン [goo] が対象としていた 3500 万件の文書である．



表 1: ノードの表す検索式とその確率値と依存度

| ノード:検索式 | ヒット件数 | 確率 (%) | 依存度 |
|---|---|---|---|
| $C_1$:京都 | 573694 | 1.639 | –.– |
| $C_2$:奈良 | 195331 | 0.558 | –.– |
| $C_3$:神戸 | 264272 | 0.755 | –.– |
| $C_4$:観光 | 458336 | 1.310 | –.– |
| $C_5$:名所 | 61591 | 0.176 | –.– |
| $C_6$:だんご | 24487 | 0.070 | –.– |
| $P_1$:(京都 OR 奈良 OR 神戸) | | | |
| AND(観光 OR 名所)AND(だんご) | 799 | 0.0023 | 0.03 |
| $P_2$:京都 OR 奈良 OR 神戸 | 876737 | 2.505 | 0.33 |
| $P_3$:観光 OR 名所 | 489043 | 1.397 | 0.50 |

$$P(A) = 1 - (0.98361 * 0.99442 * 0.99245) = 2.926 \tag{11}$$

$$\underline{P(A)} = \max\{1.639, 0.558, 0.755\} = 1.639 \tag{12}$$

そして，実際には $P(A) = 2.505$ であることから，式 (9) よりノード $P_2$の依存度[8] が式 (13) のように計算される．

$$depend(r) = \frac{2.505 - 2.926}{1.639 - 2.926} = \frac{0.421}{1.287} = 0.327 \tag{13}$$

同様にして，表1の $P_1, P_3$の依存度も計算される．依存度によって計算される単語間の相関は，どれだけ多くのページで単語が同時に用いられているかを表す指標となっている．すなわち，この依存度の値が高いということは，多くの Web ページ

[8]正確にはノード $P_2$とそれらの子ノードからなる OR ルールの依存度であるが，便宜上，親ノード $P_2$が依存度をもつとして話を進める．

作成者が同時に言葉を用いていることの証しであり，強いては大衆（Webページの作者と閲覧者）の興味が集中していると解釈できる．この意味において，大衆の興味の強く現れている相関がわかるということは，キーワード選択に関して，全く何の情報も指針も与えられない一度限りの検索においては有益な情報であり，ユーザ自身の特性を表す主観データが得られない以上，このような客観データを用いる方法は妥当であると考えられる．そこで，ここまでで得られた木構造のネットワーク，各ノードの確率値および依存度をもとに，次項のアルゴリズムによってユーザの興味を表す新仮説を生成する．

### 3.2.3 新仮説生成アルゴリズム

ユーザの興味を表す新仮説は次の手続きによって生成される [砂山 99]．

親ノード集合：$\{P_i\}$（$i = 1..n : n$ は親ノード数）

依存ノード集合：$\{D_j\}$（$j = 1..m : m$ は依存のあるノード数）

依存説明ノード集合：$\{L_{jk}\}$（$k = 1..t : t$ は依存ノードの子ノードの数）

1. 各親ノード $P_i$の依存度を計算する．
2. しきい値（詳細は後述）を超える依存度をもつノードを特定し依存ノードとする．
3. 各依存ノード $D_j$について，その各子ノード $\{C_k\}$ を起点として，$C_k$から子孫方向に辿った時に，途中の中間ノードが表す検索式が，それらの子ノードの検索式（検索語）を AND で結びつけていれば確率の小さい方へ，OR で結びつけていれば確率の大きい方へたどった先にある葉ノードを $L_{jk}$とする[9]．
4. 各葉ノードの組合せ $L_{j1}, L_{j2}, .., L_{jt}$に新仮説として子ノードを生成する．

[9]ある結果の原因として複数の要因が考えられる場合，その複数の要因が OR で結ばれるならば確率の大きい原因が重要であり，逆に AND で結ばれるなら，確率の小さい原因が重要となる．



たとえば，表1のように確率値と依存度が与えられた場合，しきい値以上の依存度をもつノードを特定する．現在この初期しきい値は 0.3 として設定している．本システムにおいては，検索時に存在する Web ページにおける検索語間の相関を求めることが目的であるから，確率誤差は考えない．そのため，依存の有無を判別するための依存度のしきい値を任意に設定できる．そこで，依存度と同様に区間 [0,1] 内の値によって正の相関を調べられる相関係数において，弱い相関があると考えられる 0.3 前後の値として，依存度のしきい値を 0.3 と設定した[10]．すると，ノード $P_2, P_3$ が依存ノードとなる．次に依存説明ノードを各依存ノードを起点として探索を行なう．ノード $P_2$に関して見ると，$P_2$の子ノードがすでに葉ノードであるために，その3つの葉ノードがそのまま依存説明ノードとなる．そこで3つのノード（検索語）間の相関の原因として新ノード $H_1$をネットワークに生成する．同様にして，ノード $P_3$に対しても新ノード $H_2$が生成される．

ここで生成された新ノード $H_1$は，「京都」「奈良」「神戸」の3ノード間の依存関係を示唆するものであり，もともとユーザの関西という部分興味から与えられた検索語である．しかし，システムが知ることができるのは「関西」ではなく「京都」「奈良」「神戸」の3つの検索語であるために，これらをもとに「関西」を推定する枠組が必要となる．すなわち，相関があると推定された検索語間の共通概念および，共通概念に関連する単語を実際の Web ページ中に求める．

## 3.3 検索支援キーワードの抽出

ユーザの興味に関連する検索支援キーワードをユーザに提供する．提供するキーワードは次の2種類である．

1. **絞り込みキーワード**：ユーザが検索結果を絞り込むためのキーワード

[10]しきい値を越える依存度を持つノードが存在しない場合は，最低1つの依存ノードを見つけるまで，しきい値が 0.01 ずつ下げられていく．

2. **興味キーワード**：ユーザが情報の漏れをなくし，幅広く情報を集めるためのキーワード

本節では，この絞り込みキーワードおよび興味キーワードを抽出する方法について述べる．

### 3.3.1 絞り込みキーワード抽出のための Web ページ獲得

絞り込みキーワードは，検索の結果得られたページを絞り込むためのキーワードであり，主に AND 条件として単語を追加することで情報を絞り込む目的に用意する．したがって，検索の結果得られたページから抽出される単語であれば，情報が確実に絞り込まれると同時に，検索された Web ページ全体がどのような内容の Web ページであるかをユーザが知るために有効となる．すなわち，サーチエンジンが出力する Web ページの要約一覧以外に検索結果をユーザに知らせる手段として，それら検索された Web ページに含まれるキーワードの集合をユーザに提示する．これらのキーワードは，検索結果となった Web ページ集合の全体像を素早く把握するために役立てられ，ユーザは検索されている情報に応じた再検索を行なえる．

### 3.3.2 興味キーワード抽出のための Web ページ獲得

本項では，3.2節で得られた興味仮説から，ユーザの興味に相当する興味キーワードの抽出元となる，Web ページを集める方法について述べる．この操作は，興味仮説が示す依存関係にある複数の検索語が，1 つの Web ページ中に同時に現れやすい原因を探るために，その原因が含まれる Web ページを集めることが目的である．したがって，依存関係にある複数の検索語を同時に含むページ中には，それらの検索語に共通する概念を表す単語や，その共通概念に関連する単語が含まれていると考えられる．そこで，興味仮説の親ノードとして存在している複数の検索語をすべて同時に含む Web ページを獲得する．これは，それらの検索語を AND 条件でつないだ検索式をサーチエンジンに与えることで実現される．たとえば，「京都」「奈良」



「神戸」の間に興味仮説が生成された場合，「京都 AND 奈良 AND 神戸」という検索式を作成してサーチエンジンに与えることで，「京都」「奈良」「神戸」の 3 つの検索語をすべて含むページを獲得する．

これは事象の共起性からの発見 [Langley 87] の考え方に基づいて，ユーザが複数の検索語を同時に用いた原因を発見することを目的として，その手がかりを得ることに相当している．こうして集められた Web ページはユーザと同じ興味である一つの概念を含み，かつその概念に基づいて構成されていると考えられるため，検索式「京都 AND 奈良 AND 神戸」から得られる Web ページには，「関西」や「近畿」といった共通概念に相当する単語や，その共通概念に関連する「大阪」「滋賀」「兵庫」などの単語が含まれると予想される．

### 3.3.3 Web ページからのキーワード抽出

サーチエンジンに検索式を入力することで得られた Web ページ集合から，キーワードを抽出する方法について述べる．キーワード抽出は次の手順で行なわれる．

1. 検索式をサーチエンジンに与えて得られる Web ページの中から，サーチエンジンの順位付けによる上位 20 ページを取り出す．(このページ集合を *documents* とする)
2. 形態素解析（[ChaSen] のシステムを適用）を行ない，各ページから名詞を抽出する．
3. 2. で取り出された各名詞に，式 (14) による評価値を与える．

ある名詞 $A$ の評価値 $value(A)$ は，以下のように定める．ただし，$tf(A,d)$ は文書 $d$ 中の名詞 $A$ の出現回数である．

$$value(A) = \sum_{d \in documents} \log(1 + tf(A,d)) \tag{14}$$

この評価関数は，*documents* 中の多くの文書に多数回出現する名詞ほど高い評価値が与えられる[11]．この評価関数は tf×idf 法 [Salton 97] のように 1 つの文書に現れ他の文書には現れにくい単語に高い評価を与えることはせずに，多くの文書に共通に出現する単語を評価している．これは，各文書に固有のキーワードよりも，ユーザの興味に関係する多くの文書に含まれる語の方が，ユーザが抱いている興味を徐々に具体化して，存在する Web 情報と必要十分に結び付けるために有効だと考えたからである．

提供するキーワードの数はまず，それぞれ 16 個ずつ抽出する[12]．また，絞り込みキーワードは絞り込みに役立つキーワードとして，絞り込みキーワードの抽出元の Web ページ集合を 2 割から 8 割に絞り込める単語に限定する．そのため出現頻度の低いユーザの望む単語が外される可能性はあるが，このような頻度の低い単語は数多く存在するために，その全てを考慮した上でユーザの興味に当てはまる単語を選ぶことは困難であり，却ってノイズとなる単語が多く含まれることを避ける．またユーザが自身の興味に当てはまる検索式を作成して絞り込み検索を進めるうちに，検索に有効な単語はいずれ出力に含まれることもある．興味キーワードも同様に，OR 条件として追加した際に情報が確実に追加されるために，興味キーワードの抽出元の Web ページの 2 割以上に含まれる単語をキーワードとする．ここで絞り込みキーワードに比べて条件が緩いのは，興味キーワードの役割として換言可能な単語は必要と考えたからである．

## 3.4 キーワード抽出実験

本節では，新仮説生成に基づく興味表現支援システムによって，どのようなキーワードが抽出されるかを実験し，提供されるキーワードを用いた検索式更新の例を簡単に述べる．(検索式更新についての詳細は 5.3節で後述する) 実験環境は，SUN SPARK Station 10GX(64MB) であり，実験プログラムは Perl で書かれている．また，サーチエンジンには goo[goo] を用いた．

[11]「ホーム」「ページ」など対象とするデータベース全体に対して，キーワードとして不適切な名詞は，予めシステムのノイズ知識として与え，キーワードとして選ばない．

[12]16 という値は，多過ぎない範囲でできるだけ多くのキーワードを提供するための数として実験により定めた．



### 3.4.1 実験 1 (テレビゲームに関する検索式による実験)

最近のテレビゲームのゲームソフトの紹介記事を探すために，あるユーザがサーチエンジンへの入力として，

「(シミュレーション OR アクション OR ロールプレイング ) AND ゲームソフト AND 紹介」

という検索式を作成した．この検索式を興味表現支援システムに入力した結果,「シミュレーション」「アクション」「ロールプレイング」が Web データベース中で同時に表れやすいことを示す興味仮説が生成され (図 7)，図 8のキーワードがシステムの出力として得られた．

出力されたキーワードのうち，絞り込みキーワードの中には，情報を絞り込むために有効なキーワードが含まれている．たとえば,「サターン」「プレイ (ステーション)」というゲームのハード名，また「タイトル」「内容」というゲームソフトに関して，ゲームソフトのタイトルを調べるのか，ゲームソフトの内容も知りたいのかを明確にするためのキーワードが出力された．

この出力されたキーワードを用いて，ユーザがハードウェアとして「サターン」を持っていた場合，

「(シミュレーション OR アクション OR ロールプレイング ) AND ゲームソフト AND 紹介 **AND サターン**」

と検索式を更新して情報を絞り込むことができる．

また興味キーワードの中には，ユーザの部分興味そのものを表す「ジャンル」というキーワードが含まれていると同時に,「シューティング」「スポーツ」「アドベンチャー」「パズル」など，検索式には含まれなかったが，ユーザの興味 (ゲームのジャ

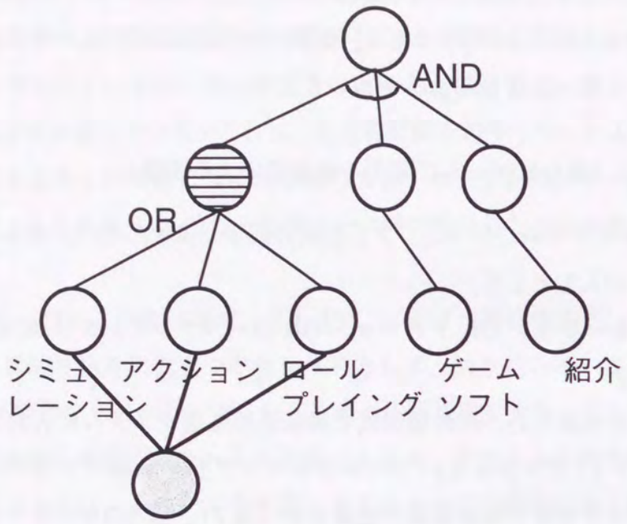


図 7: 実験「テレビゲーム」結果

ンル名）に関連するキーワードが含まれている．そこで，ユーザが「パズル」ゲームにも興味を持っていた場合，

「（シミュレーション OR アクション OR ロールプレイング **OR パズル**）AND ゲームソフト AND 紹介 AND サターン」

と検索式を更新して，ユーザの興味に関する情報をより多く獲得することができる．このように，本システムによるキーワードは情報を絞り込むための検索と，情報の漏れをなくすための検索の両方に用いられ，ユーザの興味を正確（必要十分）に表した検索式の作成が可能となる．



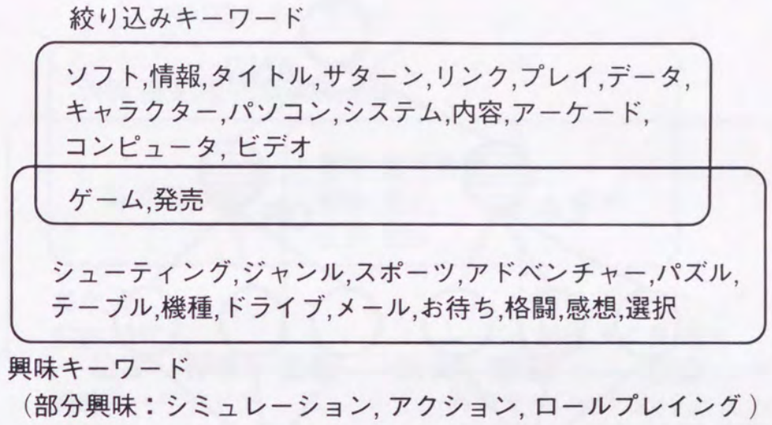


図 8: 実験「テレビゲーム」結果

### 3.4.2 実験 2（日本史に関する検索式による実験）

ユーザが，江戸時代末期の事件である「禁門の変[13]」の関連事項を調べようとしたが，その「禁門の変」という用語を思い出せなかったために，サーチエンジンへ

「（会津 OR 薩摩 OR 長州）AND（慶喜 OR 孝明 OR 容保）」

という検索式を入力として与えた．その結果,「会津」「薩摩」「長州」の間と,「慶喜」「孝明」「容保」の間にそれぞれ興味仮説が生成され（図 9），図 10のキーワードが得られた．

「会津」「薩摩」「長州」の間に存在する，部分興味に関連する興味キーワードには,「藩」という共通概念を表す単語および「幕末」「軍」という 3 つの藩に関連するキーワードが含まれている．また,「慶喜」「孝明」「容保」の 3 人に共通する概念は「幕末の人名」であるが，それに近い「幕末」という単語や関連する「政治」「徳川」

[13]禁門の変とは，江戸時代末期に長州藩が京都の幕府軍（薩摩藩＋会津藩）を攻撃して敗れた事件である．このとき，徳川慶喜は孝明天皇を守るために幕府軍の総督として戦った．また，会津藩主松平容保は京都守護職として，幕府軍を指揮した．

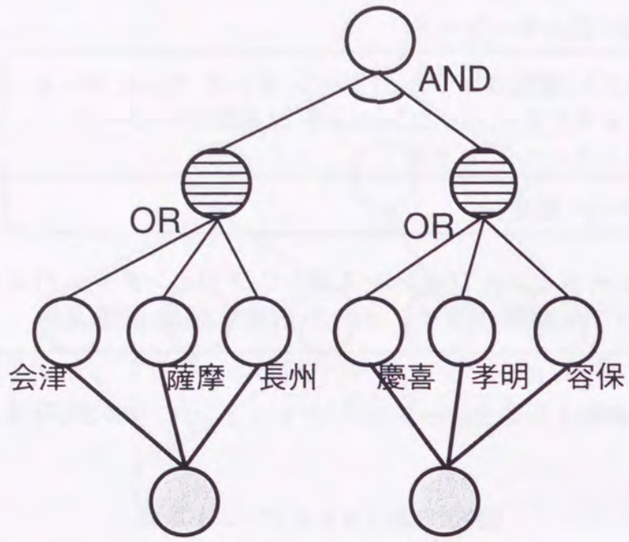


図 9: 実験「日本史」結果

「幕府」という単語，および実験当時 NHK の大河ドラマとして放映されていた「徳川慶喜」に出現していた人という共通点により，検索時における Web データベース内の相関を捉えた「ドラマ」という単語が含まれている．

しかし，ユーザの真の興味の対象が「禁門の変」にあることを，システムは知ることができないため，システムが自動的に検索語を補って検索することはできない．それゆえ，ユーザ自身が提供されたキーワードを用いて検索式を更新する．

元の検索式の入力によって，サーチエンジンは 210 件のページを出力した．ユーザは，目的の「禁門の変」について書かれたページが，出力の先頭の 10 件のページに存在しないかと思い，実際にその 10 のページを見たが，「禁門の変」という語を含むページは存在しなかった．ユーザはそれら 10 ページの中に，「禁門の変」の話を含んでいない「慶喜」の話についてのページが多いことを知り，検索式から「慶喜」を削除して，



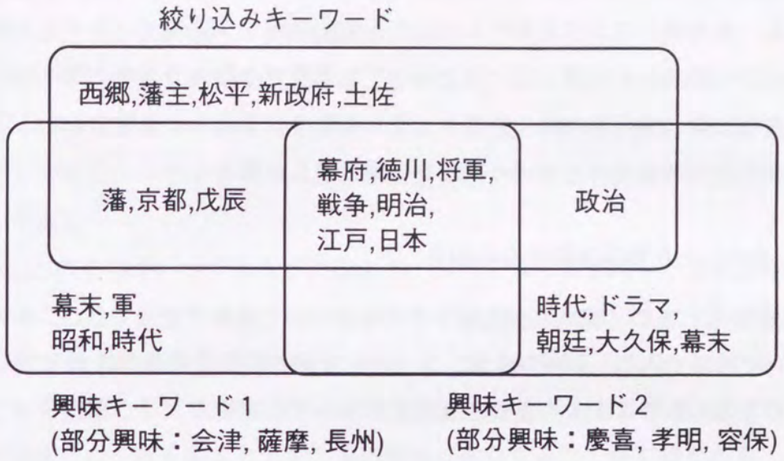


図 10: 実験「日本史」結果

「(会津 OR 薩摩 OR 長州) AND (孝明 OR 容保)」

と更新した．

また，提供されたキーワードの中から，「禁門の変」に関わるキーワードとして「京都」という「禁門の変」が起こった地名を絞り込みキーワードの中から選び，

「(会津 OR 薩摩 OR 長州) AND (孝明 OR 容保) **AND 京都**」

と検索式を更新して再検索を行なった．すると検索結果が 75 件に絞り込まれ，先頭から 5 番目に「禁門の変」というタイトルのページを発見できた[14]．

このように，システムが提供するキーワードを有効に用いることによって，確実に目的の情報に近付くことができる．また，検索時点における Web ページ内の単語間の相関をもとに，検索時点における多くの大衆の興味に基づいた検索が可能となる．この大衆とは，直接には Web ページを作成している人々を指すが，情報を公開する人に対して情報を閲覧する人が存在することも事実であり，ある話題に関する

[14] 1998/2/6 4:45 観測

多くのWebページの存在は，それらのページを必要とする多くのユーザの存在を表している．したがって，本システムにおける大衆は広く一般のインターネットユーザさらには一般の人々を指しているといっても過言ではない．この，我々が広く共有する興味に関して情報の漏れを防ぎ，また情報の絞り込みを支援するキーワードを提供することは多くのユーザの検索を支援できると考える．

### 3.4.3 キーワード抽出実験からの知見

新仮説生成に基づく興味表現支援システムについて本章で述べてきた．本システムを用いて検索式入力，新仮説生成，キーワード抽出を行なう実験を繰り返した結果，次の2点の本システムの性質と問題点が浮かび上がった．

1. OR条件で結び付いている検索語間に新仮説が生成されやすい
2. ユーザが関連語提供を望む検索語間に必ずしもWebページ内で相関がない

まず1.の点について述べると，AND条件で結びつけられる検索語は，互いに独立な単語を組合わせることが多いのに対して，OR条件で結びつけられる検索語には,「いずれか一つを含む」というOR条件の性質に基づいて換言可能な単語や，明確に一つのカテゴリーに収められる類の単語が用いられやすい．そのため，OR条件によってまとめられた検索語間にユーザの部分興味が存在すると考えられ，AND条件によってこれらの部分興味が一つにまとめられて，ユーザの興味全体が構成されると考えられる．すなわち，OR条件によってまとめられている単語は，それら全体によって一つの概念を表し，その概念が集まって一つの大きな興味ができていると捉えられる．そこで次章においては，この仮定に基づいた検索式を対象した興味表現支援システムについて述べる．

また1の項に関連して，本システムにおいては,「(京都 OR 奈良 OR 神戸) AND (観光 OR 名所) AND だんご」という検索式が与えられた場合に，OR条件内の一部の単語「京都」と「奈良」のみに新仮説（たとえば古都という概念を表す）が生成されない．また,「奈良」と「観光」に新仮説が生成されて,「鹿」や「大仏」といった奈良観光に特有のキーワードを得ることは難しい．しかしこれらは問題点ではなく，このような新仮説はユーザの検索時点における検索意図とは異なり，ユーザの検索を却って妨げるものと考えられる．なぜなら，ユーザの検索条件は「神戸」をも含む「関西」であり「古都」ではないこと，また，一般的な関西での観光を思い描いてる段階のユーザにとって，具体的な奈良観光の情報を提供するのは時期尚早だからである．

次に2.の点についてであるが，Webページ内でユーザが用いた検索語の間に相関がないことの原因としては，ユーザが用いた検索語に実際にまとまりがないことや，WWW上に存在する関連情報が少ないことなどが考えられる．この場合には，逆にこれを指標として相関が現れる検索語の組合せを入力し，積極的にWeb上の大衆の興味との一致を図ることで，情報獲得に役立てることが可能である．これとは別に，ユーザがもつ興味のうちには，大衆の興味に一致しなくともユーザ独自の観点に基づく興味が存在することがあり，そのユーザ独自の興味に関する関連語が望まれる場合がある．この場合に本章で述べたシステムは何ら検索の支援を行なうことができない．そこで，次章で述べる興味表現支援システムにおいては，上記の仮定に基づく検索式の構造をもとに，全てのユーザの部分興味に関連する検索キーワードを提供する．すなわち，検索式が一つのユーザの興味全体を表すと考え，その興味を構成する部分興味ごとに検索式を分割することで，全ての部分興味に関連する検索支援キーワードをユーザに提供するシステムを提案する．

# 4　検索式分割による興味表現支援システム

前章においては，ユーザが与えた検索語の中で，大衆の興味や理解に一致する検索語の関係に焦点を当てて，ユーザの興味に関連する検索キーワードを提供した．しかし実際には，ユーザが関連語を得たいと思う検索語を特定することは難しい．そこで，ユーザが自身の興味をもとに検索式を作成する過程に基づいて検索式の型を決める．その検索式をユーザの興味を構成する全ての要素に分割して各要素の関連語をユーザに提供する．本章におけるシステムの構成は図 11のようになる．これは「検索式分割」の処理が図 2右上の新仮説生成の工程に取って替わったものである．本章ではこの検索式分割による興味表現支援システムについて述べる．

## 4.1　ユーザの興味の構造に関する仮定

第 3章においては，任意の Boolean による検索式を入力としていたが，本章では検索式を作成するユーザの興味の構成に関する仮定に基づいて，検索式の形式を定める．

**ユーザの興味に関する仮定：**
ユーザは一つのことを頭に思い描きながら検索する

この仮定にしたがうと，検索式は OR 条件で結合された検索語の組合せを AND 条件で結合した形式となる．なぜなら，ユーザは全体として 1 つの興味を表すために，興味全体を構成する要素を AND 条件で結びつける．(このユーザの検索目的である 1 つの興味をユーザの興味の全体として**全体興味**と呼ぶ．) 加えて，全体興味を構成する各要素は同義語や類義語などによる言い換えが可能であったり，より具体的な表現を列挙することが可能であるため，それらが OR 条件で表されるからである．この仮定は，Boolean による表現全体に比べて検索式の表現を限定しているが，ユーザが検索条件として与える検索式の表現としては，そのほとんどをこの形式で表すことができる．



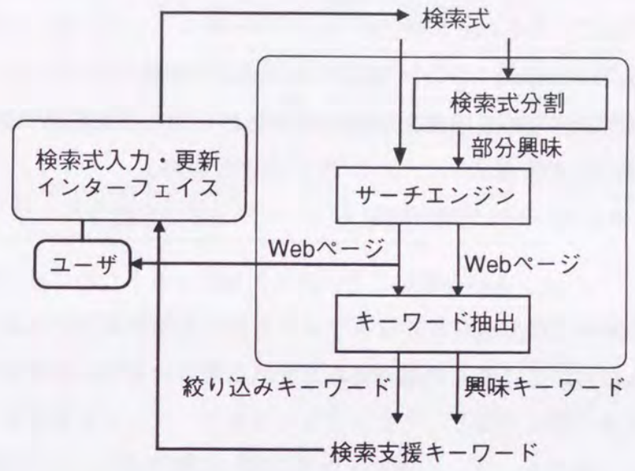


図 11: 興味表現支援システム（検索式分割）

## 4.2　ユーザの興味に基づく検索式分割

この検索式構成の過程から，OR で結合される検索語もまた 1 つの概念に基づいて与えられると考えられる．そこで，検索式全体いくつかの概念の集まりとみなし，検索式を AND の箇所で切り分けて得られる **OR 結合の検索語の組合せ**が表す 1 つの概念を，ユーザの興味の一部として**部分興味**と呼ぶことにする[15]．

たとえば，関西でだんごを食べながら観光したいユーザが「(京都 OR 奈良 OR 神戸) AND (観光 OR 名所) AND だんご」と検索式を入力したとする．すると，この検索式を AND の部分で切り分けることで，(京都 OR 奈良 OR 神戸)，(観光 OR 名所)，だんごの 3 つの部分興味に分けられ，このそれぞれは暗に関西，観光，だんごという概念をユーザ自身の言葉によって表されている．そこで，各部分興味に関連するキーワードを得るために，3.3.2節で述べた方法と同様にユーザの部分興味を表す検索式を作成する．すなわち，(京都 OR 奈良 OR 神戸) から「京都 AND

[15]OR 条件がなく検索語が 1 つの時は，検索語 1 つで部分興味を表すことになる．

奈良 AND 神戸」を,(観光 OR 名所) からは「観光 AND 名所」を，だんごはそのままの「だんご」をそれぞれ検索式としてサーチエンジンに与え，興味キーワードを抽出するための Web ページを獲得する．以下，獲得した Web ページから興味キーワードを得るための処理は 3.3.2項以降の処理と同じであるので割愛する．

## 4.3 興味キーワードの精度向上

興味キーワードは，各部分興味ごとに別々に抽出されるために統一性を欠いており，ユーザの興味とは関係のないキーワードを多く含んでいる．特に検索式分割によるシステムにおいては多くの部分興味に分けられるために，分けられた部分興味を再び一つにまとめて全体としてまとまりのあるキーワードを提供することが望まれる．そこで，興味キーワード全体としての統一を図るために次の操作を行なう．

[興味キーワードの統合]

1. 各部分興味に関連するキーワードを 3.3.3項の方法により 100 ずつ選ぶ．

2. 各名詞 $A$ の評価値を次式で更新する．

$$Value(A) = Value(A) * Number(A) \tag{15}$$

ただし $Number(A)$ は，1. の後に名詞 A が関連する部分興味の数である．

3. 各部分興味ごとに評価値の高い名詞を 20 個ずつ取り出す．

この評価値の更新は，多くの部分興味に関連するキーワードを優先して取り出すことに当たる．すなわち，各々の部分興味から得られるキーワードの中から他の部分興味との共通のキーワードを採ることで，ユーザの根本的な 1 つの興味に基づいたキーワード抽出を実現している．このようにして，検索式を分割することによっても部分興味に関連する興味キーワードが得られ，これらのキーワードが絞り込みキーワードとともにユーザに提供される．実際に抽出されるキーワードの具体例は



表 2: 新仮説生成型と検索式分割型の比較

| | 新仮説生成 | 検索式分割 |
|---|---|---|
| 部分興味の存在箇所 | Web ページ内で相関がある検索語の組合せ | 分割されたすべての検索語（の組合せ） |
| 出力キーワードの数 | 少ない | 多い |

次章の検索実験の中で見て頂きたい．そして本章の最後に，ここまでで述べた 2 種類の興味表現支援システムの比較検討を行なう．

## 4.4 新仮説生成と検索式分割の比較検討

ここまでで述べてきた二種類の興味表現支援システムについて本節でまとめる．新仮説生成と，検索式分割の違いは主に次の点である（表 2）．

それは，新仮説生成によるシステムおいては，興味キーワードが提供されない検索語が存在する点である．このキーワードが提供されない検索語について詳しくみると，新仮説は Web ページ中で相関のある単語間に生成されるために，他の検索語との関係が薄い検索語には新仮説が生成されにくく，興味キーワードが提供されない．新仮説生成の側に立つと，ユーザの興味は複数の検索語を入力している部分がユーザの興味がある部分であり，情報を漏らしたくない部分であると解釈できる．しかし検索式分割の立場に立つならば，ユーザが検索語を思いつけないから，検索語が 1 つしか入力できていないということになり，実際にはこのどちらの場合もあり得ると考える．このことはまた，提供するキーワードの数にも影響を与える．検索式分割においては，すべての興味を構成する部分興味に関して興味キーワードを提供するため，各部分興味に関連して提供する興味キーワードの数を一定とすると，新仮説生成に比べてキーワードの数は多くなる．また，興味キーワードの数は新仮説生成の時だけ出力数を多くするという単純な数合わせを行なっても，キーワード

の質に差が生じる．キーワードの量についての感じ方には個人差があり，新仮説生成と検索式分割のどちらによるキーワードを望むかは検索目的などによっても異なると考えられ，その時々に応じてユーザが一方を選択することになる．例えば，入力される検索語が少ない，検索の初期のうちは検索式分割によって多くの関連語を提供し，検索語の数が増えかつユーザが検索語が多過ぎると判断した場合には，新仮説生成による関連語提供を利用する方法などが考えられる．

しかしいずれを選択するにしても，キーワードを一次元に並べて提示するだけでは，どの部分興味とどのキーワードが関わりをもつかという，キーワードとユーザの興味との関わりを示すことができず，部分興味ごとにキーワードを取り出した意味がなくなってしまう．そこで次章においては，抽出された興味キーワードをユーザの興味によって分類して配置する二次元平面インターフェイスを提案する．



# 5　情報の視覚化：検索支援インターフェイス

本章では，ユーザが自身の興味を的確に表現するために検索支援キーワードを二次元平面上に効果的に配置した検索支援インターフェイスについて述べる．本論文で提案するインターフェイスは，検索されたWebページ情報全ての詳細な視覚化を目指すのではなく，一部のWebページ情報をWebページ集合全体の標本として，Webページから抽出されたキーワードを用いた視覚化を行なう．換言すると，ユーザに探される側のWebページを視覚化するのではなく，ユーザが積極的に探すための手がかりとしてのキーワードを視覚化する．

## 5.1　検索支援インターフェイスの意義

本節では，検索支援インターフェイスの意義について述べる．本論文で提案する二次元平面を用いた検索支援インターフェイスは，次の3点を実現する上で必要不可欠である．

1. 検索支援キーワードの提示
2. 検索式によるユーザの興味の表現
3. 検索式更新の操作性の向上

まず，1.の検索支援キーワードの提示方法に関して，検索に有効なキーワードを図12下部のように全て一列に並べて提示する場合，各キーワードとユーザによって入力された各検索語とがどのように関連しているかを，ユーザがすぐに理解することができない．また，得られた理解はユーザ独自の言葉同士の関係に基づいているため，実際のWeb上でどの検索語とどのキーワードとが関わり合っているかについての情報をユーザは知ることができない．しかし，そのような検索語とキーワードとの間の関連情報は，ユーザが提供されたキーワードを有効に活用して，存在する目的の情報にたどり着くためには必要な情報である．さらに，図12の各キーワード

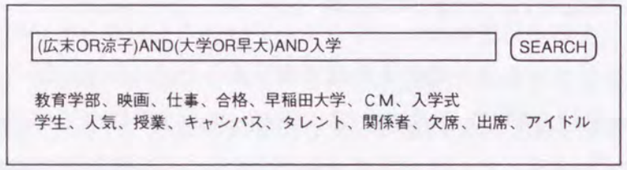


図 12: リスト型インターフェイス

をユーザが逐次吟味した上でどのキーワードを用いるかを決定する作業は骨が折れるため，提供できるキーワードの数にも制限がかかることになる．

本システムにおいては，このようなキーワード提示に関する要求を満たすべく，二次元平面上にキーワードを分類して配置する．特に本システムでは，検索支援キーワードとして絞り込みキーワードと興味キーワードをともに出力すること，および興味キーワードがユーザの部分興味ごとに与えられていることに着目して，ユーザの興味に基づくキーワードの分類を実現する．

ユーザが入力した検索語と存在する Web 情報との関係を提供するキーワードを通して理解していくことは，2. のユーザが自らの興味を表現する上でも重要となってくる．今後情報量の拡大に伴って，自身の興味をより具体的かつ正確に表す必要がますます高まると考えられることからも，多くの検索語を的確に組み合わせた検索条件作成のためには，検索語との関連に応じたキーワードの配置が不可欠となる．たとえば，提供するキーワードの配置から，存在する Web ページ情報の性質や偏りを知り，キーワードが不足する部分や具体的表現が望まれる部分を知ることは，適切な検索を行なう上で有効に働く．したがって，この点においてもキーワードの二次元配置が望まれる．

最後に，3 の検索の操作性の向上についてであるが，従来の AND,OR,NOT 条件による検索を多くのキーワードを用いて容易に行なうためには，専用のインターフェイスが必要である．特に日本語での検索を考えた場合，半角である AND,OR,NOT や ()（括弧）などの記号とともに全角である日本語を交えて検索を行なうことは大



変骨が折れる．また，できるだけ単純な作業によって検索式の更新を行ないたいと考えるであろう．そのために，キーボードを用いずにマウス操作のみによって再検索が行なえる検索支援インターフェイスを提案する．以下の節では，この検索支援インターフェイスの詳細について述べていく．ただし本章では，検索式分割による興味表現支援システムを用いてすべての部分興味に対して興味キーワードが抽出されている場合について話を進める．これは，検索式分割の場合の方が出力すべきキーワードの数が多いためであり，新仮説生成の場合は本章で述べる方法をもとにした配置が可能となるからである．

## 5.2 検索支援キーワードの二次元平面への配置

興味表現支援システムの出力は，出力される検索結果をそのまま絞り込むための**絞り込みキーワード**と，ユーザの興味の一部となっている各部分興味ごとに得られる**興味キーワード**の 2 種類のキーワードである．これらのキーワードを二次元平面上に整理して配置する方法について述べる．ただしここで，インターフェイス上に配置するために与えられているのは，絞り込みキーワードの集合 $S$と部分興味の数 $n$ に応じた（部分）興味キーワードの集合 $I_1, I_2, ..., I_n$，それからユーザが入力した検索語を含む各部分興味の検索枠である．ここでまず，絞り込みキーワードの配置方法について述べる．

### 5.2.1 絞り込みキーワードの配置

絞り込みキーワードを配置するアルゴリズムは次の手順で表される．

[絞り込みキーワードの配置]

STEP1. いずれの部分興味から得られた興味キーワード集合にも含まれない絞り込みキーワードを集合$S'$としてまとめる[16].

$$S' = S - \bigcup_{i=1}^{n} I_i \tag{16}$$

STEP2. $S'$に含まれるキーワードはユーザの興味全体に関わるキーワードであるため，すべての興味キーワード集合に属する興味キーワードとともに配置する（配置の具体的方法は後述）.

STEP3. $S - S'$に含まれるキーワードは，いずれかの部分興味に関連する興味キーワード集合にも含まれるため，絞り込みに有効であることを示すタグとともに興味キーワードの配置にしたがって並べる.

次に，興味キーワードの配置について述べる．興味キーワードには複数の部分興味に関連するキーワードがあり，各部分興味に関連するか否かで2通りずつあるため，本来興味キーワードは$n$個の部分興味に対して$2^n$種類に分類される．しかし，これらを各部分興味を基準として表示するには$n$次元の空間を要し，限られた二次元平面内に区別して配置することが不可能であるため，次項の方法によって配置する.

### 5.2.2 興味キーワードの配置（木構造）

本項では，興味キーワードに木構造の階層を持たせて二次元平面に配置する方法について述べる.

[16]集合差$A - B$は，$A \cap \overline{B}$を表す.



[興味キーワードの配置]

STEP1. 各部分興味間の関連度$R_{ij}$を計算する

2つの部分興味間の関連度を2つの部分興味に共通に関連する興味キーワードの数として計算する[17].

$$R_{ij} = |I_i \cap I_j| \tag{17}$$

STEP2. $n$個の部分興味の検索枠を一列に並べる

隣合う部分興味間の関連度の和$M$が最大になるように部分興味を一列に並べ，その並びに応じて検索枠を配置する.

$$M = \sum_{i,j \in \{1,\ldots,n\}} R_{ij} \tag{18}$$

STEP3. 興味キーワードを分類

部分興味を一列に並べた時に，隣合わない2つの部分興味に関連する興味キーワードを，その2つの部分興味の間に並べられた部分興味にも関連がある興味キーワードとして扱う[18].

STEP4. 興味キーワードを配置

各興味キーワードを属する全ての部分興味の検索枠から等距離の位置に配置する.

たとえば，3つの部分興味$A, B, C$があった場合，$A$と$B$，$B$と$C$，$C$と$A$の3通りの組合せ全てについて，共通に関連する興味キーワードの数を関連度とする（順

[17]$|A|$は集合$A$の要素数を表す.

[18]この操作によって，$n$個の部分興味に対して，各部分興味にのみ関連する興味キーワードが$n$種類，複数の部分興味に関連するキーワードが${}_nC_2$種類に分類されるため，全部で$n(n+1)/2$種類となりこれが$n$の2乗オーダであるので，二次元平面上に配置可能となる．図19では3つの部分興味に対して，興味キーワードが$3*(3+1)/2 = 6$種類に分類されている.

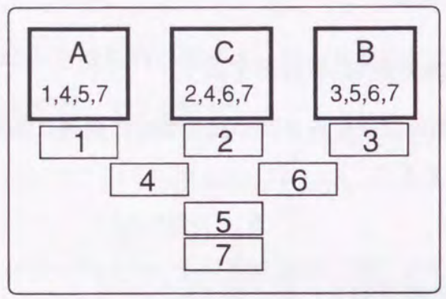


図 13: 興味キーワードの配置例

に 3,4,5 とする）(STEP1)．次に 3 つの部分興味を一列に並べる 3!通りの方法のうちで，$B-C-A(A-C-B)$ の並びが最も高い関連度 $9(=4+5)$ となる (STEP2)．最後に，図 13のように部分興味 $A, C, B$に関連する興味キーワードがそれぞれ {1,4,5,7}，{2,4,6,7}，{3,5,6,7} であった場合，キーワード 1 は部分興味 $A$ のみに関連するため部分興味 $A$ の検索枠の近くに，キーワード 4 は部分興味 $A$ と $C$に関連するため 2 つの部分興味 $A$ と $C$の検索枠から等距離の位置に，キーワード 7 は全ての部分興味に関連するため，興味全体を表すキーワードとして最もユーザの根本的な興味を表し，かつすべての部分興味の検索枠から等距離の位置に配置する (STEP4)．合わせてここに，絞り込み専用となったキーワード集合 $S'$のキーワードも配置する．また，キーワード 5 は離れた部分興味 $A$ と $B$に関連しているが，ユーザの興味は全体を通しては一つであることと，隣接する部分興味間の関連度が最も高い並びであるから，キーワード 5 は $A$ と $B$をつなぐ役割を果たしている $C$とも強い関わりがあると考え，$A, B, C$すべてに関連するキーワードと同様に配置する (STEP3)．

部分興味が 3 つの場合は図 15のような検索支援インターフェイスが表示される．（インターフェイス画面の詳細は 5.4節で後述）このキーワードの配置によって，各部分興味のみに関連するキーワードと，複数の部分興味を繋ぐキーワード，さらに



はユーザの興味の全体を表すキーワードを明らかにすることができる．

### 5.2.3 興味キーワードの配置（円型）

前項では，検索枠を横一列に並べる配置について述べた．しかし検索枠の数が増えるにつれ，離れた複数の枠に関係するキーワードの数が増して，キーワードの配置が正確でなくなるという問題点もある．そこで本項では，初めに検索枠をその枠間の関連度に基づいてグループ分けを行ない，検索語間の関連を知る上で重要な部分を正確に表した円型インターフェイスについて説明する．

そこでまず，各検索枠内の検索語で表される部分興味をクラスタリングするアルゴリズムを示す．

[部分興味のクラスタリング]

STEP1. 各部分興味間の関連度 $R_{ij}$を計算する．

2 つの部分興味間の関連度を 2 つの部分興味に共通に関連する興味キーワードの数として計算する[19]．

$$R_{ij} = |I_i \cap I_j| \tag{19}$$

STEP2. 次式の関連度の和 $S(A)$ が低い部分興味 3 つを選び，3 つのグループを作成する．

$$S_i = \sum_{j \in I} R_{i,j} \tag{20}$$

STEP3. まだ選ばれていない部分興味を，最も関連度の高いすでに選ばれた部分興味と同じグループに入れる[20]．

[19] $|A|$ は集合 $A$ の要素数を表す．

[20] ただしグループの部分興味の数が 4 以上になる時には，関連度の順位が 4 位以下の部分興味を次に関連度の高い部分興味を含む別のグループに入れる．

この方法によって，部分興味は3つのグループに部分興味間の関連に基づいてクラスタリングされ，部分興味が3つの場合は図16のような検索支援インターフェイスが表示される．画面のKey1からKey3の3つの検索枠がユーザの部分興味を表し，各検索枠の中には部分興味を構成する検索語が入れられる．各枠の近くには各部分興味にのみ関連するキーワード，2つの検索枠の間には2つの部分興味に関連するキーワードが，またインターフェイス中央には全ての部分興味に関連するキーワードが配置される．先の絞り込み専用のキーワードも，検索式全体から得られるキーワードであるので中心に配置する．

また部分興味の数が4以上の場合においては，クラスタリングされて複数の部分興味を含むグループを，図16の円を縮小した相似形に上述の方法と同様に配置して，図16の1つの検索枠と置換した図17のようなインターフェイスが構築される．すなわち，3つの検索枠（のグループ）を一つの単位として，その3つの検索枠の関係については情報を漏らさずにキーワードを分類して表すことができる．

## 5.3 検索式の更新と情報獲得の関係

ユーザの望む情報が一回の検索で得られれば検索は終了となるが，多くの場合，何度かの検索式の更新を経て目的の情報を見つけることになる．本節では，ユーザがインターフェイス上に提示されたキーワードを用いて，ユーザがどのように検索式を更新することができるか，また更新するべきかについての具体的な方法について述べる．

### 5.3.1 検索における適合率と再現率

検索の精度を測る基準には適合率（Precision）と再現率（Recall）とがある．適合率は，検索されたWebページ中でユーザの希望に合うWebページの割合，再現率は，Webページのデータベース中でユーザの希望を満たすWebページ全てのうち，検索によって抽出されたWebページの割合である．これを式で表すと図14の



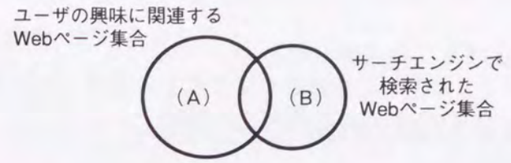


図 14: 適合率と再現率

記号A，Bを用いて，式（21），(22）のように表される．

$$Precision = \frac{|A \cap B|}{|B|} \tag{21}$$

$$Recall = \frac{|A \cap B|}{|A|} \tag{22}$$

検索においては，これら適合率と再現率の両方の値を向上させることを目指す必要があるが，検索条件を変えることによって同時に両者を向上させることは難しい[Salton 97]．それゆえユーザは，自身の興味を正確かつ的確に表すために検索対象となる文書集合に合う検索語を用いて，図14の集合Aをちょうど表す検索式を作成する必要がある．そのための具体的検索式更新の方法を次項で述べる．

### 5.3.2 検索式更新の具体的方法

情報の絞り込みや情報の範囲を広げながら適合率と再現率を向上させるのための，検索式更新の方法には以下のものが挙げられる．

**[検索式更新の方法]**

- 情報の絞り込みを行なう（適合率の向上）

  1. [絞り込みキーワード] をAND条件として追加
  2. 検索式中のOR条件を削除

3. [絞り込みキーワード] を NOT 条件[21] として追加

- 情報の範囲を広げる（再現率の向上）

4. [興味キーワード] を OR 条件として追加
5. 検索式中の AND 条件を削除

- 欲しい情報だけを獲得する（適合率と再現率両方の向上）

6. 検索式中の 1 つの検索語を 1 つの [絞り込み/興味キーワード] と交換
7. 検索式中の 1 つの検索語を複数の [絞り込み/興味キーワード] と交換
8. 検索式中の複数の検索語を 1 つの [絞り込み/興味キーワード] と交換

これら更新の方法について以下で少し説明を加える．絞り込みキーワードは，ユーザが与えた検索式にマッチする Web ページ集合の一部に含まれるキーワードとして抽出されているため，AND 条件や NOT 条件として追加することで，情報の絞り込みが行なえる（更新 1,3）．また興味キーワードは，もとの検索式にマッチする Web ページ集合とは別の，Web ページ集合から選ばれたキーワードであるため，OR 条件として追加することで獲得する Web 情報の範囲を広げることができる（更新 4）．また，これらのキーワードを追加する作業とは逆に余分なキーワードを削除することによっても検索式を更新できる．過度の絞り込みを行なっていた AND 条件の検索語を削除することでより多くの情報を得ること（更新 5）や不要な情報となっている OR 条件を削除すること（更新 2）でも，情報を絞り込むことが可能である．ここまでの，更新 1 から更新 5 は適合率か再現率のいずれか片方を向上させる操作であったが，削除と追加を組み合わせた検索語の取り替えによって，適合率と再現率の両方の向上を目指して，存在する Web 情報に合わせた的確な検索式の作成を目指すことができる．すなわち，検索式中の検索語をより適切な単語で置き換えること（更新 6）や，よく知らなかった興味の対象について，具体的な複数の検索語で表すこと（更新 7），そして，曖昧な表現で羅列していた検索語群を適切な 1 つの単語で表現すること（更新 8）が可能である．

[21] 指定した検索語を含まないページを集める検索条件



ここで挙げたように，検索式をさまざまな検索語で更新していく可能性が存在する．しかしいずれの方法で，どの提供された検索語を用いて検索式を更新するかを計算機が単独で判断することは極めて困難である．それゆえ，ユーザ自身が表現し尽くせていない興味の内容によってこれを決める．なぜなら，ユーザの多様な興味を検索語の組合せという限られた情報のみから，計算機が正確に判断することは不可能だからである．そこで，検索されたページを実際にユーザが閲覧し，また提供された検索語を眺めながら検索条件を吟味することで，ユーザ自身が自分の興味表現となる検索式を完成させる．

また本システムにおいては，二度三度と何度でもキーワードの提供と検索式の更新を繰り返すことができる．この繰り返しによって，ユーザが心に抱いている興味を，存在する Web 情報に合った表現を用いて，徐々に具体的な興味表現として検索式で表すことができる．それはこの繰り返しが，単なる同じ作業の繰り返しではなく，興味を具体化する一連の流れの一部として存在しかつ，Web 情報とのインタラクションをもとにして，存在する情報の位置を確実に突き止めるための 1 ステップとして存在しているからである．

## 5.4 検索支援インターフェイスの外観

本節では，実際に作成した検索支援インターフェイスの構成とその操作法について述べる．

図 15 が 5.2.2 項で述べた木構造の検索支援インターフェイス，図 16 と図 17 が 5.2.3 項で述べた円型の検索支援インターフェイスである．図 15 は「サッカー AND 五輪 AND 予選」という検索式を入力したときの，検索支援インターフェイスである．この図において画面上部の Key1 から Key3 の 3 つの各検索枠の中に，それぞれ部分興味を構成する検索語が入れられている．枠の下には 5.2 節で述べた方法にしたがっ

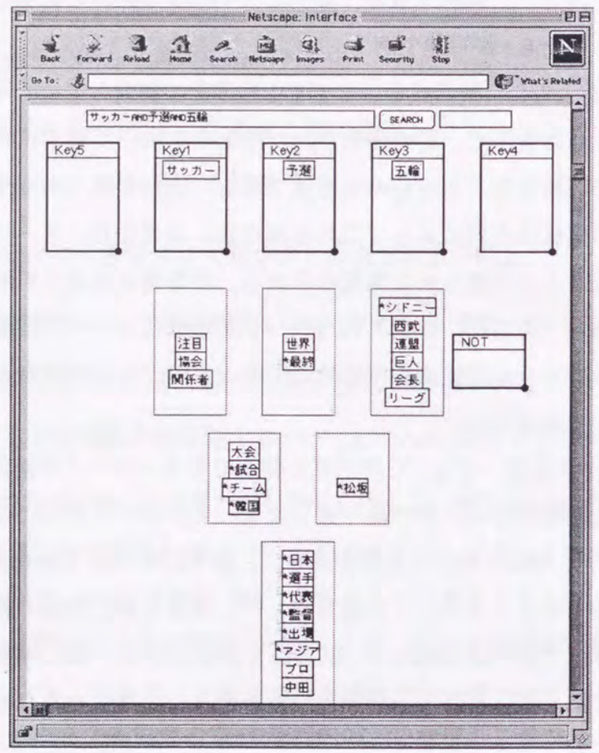


図 15: 検索支援インターフェイス（木構造）

て興味キーワードが配置されており，“*”印のついているキーワードが絞り込みキーワード[22]となっている．

ユーザはこのインターフェイス上で，キーワードをマウスでドラッグして，各部分興味を表す検索枠の内外に移動させることにより検索式を更新できる．すなわち1つの検索枠の中に含められたキーワードはOR結合で結ばれた検索条件となり，そのOR結合で表されている検索枠同士がAND結合で結ばれ，図15上部のような検

[22]実際のインターフェイス上では色分けして表示している



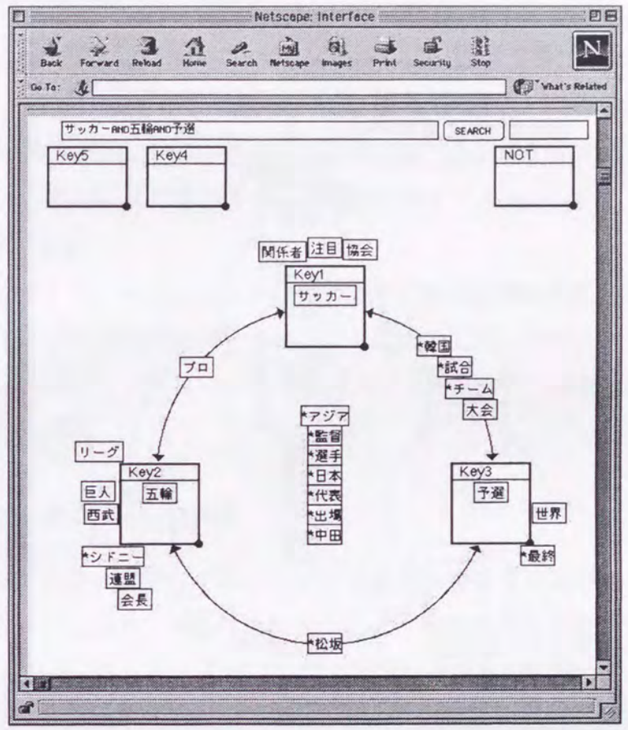


図 16: 検索支援インターフェイス（円型）

索式となる．また，画面右にはNOT条件のための検索枠も用意されている．この枠に含められた検索語はNOT条件として扱われる．ユーザは検索式を更新した後に，画面右上のSEARCHボタンを押すことで再検索を実行できる．

図16は，同じ検索式を与えた時の円型の検索支援インターフェイスである．検索語「サッカー」と「五輪」にのみ関連する「プロ」というキーワードが，木構造のインターフェイスでは図15の最下部に配置されているのに対して，円型の図16では正確に表現されている．また，このインターフェイス上で「日本」と「代表」というキーワードをそれぞれKey4とKey5の検索枠に追加して検索を行なった後に

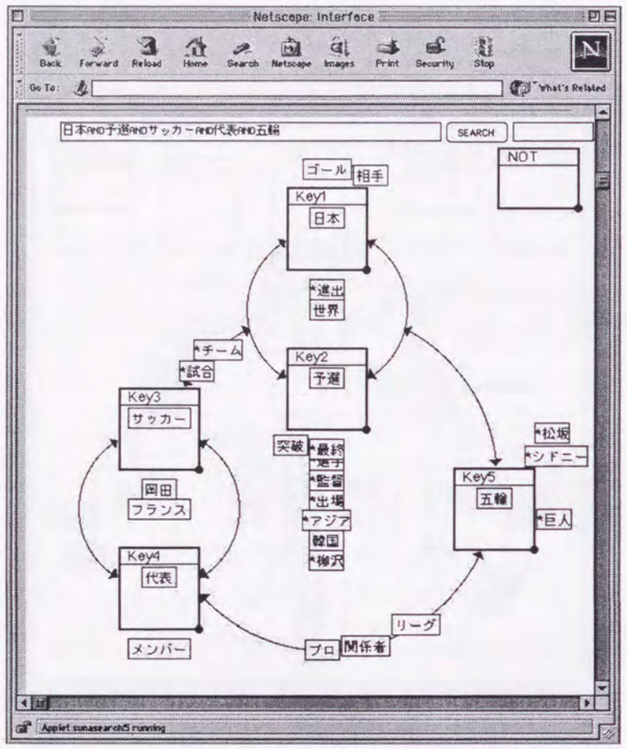


図 17: 検索支援インターフェイス（円型）

得られる検索支援インターフェイスが図17である．5つの部分興味が3つにクラスタリングされて円型に配置され，その1つのクラスタにおいても同様に円型の配置に基づいてキーワードが提供されている．



## 5.5 検索式更新実験

本節では，興味表現支援システムによって実際に行なった検索の結果を示す．実験環境は，Pentium II 350MHz (256MB) OS-Linux であり，プログラムは検索サーバが Perl，検索実行部が C 言語，検索インターフェイスが JAVA でインプリメントされている．データベースは芸能スポーツ関係のニュース [zakzak] など約 10000 件の Web ページである．

あるユーザがアイドルの広末涼子の大学生活に関する情報を得ようとして,「広末 AND 大学」という検索式をサーチエンジンに入力した．本システム内のサーチエンジン[23] はこの入力に対して 42 件の Web ページを出力し，検索式更新のために図 18の画面を出力した．

### 5.5.1 検索支援キーワードの提示

「広末」と「大学」がそれぞれユーザの部分興味を表す検索語として Key1 と Key2 の検索枠の中にあり，それらの枠の下には，各部分興味に関連する興味キーワードが配置されている．たとえば,「広末」の下には彼女の大学である「早大」や「CM」といった関連語が,「大学」の下には「卒業」「合格」などの関連語が現れている．

また,「広末」と「大学」両者の関連語として広末の名前である「涼子」や大学関連の広末の大学生活の情報を得るために有効なキーワードが現れている．これらのキーワードは実際の Web ページから抽出されているため，それらのキーワードを含む複数のページの存在が保証されている．

### 5.5.2 検索支援キーワードと Web データベースとの関係

これらのキーワードはデータベースや検索結果の特徴を表している．すなわち，絞り込みキーワードは検索結果の上位ページから得られたキーワードであるため，

[23] 検索語の Web ページ内の出現頻度によってランクづけしている．

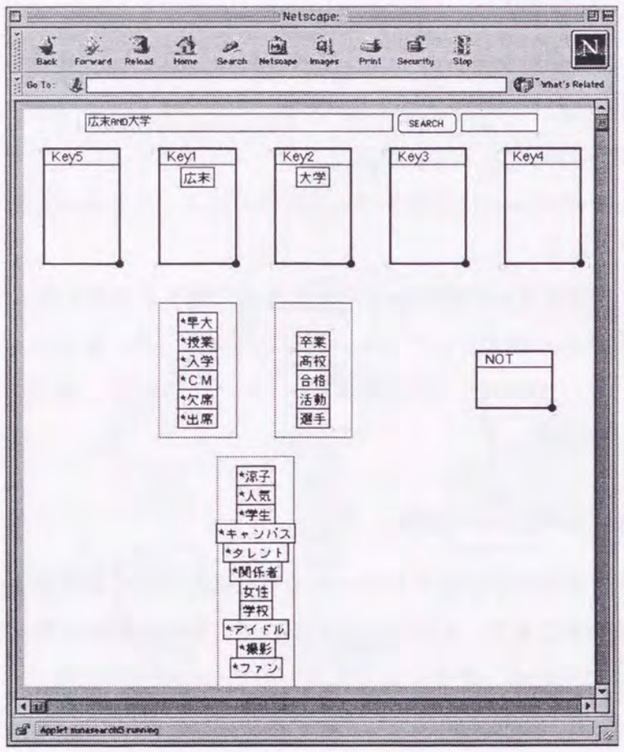


図 18: 検索支援インターフェイス

実際の Web ページを見る代わりにキーワードを見ることで，検索結果として得られたページの特徴を知ることができる．図 18においては，絞り込みキーワードは「広末」の下にはあるが,「大学」にのみ関連するキーワードの中には含まれていない．このことから検索結果は大学の情報よりも広末の情報を主としたページが多く得られていることがわかる．これは,「涼子」というキーワードが「広末」だけでなく「大学」にも関連するキーワードとして得られていることからも知ることができる．このようにユーザが，出力されているキーワードの偏りを見ることによっても，存在



する情報の性質を知ることができる．

### 5.5.3 ユーザの興味に基づく検索式の更新

ユーザは提示されているキーワードを用いて，自らの興味を明確にし，具体的に検索式として表現することができる．たとえば,「涼子」を AND 条件として追加 (5.3.2項の更新の方法 1) したり,「大学」を「早大」と取り換え（更新 6）て具体的な検索式を作成できる．また，大学生活に加えて高校の情報も得たい場合には「大学」の枠に「高校」を加えて OR 条件として（更新 4）追加することで,「大学生活」という部分興味を「学生生活」と変更することもできる．

図 18のユーザは，提供されたキーワード「涼子」「早大」「入学」をインターフェイス上でマウスで移動させて，検索式を「(広末 OR 涼子) AND (大学 OR 早大) AND 入学」と更新した．すると，32 件の Web ページが検索結果として得られた．検索式更新後のインターフェイス画面は図 19のようになった．

このインターフェイス上においても，ユーザが入力した検索語に関わるキーワードが提供されている．ユーザはこれらの提供されたキーワードの中から自身の興味をより的確に表すキーワードを発見することで，徐々に曖昧な興味を具体化していく．そこでユーザは,「入学」「大学」「早大」という単語の代わりに,「入学式」というキーワードを用いて,「(広末 OR 涼子) AND 入学式」と検索式を更新した（更新 7）．その結果 14 件の Web ページが得られ，具体的な「入学式」という興味を満たす情報を得るに至った．

### 5.5.4 ユーザの興味に関連する新たな情報の取得

ユーザはすでに目的の Web ページを獲得したが，ユーザは関連する別の情報を得ることも可能である．そこでユーザは，広末に関係するが大学には関係しない情報を得ようと考え，NOT 条件として大学関連の語「早大」「大学」「教育学部」「入学式」「キャンパス」を与えた．新たな検索式は「広末 OR 涼子」と上記 5 つの語

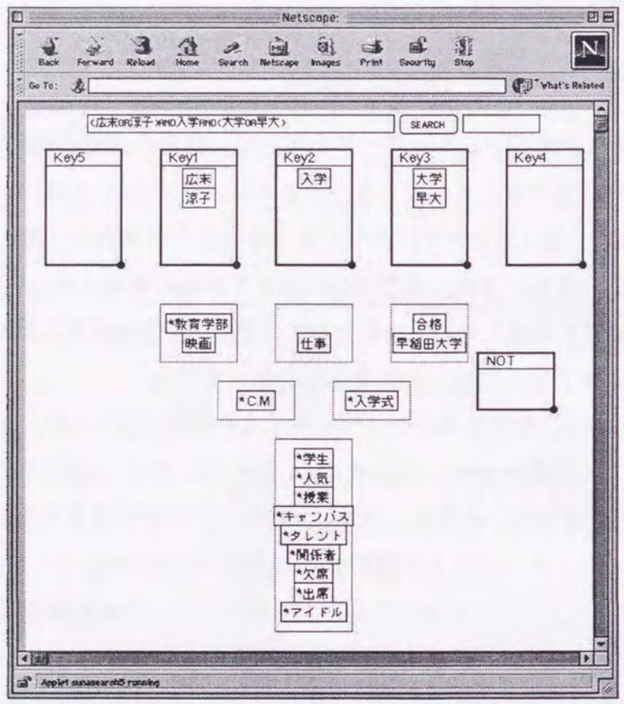


図 19: 検索式更新後のインターフェイス

による NOT 条件である．

この検索の結果 80 件のページが出力された[24]．また，新たなキーワードとして，「映画」「デビュー」「人気」「ＣＭ」「ＴＢＳ」「発売」「アイドル」といった芸能活動に関するキーワードが多く得られ，ユーザはこれらのキーワードを用いてさまざまな情報を得ることが可能となった．

NOT 条件は本システムにおいて特に効果的である．絞り込みキーワードは，現在検索されている多くのページに含まれるキーワードであるから，そのようなキー

[24]NOT 条件を用いない場合は 144 件の Web ページにマッチした．






図 20: NOT 条件による検索式更新後のインターフェイス

ワードを NOT 条件として追加することで，隠れている少数派のページを検索することができるからである．また，インターフェイスの特徴を生かして，多くのキーワードを同時に入力として扱える利点もある．ユーザはこの後「(広末 OR 涼子) AND 映画」と検索式を更新したのち，さらに「映画」に関連する「主演」というキーワードを得て「(広末 OR 涼子) AND 映画 AND 主演」という検索式を作成し，広末が主演する映画の情報獲得に成功した．

次章においては，この作成した興味表現支援システムに対して客観的な評価を行ない，本システムの有効性を確認する．

# 6　興味表現支援システムの評価

本章においては，構築した興味表現支援システムをさまざまな角度から考察した上で定性，定量的な評価を与える．

検索システムを評価する最も一般的な指標に，適合率[25]と再現率[26]がある．しかし本システムは，ユーザにキーワードを提供することでユーザの検索を支援するシステムであり，キーワードを選ぶという大事な操作はユーザ自身に任されている．そのため，これら適合率と再現率を大きく左右する単語の選択は，ユーザ自身の技量（有効な検索語を提供されたキーワードの中から選び出す）によるところが大きく，適切なキーワードを選べばこれらの値が改善されるのは当然である．実際3.4.2節の実験例において,「禁門の変」という語を含む文書を適切な文書として適合率を計算した結果，初めの検索結果時に出力された210のWebページ中に21の「禁門の変」を含むページがあり，適合率が21/210(= 0.1) であったのに対して，検索式更新後に11/75(= 0.147) に改善されている．したがって本システムを再検索支援という観点に基づいて，本システムが行なうキーワード提供と構築したインターフェイスがどの程度検索に役立てられ，ユーザの検索がスムーズに行なわれているかを評価する．

## 6.1　アンケート調査によるシステム評価

まず，実際にユーザに本システムを使って検索を行なってもらい，使用時の操作や印象に関してアンケート調査を行なった．アンケートに答えてもらったのは20代の大学生および大学院生の男女計34名である．

### 6.1.1　検索支援キーワードに関するアンケート結果

本システムが提供する検索支援キーワードについて，表3に示すアンケート結果を交えて評価を行なう．まず,「Web上に存在する情報（情報の有無や，情報の傾向

[25]検索結果に含まれる正解となる文書（Webページ）の割合
[26]正解となる文書全体の中から検索された文書の割合



表 3: 新仮説生成型と検索式分割型の比較

| 質問 | Yesと答えた人の割合 (%) |
|---|---|
| Web上に存在する情報について知ることができた | 83 |
| 興味を表すキーワードが得られた | 72 |
| キーワードを見て新たな興味が湧いた | 69 |

と性質）について知ることができましたか？」という質問に対しては，83%の人がYesと答えている．これは，3.3.3項のキーワード抽出法でも述べたように，多くのWebページに共通して出現する単語を取り出してキーワードとしているため，出力されたWebページ全体の性質を，キーワードから容易に類推できたためと考えられる．これにより，存在する情報を確認しながら再検索を行なうことが可能であるため，後戻りや試行錯誤の少ない素早い検索の実現が見込まれる．

次に,「初めは検索語として与えられなかったが，興味を表すキーワードが得られましたか？」という質問には72%の人がYesと答えている．これは，人が即座に単語を思い起こすのが不得手であることや，興味が具体化されないままで検索を始めていることを示唆する結果である．ユーザは入力し損ねた単語をシステムから提供される単語で補完することができ，具体的な興味表現を行なうことによって望む情報を素早く獲得することができる．これら2つの質問に対する結果の解釈をもとにした，本システムを用いて検索を素早く行なえることの裏付けは6.2節で再び述べる．

3つ目に「キーワードからの示唆で新たな興味が湧きましたか？」という質問に対して69%の人からYesの回答が得られた．これはすなわち，初めに意図していた興味に対する結果が得られるだけでなく，提供されるキーワードによって，興味に関連する別の情報の存在を知ったことに相当する．ユーザは検索結果として出力されたWebページ中から，目的の情報以外でも興味のあるWebページへのリンクを

辿ることがよくあり，この関連情報の取得を支援する意味においても本システムは有効に働いている．

これらの結果は，提供するキーワードがユーザ自身の興味との関わりに応じてインターフェイス上に配置されていることと，4.3節で述べた興味キーワードの精度向上の効果によって，1つあるいは少数の部分興味にのみ関連するとして出力されていたキーワードが複数，あるいは全ての部分興味に関連するキーワードとして出力され，部分興味間にまとまりが生じたことにもよっている．実際，4.3節のアルゴリズムを適用しなかった場合に比べて，興味キーワードの総出力数が部分興味2つの時は85%に，また部分興味3つの時には75%に絞り込まれている．

### 6.1.2 検索支援インターフェイスに関するアンケート結果

次にシステムが提供するキーワードが，ユーザが入力したどの検索語と関連して得られたかを理解しやすく，また再検索の操作を行ないやすいキーワードの配置を調べるため，インターフェイスの形について比較してもらうアンケート調査を行なった．用意したインターフェイスは次の5つである．まず，5章で用いた図18の木構造のインターフェイスが1つ目であり，この図の上下を反転させたピラミッド型として図21のインターフェイスが2つ目である．これらのインターフェイスにおいては，ユーザの検索語を含む検索枠が直線上に並べられ，その検索枠の上下にキーワードが規則的に配置されている．図21のインターフェイスは，ユーザの興味に関連するキーワードが下になるほど細分化されていき，上から下への階層構造をもつという点で [塩沢 97] と同形式のインターフェイスである．

また，3つ目は5.2.3項で述べた方法をもとに，図17のように検索枠を円状に配置した円型で，中央にユーザの根本的な興味を表すキーワードが配置され，外側の周辺になるほど個別の検索語に関連するキーワードが配置される．4つ目としてクラスタリングを行なわない円型のインターフェイスとして，図22のインターフェイスを構築した．これは複数の3つ以上の部分興味に関連するキーワードを，キーワー



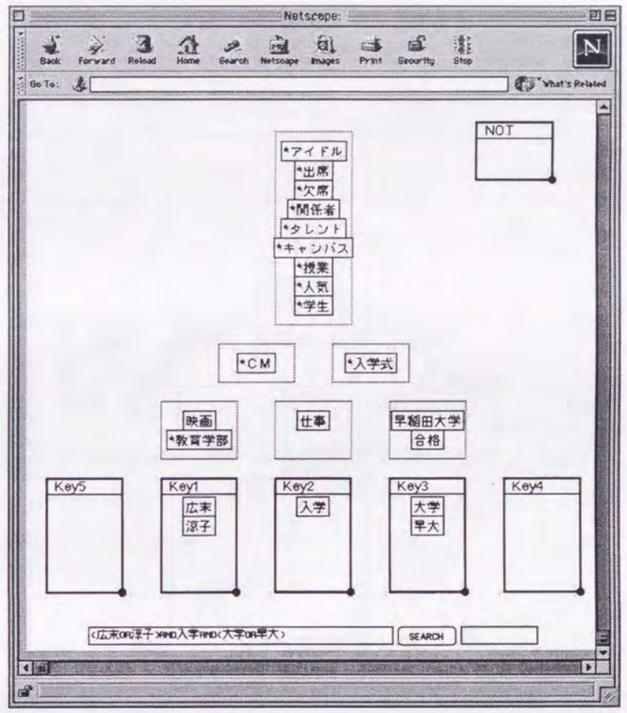


図 21: ピラミッド型インターフェイス

ドの関連する全ての部分興味の検索枠の重心位置に配置している．これらは，中心から周辺部への階層構造をもつ [Lamping 95] と同形式のインターフェイスである．

そして最後に，従来のサーチエンジンでよく行なわれている出力に似せて，全てのキーワードをまとめて提供する図12のリスト型インターフェイスを用意した後，これら5つのインターフェイスを比較するための質問をユーザに対して行なった．それぞれの質問に対して最も好まれるインターフェイスを選んでもらった結果を表4に示す．

まず「最もキーワード間の関連および検索語との関連を理解しやすい」インター

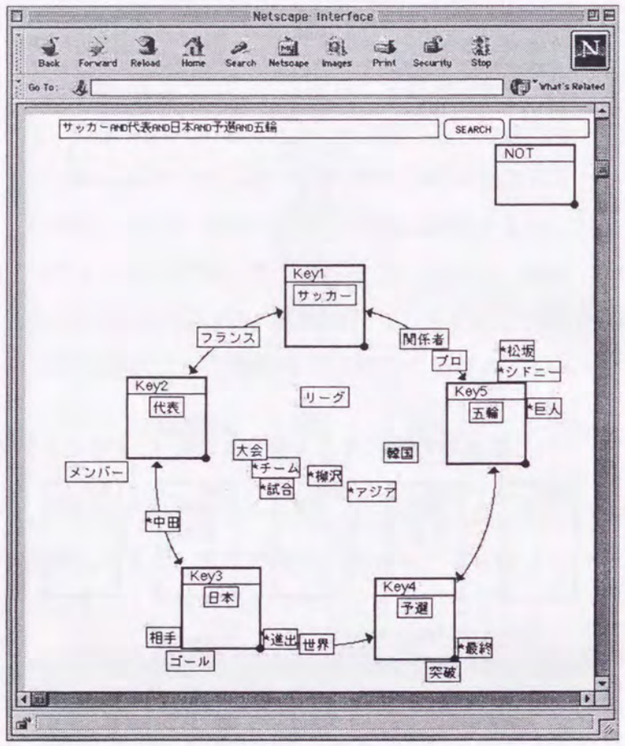


図 22: 大円型インターフェイス

フェイスはどれかを尋ねたところ，67%の人が「大円型」もしくは「円型」と答え，続いて海型の21%，ピラミッド型の9%，リスト型の3%となった．これは円型のインターフェイスにおいては，ユーザの考えの中心となっている根本的な興味に関する単語が中心に来ていることと，興味が分解されるにしたがって，単語が徐々に周辺に散らばりながら離れていくことの2つの直観に一致する配置によると考えられる．クラスタリングを行なわない大円型の人気は，円型に比べてシンプルな構造のためと考えられるが，少しインターフェイスに慣れてくると，キーワードの出所が



表 4: インターフェイスの型の比較（%）

| | 木構造 | ピラミッド型 | 大円型 | 円型 | リスト型 |
|---|---|---|---|---|---|
| 検索語とキーワードの関連のわかりやすさ | 21 | 9 | 35 | 32 | 3 |
| 再検索のしやすさ | 35 | 9 | 23 | 18 | 15 |

表 5: 円型と海型の比較

| | 円型 | 海型 |
|---|---|---|
| 根本的な興味を表すキーワードの配置 | 中央 | 下方 |
| 興味の細分化の方向 | 中心から周辺 | 下方から上方 |
| 単語の集合としての配置 | 雑然 | 整然 |

より明確なクラスタリングありの手法が望まれる傾向にあった．シンプルさが望まれるという傾向は，次の質問の結果にもよく表れている．

「最も再検索しやすい」のはどのインターフェイスかと尋ねた場合においては，木構造のインターフェイスが35%と最も多く，円型と大円型の合計の41%とあまり変わらない結果が得られた．この結果は円型インターフェイスにおいて，円状に並べられた検索枠に単語を配置していくことで線形の検索式をイメージしにくいことや，キーワードがまとめて配置されていないために，キーワードの全体を把握するための時間がかかること，検索式更新のためのマウス操作の方向が一定でないなどの原因が挙げられる．また，リスト型インターフェイスもシンプルさでは負けていないため若干名の嗜好するユーザが存在するが，その値も15%に留まっており二次元検索インターフェイスの必要性を裏付けている．

アンケート結果により検索に効果的とみられる円型と海型のインターフェイス

について，ここまでで述べた比較をまとめたのが表5である．すなわち，ユーザの興味をもとにした検索語と検索支援キーワードの間の関わりを良く表す配置は円型であるが，再検索のためにはシンプルで見やすい海型のような配置が望まれている．この表以外の点について両者の比較を以下で述べる．検索枠の並びに関していえば，円型には端がなく環状であるが海型は直線上であるために端が存在する．このことは検索枠間の関連を表現する上で，直線の端と端をつないだ環状の方が枠間の組合せを一つ多く表すことができるためにやや優れるが，海型においては検索枠間の距離を計算して一列に並べているために，意味が最も離れた検索語を容易に確認できる面もある．

また円型における検索枠は左右だけでなく上下の座標ももつために，ユーザが初めに入力した検索語を最重要枠として，最も上に明示的に配置することができる[27]．ここまで述べてきたように両インターフェイスは一長一短であり，どちらを採るかはユーザの好み次第なのが現状であるため，さらに優れたインターフェイスの構築を目指しつつ，現時点では2つのインターフェイスをボタン一つで切替可能にして対応する．

## 6.2 検索実験によるシステム評価

のべ149の質問に対してサーチエンジンを用いて解答してもらう実験を行なった．これは同一のデータベースを対象として，興味表現支援システム（検索支援キーワードとインターフェイス）を用いる場合と用いない場合とで，検索を終了するまでの検索式の更新回数を測定した．ここでいう検索終了とは，質問の回答を得るまたはユーザが回答は存在しないと判断して検索を打ち切ったかのいずれかである．検索式の更新なしで解答が得られた質問を除いた実験結果を図23に示す．

この結果，本システムを用いた場合の方が更新の回数が少ないという結果が得

[27]海型においても色を変えることによって最重要枠を示すことはできるが，配置によって直観的に理解できる円型には劣る．



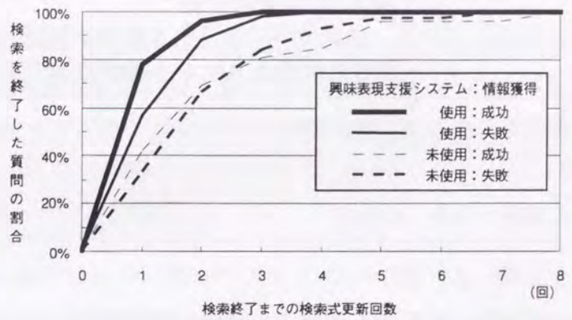


図 23: 検索式更新回数の比較

られた．これは，質問の解答が得られる場合においてもあきらめる場合においても，存在するWebページ情報をいち早く摑むことができ，目的の情報の有無および，目的の情報を得るために必要なキーワードを知ることができたことによる．特に質問の解答が得られる場合には，提供されるキーワードを用いて徐々に情報に近付くという目的が達せられているといえる．この「徐々に」とは「と金の遅早」[28]という将棋の格言のごとく「ゆっくり」の意味ではなく「確実に」という意味で本システムが作用していることに注目されたい．また問題の回答が存在するにも関わらず，回答のページにたどり着けなかった割合は本システムの有無に関わらず約20%であり，キーワード検索の問題点を解決し切れていない面もあるが，純粋に検索式更新回数の改善に効果を発揮したと結論付けることができる．

検索式の更新回数が少ないということはすなわち，検索における後戻りが少ないともいえる．具体的には，過度の絞り込みや単語の言い換えなどによる検索のやり直しや，不適切な単語による検索を自然に避けることができるということである．何の手がかりもなく関連語も与えられず，試行錯誤によって回数を重ねる従来の検索の不便さを思えば，提供されるキーワードによって存在する情報を確認しながら，

[28]一見遅くみえるような攻めでも，確実な攻めほど実は早いということ

表 6: 計算量の比較

| 処理 | 関連度学習システム | 興味表現支援システム |
|---|---|---|
| 事前学習コスト | $O(m^2 * n)$ | $O(n)$ |
| 検索コスト | $O(n)$ | $O(n)$ |
| キーワード選択コスト | $O(m)$ | $O(1)$ |

自身の興味を表すキーワードを自ら選びとる手間と引き換えに素早く情報を獲得できる本システムの方が総合的に勝るといえる．

## 6.3 興味表現支援システムの計算時間

本節では，興味表現支援システムと従来の関連度学習 [Chen 96] に基づいてキーワードを提供する関連度学習システムとの計算時間を比較する．この計算時間をまとめたのが表6である．ただし，$m$ は検索キーワードとして扱う単語の総種類数（日本語辞書に含まれる全名詞数）であり，$n$ は検索データベースがもつ全 Web ページの数である．表6の各項目について順に説明を行なう．

### 6.3.1 検索のための事前学習コスト

まず，検索を行なう前の段階においてどれだけの事前学習が必要であるかについて述べる．単語間の関連度を学習するシステムの場合，すべての名詞の組合せ（${}_mC_2$ 通り）に関して，データベース内の全 Web ページにおける共起頻度を計算するため，$m^2 * n$ に比例する計算量（$O(m^2 * n)$ と記述する）を要する．任意の語による検索が行なわれるサーチエンジンにおいて，この計算量は現実的ではない．ある形態素辞書における名詞の数は 10 万を越えるため，その各名詞ペアの関連度として 50 億通り以上の情報を保存しかつ，Web ページ情報の更新に伴なって逐次関連度を



計算し直す必要があるからである[29]．本システムでは各 Web ページごとに，含まれている名詞とその頻度を計算するだけであるため，単純に Web ページの数に比例した計算量で $O(n)$ となる．これは一般のサーチエンジンにおいても行なわれている，キーワード抽出やタグづけ，抄録作成等の前処理の最低限の計算量と同じである．

### 6.3.2 検索コスト

次に，実際に検索を行なう際のコストについて述べる．検索そのもののコストは存在する Web ページの総数に比例して $O(n)$ となる[30]．新仮説生成を行なう興味表現支援システムにおいては，関連語取得のために（検索式を表す木構造のネットワークのノード数 $t$ ＋興味仮説の数 $h$）の回数，サーチエンジンを用いている．サーチエンジンの検索コストは $O(n)$ であり，$t+h$ は通常 20 を越えない（3.4.2節の例では 9+2=11）ため，$O((t+h)n) < O(20n) = O(n)$ となり検索のオーダーとしては変わらない．また，検索式分割による場合では，分割された部分興味の数の回数だけサーチエンジンが用いられる．しかしこれも 10 個以上に分割されるような壮大な興味に関わる検索をユーザが行なうことは考えにくいため，検索コストはやはり $O(n)$ となる．このように，本システムにおける検索コストのオーダーは，通常の検索システムと変わらないものの実際には検索に数倍の時間がかかる可能性がある．そこで，まず通常の検索結果と同様に検索された Web ページのみを表示して，ユーザが出力された Web ページを見てから検索式の更新要求を起こすまでの時間内に検索支援キーワードを提供するために必要な検索を行なうことによって，ユーザに遅さを感じさせずに本システムを動作させる．

[29] 50 億回サーチエンジンで検索する計算量 (一秒間に 100 回検索ができるとして 579 日かかる) に相当する．

[30] 実際には，$O(\log n)$ や $O(\log m)$ にまでコストを下げることは可能である．

### 6.3.3 関連キーワード選択のコスト

最後に，ユーザの興味に関連するキーワードを選択する際のコストについて述べる．関連度学習システムにおいて，検索語と関連度の高い名詞をデータベース中の全名詞の中から一定数選ぶための計算コストは $O(m)$ である．なぜなら，各名詞との関連度を少なくとも一度は参照する必要があるからである．これに対して本システムでは，検索されたWebページの中から一定数のWebページ（本論文では20ページ）を抜き出し，その抜き出されたWebページ中にのみ存在する名詞の順位付けを行なって上位の名詞を取り出すため，出現する名詞の数は全名詞数に比べて十分小さく，処理時間は一定で定数オーダの $O(1)$ となる．

## 6.4 現在の興味表現支援システムの限界と課題

興味表現支援システムについて，本章ではその有効性を示すためにさまざまなことを述べてきたが，ここで本システムの限界についても述べる必要がある．特に本システムはWebページからキーワードを抽出してユーザに示し，キーワードをユーザが目的の情報へ辿り着くための道標としてきた．そのため，次の場合には目的のページを得るための有効なキーワードが得られず，目的のページに到達し難い．

1. ユーザが望むWebページが極端に少ない．
2. ユーザが望むWebページに共通のキーワードがない．

1. の点はすなわち，目的の情報の絶対数が少ないことが原因で目的とするWebページからはキーワードが抽出されないという問題である．これは，検索ユーザの興味が特殊な場合であり，多くの検索語によって正確に目的の情報を表す検索語の入力が必要となる．しかし，そのような特殊な情報の集合は2. の問題点であるユーザの興味を表す的確な言葉自体が存在しない可能性が高い．なぜなら，絶対数が少なくなるほどの特殊な情報の名称として，決まった通称が存在するとは考えにくいからである．このような場合にはユーザの希望を満たすページだけを得るための検索語がもとから存在しない，あるいは非常に複雑になるため，本システム以外でもユーザが入力した検索語を用いる検索技術全般において検索が難しいと考えられる．Webページを検索するために用いられる情報は，通常Webページに含められた言葉のみであるため，Webページに用いられている言葉と，ユーザの興味を表す言葉との間で関わりが生じなければ，目的の情報を得ることが叶わない．

その他の問題点として，本実験で用いたデータベースは実際のWWW上に存在するWebページの総数に比べて非常に少ないことが挙げられる．しかし，さまざまな実験結果によって示してきたように，データベースが含むWebページの総数が増した場合においても，提供される語をNOT条件として用いることで情報を絞り込み，不要な情報を除いたWebページの中からの関連キーワードを得て，目的のページに達することができると考えている．今後は，これを実証するために大規模なデータベースを構築して，より多くのユーザに実際に使ってもらうことによってシステムの実用化と完成を目指したい．

# 7　結論

本論文ではWWW上でサーチエンジンを用いて検索を行なうインターネットの検索ユーザが，ユーザ自身の興味を具体化して的確な検索式作成を行なえる興味表現支援システムを提案した．本システムにおいては，ユーザの興味に関連する単語を獲得，分類し，二次元平面上に配置した検索インターフェイスを備えている．ユーザはこのインターフェイスを用いて，ユーザの頭の中にのみ存在する興味を具体的な言葉として表現することができ，ユーザ自身の興味の中で未だ具体的でない部分の示唆を受けて，興味の具体化が行なえることを実験によって検証した．また，ユーザが提供される単語を用いて，後戻りや試行錯誤のない素早い目的の情報への到達が実現できることを示した．

情報の電子化とコンピュータの普及に伴なって，今後もさらに膨大なWWWなどの文書データベースへのアクセス要求が高まり，さまざまな情報が容易に獲得できる時代になる．新たな時代に向けて感性豊かな人間の興味は尽きることなく，ますます多様な人間の興味に対応した情報検索の手段が渇望されるであろう．そのような未来に向け，快適な情報アクセスを実現するための研究の一端として，本研究で提案する興味表現支援システムが貢献できれば幸いである．



# 謝辞

大阪大学における研究生活におきまして，学部4年生の時より現在に至るまでの6年間の長きに渡り，御指導・御鞭撻・有益な示唆を賜わりました谷内田正彦教授に深く感謝の意を表明します．また，本論文によって博士の学位を取得する以前に助手として採用して頂き，一人前の研究者としての第一歩を与えて下さいましたことに心より感謝申し上げます．

本論文の学位審査を行なって頂き，本論文を完成させる上で有益な御助言と示唆を賜わりました新井健生教授，産業科学研究所溝口理一郎教授に深く感謝申し上げます．

大澤幸生先生（現筑波大学大学院）におかれましては，星や砂の数にも匹敵するご厚情を賜わり，本当にお礼の申し上げようもございません．また日々，叱咤激励の連続で研究への意欲を駆り立てて頂きましたことは忘れられません．ありがとうございます．

八木康史先生には，研究内容に関する多くの御助言，ならびに仕事上の御助言を多数，賜わりましたことを深く感謝申し上げます．

岩井儀雄先生には学生時代，そして職場となった大学研究室におきまして，良き先輩の模範として多くを学ばさせて頂きました．深く感謝致します．

山口智浩先生（現奈良工業高等専門学校）には学部4年生の折りに，人工知能研究とは何かを深く考えるきっかけを与えて頂き，現在に至るまで多くの御助言を頂きましたことを深く感謝申し上げます．

また，本研究に関わる議論に加わって下さり，さまざまな有益な示唆と御助言を賜わりました大阪大学産業科学研究所の鷲尾隆先生，大阪市立大学の北村泰彦先生，NTTの松澤和光氏，阿部明典氏，藤本和則氏に深く感謝申し上げます．

楽しく快適な研究生活を送らせて頂きました谷内田研究室の先輩，同輩，後輩諸氏に感謝を申し上げます．本研究の一部を共に行ない，研究成果を共有させて頂き

ました野村勇治氏（現松下通信工業），中田正樹氏（現ミノルタ）に感謝致します．また，松村真宏氏（現東京大学博士後期課程）との研究に関するさまざまな議論は，研究の方向や不足する点を探る上で非常に有益であったことをここに記し，感謝致します．横山太郎氏，長原一氏，岡本充義氏（現松下電器）らとの研究その他の議論の中では，鋭気を養なうことができました．感謝致します．

最後に，常に暖かく見守ってくれた両親と，すべてを導き，すべてを益となして下さった神様に心から感謝します．



# 参考文献

[Armstrong 95] Armstrong, R., Freitag, D., Joachims, T. and Mitchell, T.: "A Learning Apprentice for the World Wide Web", *AAAI Spring Sympo. Series on Information Gathering from Distrubuted, Heterogeneous Environments*, (1995).

[渥美 97] 渥美雅保: 「Web ページからのユーザの興味の遺伝的アルゴリズムに基づく抽出」, 情報処理学会研究報告, 97-ICS-18, pp.13 – 18, (1997).

[Balabanovic 95] Balabanovic, M., Shohham, Y.: "Learning Information Retrieval Agents: Experiments with Automated Web Browsing", *AAAI Spring Sympo. Series on Information Gathering from Distrubuted, Heterogeneous Environments*, (1995).

[Beaulieu 97] Beaulieu, M.: "Experiments on interfaces to support query expansion", *Journal of Documentation*, Vol.53, No.1, pp.8 – 19, (1997).

[Buckley 95] Buckley, C. and Salton, G.: "Optimization of Relevance Feedback Weights", *in Proc. SIGIR'95*, pp.351 – 357, (1995).

[Bonissone 87] Bonissone, P.P. et al: "A Layered Architecture for Reasoning with Uncertainty", *in Proc. International Joint Conference of Artificial Intelligence (IJCAI'87)*, pp. 891 – 898, (1987).

[Bruza 97] Bruza, P.: "Query ReFromulation on the Internet: Empirical Data and the Hyperindex Search Engine", *in Proc. RIAO-97 Computer Assisted Information Searching on the Internet*, (1997).

[Charniak 91] E.Charniak: "Bayesian Network without tears", *AI Magazine*, Winter, pp.50 –63, (1991).

[Chen 96] Chen, H., Schatz, B., Ng, T., Martinez, J., Kirchhoff, A. and Lin, C.: "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project", *IEEE Trans. Pattern Analysis and Machine Intelligence(PAMI)*, Vol.18, No.8, pp.771 – 782, (1996).

[Cohen 95] Cohen, J: "Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting", *J. American Society for Information Science, 46*, pp.162 – 174, (1995).

[Eguchi 99] Eguchi, K., Ito, H., Kumamoto, A. and Kanata, Y.: "Adaptive Query Expansion Based on Clustering Search Results", 情報処理学会論文誌, Vol.40, No.5, pp.2439 – 2449, (1999).

[Fayyad 96] v Fayyad, U., Piatetsky-S., G., and Smyth, P.: "From Data Mining to Knowledge Discovery in Databases", *AI magazine*, Vol.17, No.3, pp.37 – 54, (1996).

[Furnas 87] Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T.: "The vocabulary problem in human-system communication", *Communications of the ACM*, Vol.30, No.11, pp.964 – 971, (1987).

[Goldberg 89] Goldberg, D.E.: "Genetic Algorithms in Search", Addison-Wesley, (1989).

[原田 97] 原田昌紀: 「サーチエンジン徹底活用術」, オーム社, (1997).

[Hearst 95] Hearst, M. A, Karger, D. R, and Pederson, J. O: "Scatter/Gather as a Tool for Navigation of Retrieval Results" *AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, pp.65 – 71, (1995).



[帆足 99] 帆足啓一郎, 松本一則, 井ノ上直己, 橋本和夫: 「文書間の類似度における単語寄与度を利用した検索式拡張手法」, 情報処理学会研究報告, 99-DBS-118, pp.17 – 24, (1999).

[インターネット白書 98] 日本インターネット協会:「インターネット白書 '98」, インプレス社, (1998).

[Krulwich 95] Krulwich, B.: "Learning User Interests Across Hetegenous Document Databases", *AAAI Spring Symposium on Information Gathering from Heterogeneous*, SS-95-08, pp.106 – 110, (1995).

[Lamping 95] John Lamping, Ramana Rao, Peter Pirolli: "A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies", Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'95), (1995).

[Langley 87] Langley, P., Simon, H.A., et al: "Scientific Discovery - computational Explorations of the Creative Processes", *MIT Press*, (1987).

[増井 91] 増井俊之: 「シグナチャと曖昧検索を用いた文書検索システム」, 日本 UNIX ユーザ会, 第 18 回 jus UNIX シンポジウム論文集, pp.9 – 16, (1991).

[Mori 99] M.Mori and S.Yamada: "Bookmark-Agent: Information Sharing of URLs", *Poster Proceedings of The 8th International World Wide Web Conference (WWW-8)*, pp.70 – 71, (1999).

[森田 96] 森田昌宏, 速水治夫: 「情報フィルタリングシステム -情報洪水への処方箋-」, 情報処理学会誌, Vol.37, No.8, pp.751 – 758, (1996).

[村田 99] 村田剛志: 「サーチエンジンを用いた Web ページ集合の視覚化」, 第 13 回人工知能学会全国大会論文集, pp.541 – 544, (1999).

[仲川 99] 仲川こころ，高田喜朗，関浩之:「検索目的を反映したカテゴリ構造に基づく WWW 検索支援」, 情報処理学会研究報告, 99-HI-82, pp.59 – 64, (1999).

[新田 99] 新田清，蓬莱尚幸，園部正幸:「文書クラスタリングを利用した検索質問展開手法の開発と評価」, 情報処理学会研究報告, 99-DBS-118, pp.9 – 16, (1999).

[Ohsawa 97] Ohsawa, Y. and Yachida, M.: "An Index Navigator for Understanding and Expressing User's Coherent Interest", *in Proc. International Joint Conference of Artificial Intelligence(IJCAI'97)*, Vol.1, pp.722 – 729, (1997).

[Pearl 93] Pearl,J.: "Aspects of Graphical Models Connected With Causality", *in Proc. of the 49th Session of the International Statistical Institute*, Tome LV, Book 1, Florence, pp.399 – 401, (1993).

[Robertson 94] S.E.Robertson, S.Walker, S.jones, M.M.Hancodk-Beaulieu and M.Gatford:"Okapi at TREC-3", *TREC3 Proceedings*, pp.109 – 125, (1994).

[Rocchio 71] J.Rocchio: "Relevance Feedback in Information Retrieval", *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall Inc., pp.313 – 323, (1971).

[Salton 83] G.Salton and M.J.McGill: "Introduction to Modern Information Retrieval", *McGraw-Hill Advanced Computer Science Series*, McGraw-Hill Publishing Company, (1983).

[Salton 88] G.Salton: "Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer", Addison-Wesley, (1988).

[Salton 90] G.Salton and C.Buckley: "Improving Retrieval Performance by Relevance Feedback", *Journal of the American Society for Information Science*, Vol.41, No.4, pp.288 – 297, (1990).



[Salton 97] Salton, G. and Buckley, C.: "Term-Weighting Approaches in Automatic Text Retrieval", *Readings in Information Retrieval*, pp.323 – 328, (1997).

[Santos 99] Santos Jr. E., et al: "Dynamic User Model Construction with Bayesian Networks for Intelligent Information Queries", *in Proc. FLAIR'99*, pp.3 – 7, (1999).

[塩沢 97] 塩沢秀和, 西山晴彦, 松下温 :「納豆ビュー」の対話的な情報視覚化における位置づけ, 情報処理学会論文誌, Vol.38, No.11, pp.2331 – 2342, (1997).

[砂山 99] 砂山 渡, 大澤 幸生, 谷内田 正彦:「事象毎の生起確率から未知事象発見を支援する手法とそのアンケート調査への適用」, 人工知能学会誌, Vol.14, No.2, pp.349 – 358, (1999).

[角 94] 角康之，堀浩一，大須賀節雄:「テキストオブジェクトを空間配置することによる思考支援システム」, 人工知能学会誌, Vol.9, No.1, pp.139 – 147, (1994).

[臼澤 99] 臼澤基紀，新垣紀子，野島久雄，石崎雅人:「WWW 検索行動における『戻る』行動と検索方針の変化との関係」, 情報処理学会研究報告, 99-HI-83, pp.61 – 66, (1999).

[吉川 98] 吉川耕平，西村英樹，稗田薫，宇都宮速人:「ネットワーク上の情報検索とブラウジング」ソフトウェアシンポジウム, (1998).

[渡部 99] 渡部勇, 三末和男:「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, pp.65 – 72, (1999).

[AltaVista] A search engine: AltaVista,
(URL) http://www.altavista.com/cgi-bin/query

[ChaSen] Chasen home page:
(URL) http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

[ditto] A search engine: ditto,
(URL) http://www.ditto.com/

[Excite] A search engine: Excite,
(URL) http://www.excite.com/

[goo] A search engine: goo,
(URL) http://www.goo.ne.jp/

[InfoNavigator] A search engine: InfoNavigator,
(URL) http://infonavi.infoweb.ne.jp/

[Infoseek] A search engine: Infoseek,
(URL) http://www.infoseek.co.jp/

[Lycos] A search engine: Lycos,
(URL) http://www.lycos.com/

[Mondou] A search engine: Rcaau-Mondou,
(URL) http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/

[NTT Directroy] NTT Directory,
(URL) http://navi.ocn.ne.jp/

[Thesaurus] Thesaurus-step2:
(URL) http://search.kcs.ne.jp/the/

[Yahoo] Yahoo! JAPAN: (URL) http://www.yahoo.co.jp/

[zakzak] News:zakzak,
(URL) http:/www.zakzak.co.jp/



# 研究業績

●学術論文誌

1-1 砂山渡・大澤幸生・谷内田正彦：事象ごとの生起確率から未知事象発見を支援する手法とそのアンケート調査への適用, 人工知能学会誌, Vol.14, No.2, pp.349 – 358, (1999).

1-2 砂山渡・野村勇治・大澤幸生・谷内田正彦：Web ページ検索におけるユーザの興味表現支援システム, 電子情報通信学会論文誌, Vol.J82-D-I, No.12, pp.1394 – 1402, (1999).

1-3 砂山渡・大澤幸生・谷内田正彦：ユーザの興味の構造を用いて関連検索キーを提示する検索支援インターフェイス, 人工知能学会誌, (投稿中).

●国際会議（査読付き）

2-1 Wataru Sunayama, Yukio Ohsawa and Masahiko Yachida: Validating Questionnaire Data Analysis System by Shrinking Iterative Questinnaires, In Workshop Notes on Validation, Verification and Refinement of AI, International Joint Conference of Artificial Intelligence(IJCAI'97), Nagoya, pp.65 – 66, (1997).

2-2 Wataru Sunayama, Yuji Nomura, Yukio Ohsawa and Masahiko Yachida: Refining Search Expression by Discovering Hidden User's Interests, in Proc. International Conference on Discovery Science'98, Fukuoka, pp.186 – 197, (1998).

2-3 Wataru Sunayama, Yukio Osawa and Masahiko Yachida: Search Interface for Query Restructuring with Discovering User Interest, in Proc. Knowledge-Based Intelligent Information Engineering Systems(KES'99), IEEE, Adelaide, pp.538 – 541, (1999).

2-4 Wataru Sunayama, Yukio Osawa and Masahiko Yachida: Support System for Refining Search Expression with Discovering User Interest, in Proc. International Conference on Computer Communication'99, Tokyo, I-1-14, (1999).

2-5 Wataru Sunayama, Yukio Osawa and Masahiko Yachida: Computer Aided Discovery of User's Hidden Interest for Query Restructuring, in Proc. International Conference on Discovery Science'99, Tokyo, pp. 68 – 79, (1999).

●国内発表（研究会）

3-1 大澤幸生・砂山渡・谷内田正彦：時系列観測データからの新仮説創発の支援，情報処理学会人工知能研究会資料 102-1, pp.43 – 48, (1995).

3-2 砂山渡・大澤幸生・谷内田正彦：事象毎の生起確率からの未知原因推定による大衆心理の推定，第 29 回人工知能学会人工知能基礎論研究会資料, pp.13 – 18, (1997).

3-3 砂山渡・野村勇治・大澤幸生・谷内田正彦：Web ページ検索における改善検索キー教示システム，第 33 回人工知能学会人工知能基礎論研究会資料, pp.49 – 54, (1998).

3-4 砂山渡・大澤幸生・谷内田正彦：Web ページ検索における検索式の構造を利用したユーザの興味表現支援システム，人工知能学会第 9 回 AI シンポジウム資料, pp.79 – 84, (1998).

●国内発表（全国大会）

4-1 砂山渡・大澤幸生・谷内田正彦：断片的な確率データからの学習システム，第 10 回人工知能学会全国大会論文集，pp.313 – 316, (1996).

4-2 砂山渡・大澤幸生・谷内田正彦：事象毎の生起確率からの未知原因推定による大衆心理の推定，第 11 回人工知能学会全国大会論文集，pp.563 – 564, (1997).



4-3 野村勇治・砂山渡・大澤幸生・谷内田正彦：ネットワーク情報から社会の新しいニーズを発見する大衆心理分析システム，情報処理学会第 55 回全国大会 (平成 9 年後期) 講演論文集 (4)，pp.435 – 436, (1997).

4-4 砂山渡・大澤幸生・谷内田正彦：Web ページ検索におけるユーザの興味表現支援システム，第 12 回人工知能学会全国大会論文集，pp.372 – 375, (1998).

4-5 砂山渡・大澤幸生・谷内田正彦：Web ページ検索におけるユーザの興味表現支援システム，第 13 回人工知能学会全国大会論文集，pp.498 – 501, (1999).