Webページ間の相対的な具体抽象関係の視覚化による情報収集支援

砂山　渡[†]　鮫島　聡志[†]　西原　陽子[††]

Information Collection Support by Visualizing Relative Concrete-Abstract Relationship among Web Pagees

Wataru SUNAYAMA[†], Satoshi SAMESHIMA[†], and Yoko NISHIHARA[††]

あらまし　近年のインターネットの普及により，個人が獲得できる情報量が飛躍的に増大している．そのため，情報を探す際に一般に用いられる検索エンジンにおいても，検索結果の中から有効な情報を選別する必要があり，探している情報と関係しそうなリンクをたどった上で，そのリンク先の内容を確認することが一般に行われている．しかし，リンク先の情報がどの程度具体的な記述となっているのか，若しくは抽象的，一般的な情報だけであるのかを，そのリンクのアンカータグや簡潔な要約のみから判断することは難しい．そこで本研究では，Webページ間の相対的な具体抽象関係を定義し，検索語について集められた複数のWebページを，ユーザが着目するWebページに比べて「具体的なWebページ」と「抽象的なWebページ」の評価を与えた上で，視覚化するシステムを提案する．また実験により，本システムが提示する，Webページ間の具体抽象関係の情報が，効果的な情報収集に役立てられることを確認した．

キーワード　相対的な具体抽象関係の評価，情報収集支援，情報検索，情報視覚化

1. まえがき

近年のインターネットの普及により，個人が獲得できる情報量が飛躍的に増大している．そのため，情報を探す際に一般に用いられる検索エンジンにおいても，検索結果の中から有効な情報を選別する必要があり，探している情報と関係しそうなリンクをたどった上で，そのリンク先の内容を確認することが一般に行われている．

リンク先の情報の，検索目的との類似性もさることながら，その情報の具体性が問題になることもある．例えば，ある用語について調べる際に，その用語の意味だけが知りたい場合と，意味を含めた背景情報まで幅広く集めたい場合とでは，必要とするWebページが異なってくる．すなわち，リンク先の情報がどの程度具体的な記述となっているのか，若しくは抽象的，一般的な情報だけであるのかを，そのリンクのアンカータグや簡潔な要約のみから判断することは難しい．

そこで本研究では，Webページ間の相対的な具体抽象関係を定義し，検索語について集められた複数のWebページを，ユーザが着目するWebページに比べて「具体的なWebページ」と「抽象的なWebページ」の評価を与えた上で，視覚化するシステムを提案する．本システムによって，例えばユーザが具体的な情報を望む場合には「具体的なWebページ」と評価されたWebページの内容を確認していけばよく，情報収集の効率化が期待できる．

本研究では，相対的な「具体抽象関係」の定義を以下のように与える．すなわち，二つのテキストAとBがあったときに，Aに比べてBが，「異なる内容を含む」または「より詳しい説明を含む」ときに，BはAに比べて具体的，AはBに比べて抽象的とする．これは，必ずしもオントロジカルな意味での，下位概念，上位概念などには対応しない．

例えば「国会」について，「国会とは法律を作るところ」とだけ記述されているテキストに比べて，この記述に加えて「a. 国会を構成するのは衆議院と参議院」「b. 国会は国会議事堂で行われる」「c. 国会は憲法による国権の最高機関，かつ唯一の立法機関」など，国会

[†] 広島市立大学大学院情報科学研究科，広島市
Graduate School of Information Sciences, Hiroshima City University, Hiroshima-shi, 731–3194 Japan
[††] 東京大学大学院工学系研究科，東京都
School of Engineering, The University of Tokyo, Tokyo, 113–8656 Japan

について他の説明が含まれているテキストはより具体的となる．また，a. のテキストと比較して，衆議院と参議院のそれぞれの議員の定数や任期，衆議院の優越について，より詳しい説明がなされているテキストも具体的となる．

以下本論文では，**2.** でほしい情報を探す支援となる研究の背景について述べ，本研究の位置付けを明らかにする．**3.** で，Web ページ間の相対的な具体抽象関係の定義を与えたのち，提案する Web ページ集合の具体抽象関係の視覚化システムについて述べる．**4.** で，システムが抽出する具体抽象関係の精度評価実験，**5.** で，相対的な具体抽象関係の視覚化による情報収集支援についての実験を述べた上で，**6.** で本論文を締めくくる．

2. 研究背景

検索結果として得られる Web ページ集合から，欲しい情報を探すための検索支援は古くから行われて来ている．その多くは，限られた検索語によって検索意図を表現することの困難さから生じる，検索意図のあいまいさや，検索語の多義性を解消するためのもので，内容の類似性に基づく分類や情報収集を行っている．例えば，関連フィードバック [3] によって検索結果に含まれる単語をもとに再検索を行う研究や，Web ページ間のリンクやキーワードの類似性から Web ページ間の関係を視覚化する手法 [4] があり，内容の類似度を測ることを目的とした手法においては，内容が全く同じ Web ページ同士の類似度が最も高くなる．しかし本研究においては，より情報量が多い，若しくは少ないページを探すことの支援を目的としているため，これらの手法を単純に適用することはできない．

Web ページ間の関連をリンク情報から測る研究として，比較対象となる二つの Web ページの URL それぞれを検索語として，それぞれの検索結果として得られる Web ページ集合を比較して，共通に含まれるリンクの割合を用いる研究 [1] やユーザが着目している Web ページからリンクをたどり，たどった回数の少ない Web ページを関連のある Web ページとして視覚化する手法 [2] がある．しかし，リンク情報は Web ページの内容を直接表していないため，Web ページ間の類似度，更には具体抽象関係を推定することは難しい．

また，検索結果として提示される，各 Web ページの要約や，各 Web ページから抽出されるキーワードによって，人間が具体抽象関係を推定することを期待する方法も考えられる．しかし，具体的な情報を表すキーワードはマイナーかつ低頻度な専門用語となることが多く，キーワードとして提示することが困難であることから，テキストが含む名詞の種類数を用いることによって，キーワード抽出を行わずに，それらのキーワードを間接的に含んだ指標によって，具体抽象関係の抽出を図る．

すなわち本研究では，検索結果の Web ページ集合に対して，Web ページ間の類似度を測る従来手法の利用によって，一定以上の類似度がある Web ページを集め，それら各 Web ページに含まれる単語をもとに，具体，抽象の度合を表す数値を付与する．また，具体抽象の度合に応じた色付けと，Web ページ間の類似性をもとに，二次元平面上に Web ページを配置したインタフェースを実装し，情報収集の効率化を目指す．

3. 具体抽象関係の視覚化システム

本章では，Web ページ間の相対的な具体抽象関係を評価して，Web ページを分類して視覚化するシステムについて説明する．

3.1 システムの全体構成

図 1 にシステムの全体構成を示す．入力はユーザが与えた検索語による検索結果の Web ページ集合，及びその中の，ユーザが着目する特定のページとする．集められた Web ページ集合は，ユーザが与えた検索語を含んでおり，この検索語を具体抽象関係の基準となる話題として扱う．次に，ユーザが着目する Web ページ，及び Web ページ集合中の各 Web ページから，検索語に関する話題部分のテキストを抽出する．

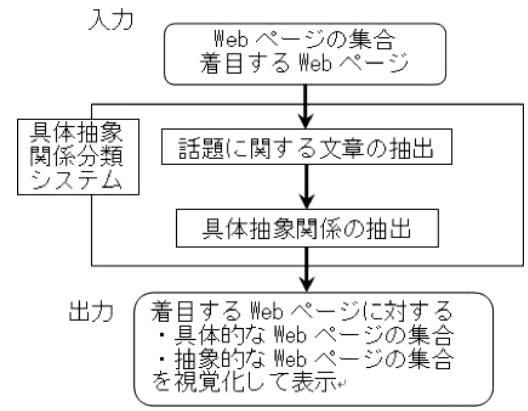

図 1　具体抽象関係の視覚化システム
Fig. 1　Concrete-abstract visualization system.

システムは抽出されたテキストをもとに，ユーザが着目する Web ページとの具体抽象関係を表す数値を，検索結果の各 Web ページに与える．最後に，二次元平面インタフェース上に，得られた具体抽象関係を出力する．ユーザはこの出力画面を見て，現在着目しているページよりも具体的な情報がほしい場合には，具体的と表示されている Web ページを積極的に確認していく．以下の各節で，各モジュールの詳細について述べる．

3.2 システムへの入力

本システムへの入力として，ユーザが与えた検索語による検索結果の Web ページ集合，及びその中のユーザが着目する特定のページを与える．ただし，検索結果のすべてのページを対象としても，ディスプレイ上の出力をユーザが十分に認知できないおそれがあるため，検索エンジンの出力の上位 100 件を用いる．

3.3 話題に関する文章の抽出

各 Web ページには，検索意図を表す検索語に関係する記述が含まれる部分と，それ以外の話題について書かれている部分とがある．検索意図に合わない内容について，詳しい説明がなされている Web ページが，具体的な情報として判断されることを避けるため，各 Web ページから検索語に関係する話題について書かれた文章を抽出する．

そのため，単語の集合で表されるある話題について，関連する文章を抽出する手法 [6] を適用する．アルゴリズムは以下のようになる．

1. Web ページ中で検索語と関連の強い単語を抽出し，その単語集合を話題とする．
2. Web ページ中の各文に，話題との関連度を表す数値を付与する．
3. Web ページ中で，しきい値以上の値をもつ最初の文から，しきい値以上の値をもつ最後の文までを，話題に関連する文章として抽出する．

単語の抽出には形態素解析器 ChaSen [7] を用い，以下では名詞のみを評価の対象として扱う．以下の項で，このアルゴリズムの各ステップについて述べる．本アルゴリズムは従来研究 [6] によるものであり，本研究の目的に応じて，係数などの定数が異なる以外は同一となっている．

3.3.1 話題を表す単語の抽出

話題を表す単語の抽出には，入力された検索語に関連する重要文を抽出する展望台システム [5] を用いる．展望台システムは，検索語を，話題を表す単語の初期集合 V として，Web ページ中のすべての単語 w に，式 (1) による評価値 $Val(w)$ を与える．また，この評価値による上位（単語種類数の 4%で最大 5 個）に含まれかつ，まだ V に属さない単語を，新たに話題を表す単語として集合 V に追加する．ただし，式中の条件付き確率 P は，句点で区切られる 1 文中に単語 v と単語 w が同時に現れる可能性を表す．

$$Val(w) = \prod_{v \in V} P(v \mid w) P(w \mid v) \tag{1}$$

3.3.2 各文の話題との関連度の評価

前項で得られた，話題を表す単語集合 V を用いて，各文に話題との関連度を表す数値を与える．すなわち，単語集合 V に含まれる単語が現れる文を高く評価するとともに，その近くにある文も話題に関係する文として評価する．この話題を表す単語からの距離を，句点と，Web ページ中に存在するタグ（ブロックレベル要素（表 2））をもとに，表 1 のように与える．

各文 S_i の評価値 $Sent(S_i)$ を，その文が含む話題を表す単語の評価値 $Base(v)$ と，他の文が含む話題を表す単語の評価値 $Eff(v)$ との総和として，式 (2) で与える．

$$Sent(S_i) = \sum_{v \in S_i} Base(v) + \sum_{S_{j,\ j \neq i} \in D} \sum_{w \in S_j} Eff(v) \tag{2}$$

$$Base(v) = 100 \times Val(v) \tag{3}$$

$$Eff(v) = max\{80 \times Val(v) - Dis_{ij},\ 0\} \tag{4}$$

ただし，v は話題を表す単語集合 V の要素，i, j は文章中の文の番号，D は文章全体を表す．また，Dis_{ij}

表 1 Distance の値
Table 1 Values of distance.

区切りの種類	話題からの距離
句点	1
ブロックレベル要素	10

表 2 ブロックレベル要素
Table 2 Block-level tags.

ADDRESS, BLOCKQUOTE, CENTERDIR, DIV, DL FIELDSET, FORM, H1, H2, H3, H4, H5, H6, HR ISINDEX, MENU, NOFRAMES, NOSCRIPT, OL, P, PRE, TABLE, UL, DD, DT, FRAMESET, LI, TBODY TD, TFOOT, TH, THEAD, TR

は，文 S_i と文 S_j の間にある句点とタグの数によって，表 1 の値を積算した値を表す(注1)．

3.3.3 話題を表す文章範囲の特定

話題を表す文章として，前項で与えた各文の評価値が，しきい値以上となる最初の文からしきい値以上の最後の文までの連続した文を抜き出す．しきい値の決め方の詳細については次章で述べる．

3.4 具体抽象関係の評価

前節の方法によって抜き出された，各 Web ページの検索語に関する文章をもとに，Web ページ間の具体抽象関係を評価する．すなわち，ユーザが着目する Web ページ P_v と，検索結果に含まれる各 Web ページ P_i を比較し，検索結果の Web ページに具体抽象の度合を表す，次式の評価値 $CA(p_v, p_i))$ を与える．ただし，$Nouns(P_i)$ は Web ページ P_i の検索語に関する文章中の名詞種類数を表す．

$$CA(P_v, P_i) = Nouns(P_i) - Nouns(P_v) \qquad (5)$$

結果として，この式による値が大きい Web ページほど，検索語に関して具体的な記述がなされているページとして，逆に値が小さい Web ページほど抽象的なページとして評価する．

名詞種類数を評価に用いた理由としては，より多くの具体的な情報を記述するためには，多くの種類の単語を使う必要があると考えたことによる．すなわち，文章量が多くても，ただ同じことを繰り返し記述しているだけでは，情報が具体的とはいえない．また，一般的にだれもがよく知っている単語がたくさん並んでいるだけの，絶対的評価では抽象的と考えられる Web ページであっても，本研究で抽出する具体抽象関係は，ユーザが着目する Web ページに対して相対的に与えることを目的としているため，ユーザが着目する Web ページに比べるとより多くの情報を含む具体的な Web ページとして評価されることがある．

一方，式 (5) による評価値の絶対値が 0 に近い場合には，両者にははっきりとした具体抽象関係がないと考えられる．どの程度の絶対値がある場合に，具体抽象関係が認められるかについては，次章で実験により検証する．

3.5 具体抽象関係の視覚化

図 2 にインタフェースの出力画面を示す．インタフェースでは，中心にユーザが着目している Web ページ（星印）が配置され，その他の Web ページは各 Web ページ間の式 (5) による名詞種類数の差の絶対値をばねに与える力として，ばねモデルのアルゴリズム [8] を用いて配置する．ばねモデルを用いる理由は，できるだけ多くの情報を重ならないように表示するため，また具体若しくは抽象の程度が近い Web ページを近くに配置するためである．各 Web ページの表示には，タイトルの先頭 6 文字を用い，実際のインタフェース上では，表 3 の色と印によって，ユーザが着目している Web ページ（User noticed），ユーザが現在見ている Web ページ（User gazing），具体的な Web ページ（Concrete），抽象的な Web ページ（Abstract），その他の Web ページ（Others）が明確に区別できるように表示される．

インタフェース上では，ユーザはマウスポインタを移動させることができ，ユーザがカーソルを合わせて見ている Web ページが赤い■印で表示され，**3.3.1** で抽出した話題を表す単語が画面下部に表示される．その状態で右クリックを押すと，別ウィンドウのブラウザ上に実際の Web ページの内容が表示される．また，カーソルを合わせた状態でダブルクリックを行うと，その Web ページを新たな着目 Web ページとして，表示中の Web ページ集合に対して具体抽象関係の再計算を行い，再描画を行う．

3.6 想定するシステムの使用方法

本システムでは，情報を探すユーザが，現在着目している Web ページよりも詳しい（具体的），若しくは簡潔な（抽象的）情報を含む Web ページを探す支援を行う．

例えば,「総理大臣」に関する情報を探しているユーザが「総理大臣」を検索語として検索を行い，ユーザが図 2 の出力の中から，総理大臣という言葉の意味について書かれている図 3 の Web ページを閲覧していたとする．

総理大臣に関する，より詳しい情報を望む場合には，インタフェース上で黄色い●印で表示されている Web ページの中から，例えば，内閣総理大臣に関する地位や資格，代理の話など，もとのページにはなかった情報を含むページ（図 4）を発見できる．また，更に詳しい情報を望む場合には，今見ていたページを再び着目 Web ページとして再描画を行い，再び黄色の●印の Web ページを探すことで，歴代の総理大臣について書かれた Web ページなどを発見できる．逆に，総

(注1)：表 1 の距離の数値や，式 (3) と式 (4) の係数は，経験的に良好な値をとる数値として定めた．

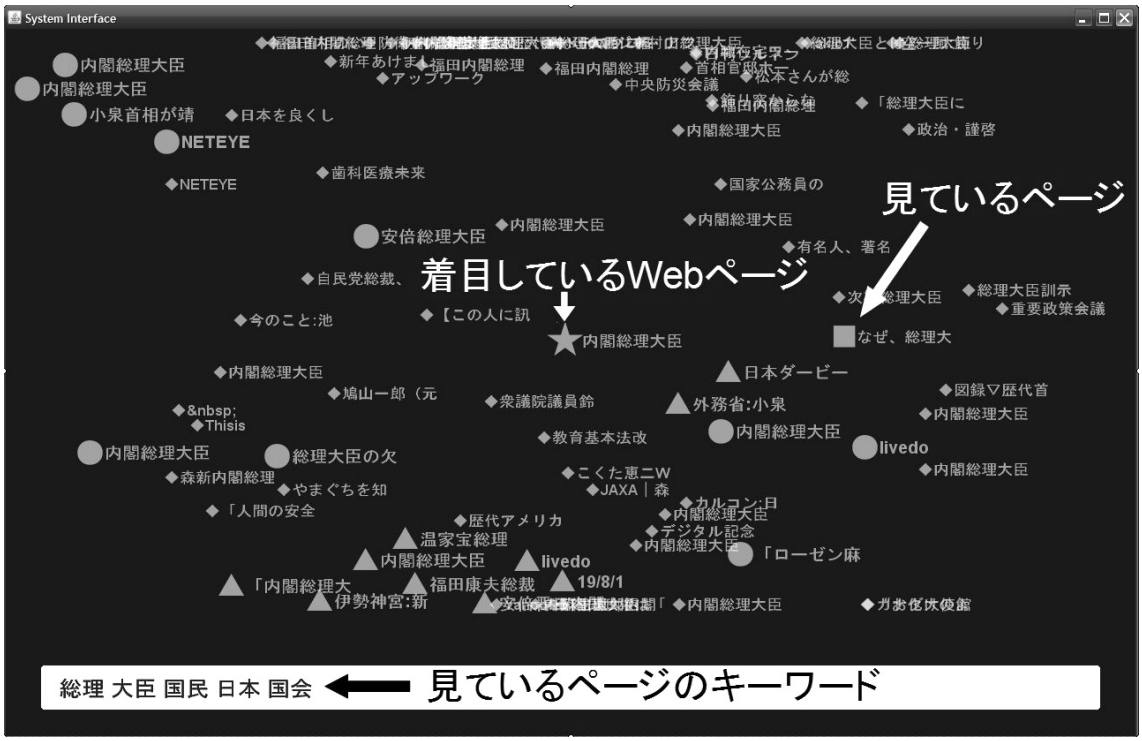

図 2　システムの出力インタフェース画面
Fig. 2　Output interface for proposed system.

表 3　Web ページの色と記号
Table 3　Colors and symbols of Web pages.

Web page	Color	Symbol
User noticed	green	★
User gazing	red	■
Concrete	yellow	●
Abstract	light blue	▲
Others	white	◆

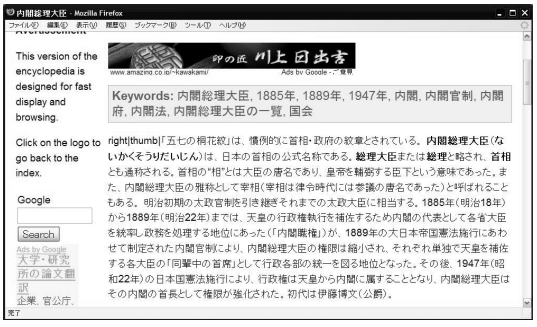

図 3　基準 Web ページの例
Fig. 3　Example of a basic Web page.

理大臣についての簡潔な情報を得たい場合には，青い▲印の Web ページを確認することで，図 5 のように，簡潔に記述された総理大臣の記事などが発見できる．

現在のシステムは，ユーザが望む特定の話題に関する情報の場所を，直接提示するには至っていない．しかしこの探索の最初の段階として，閲覧中のページと比較して，ユーザが望む Web ページの具体抽象の程度が異なる Web ページの集合を示すことで，情報源の発見や情報収集の効率化を図ることができると考えている．

4.　相対的な具体抽象関係の抽出精度評価

本章では，Web ページ間の相対的な具体抽象関係を抽出する手法の精度，並びに評価式のしきい値を決めるために行った実験について述べる．すなわち，被験者によって認定された具体抽象関係のペアと，システムが出力する具体抽象関係との比較を行って精度を評価する．その際，システムのしきい値を様々に変化させた結果とともに，しきい値についての考察を加える．

4.1　被験者による具体抽象ペアの作成

表 4 に示す六つの検索語それぞれを検索エンジン

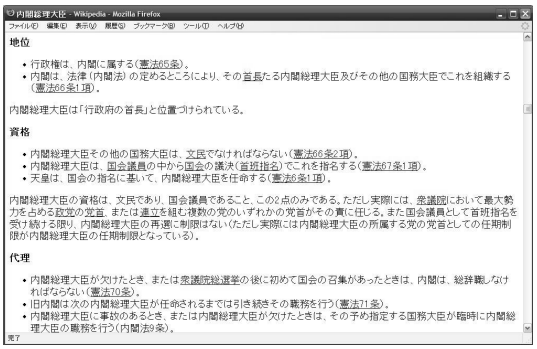

図 4　基準ページより具体的な Web ページの例
Fig. 4　Example of a concrete Web page.

図 5　基準ページより抽象的な Web ページの例
Fig. 5　Example of an abstract Web page.

表 4　精度評価実験に用いた検索語と具体抽象ペア数
Table 4　Search keywords for concrete-abstract evaluation.

Keywords	Web hit numbers	# of conc-abst pairs
イベリア半島	112,000	24
国会議事堂	1,080,000	26
センター試験	2,550,000	37
総理大臣	2,600,000	39
裁判員制度	3,070,000	37
初音ミク	7,100,000	19

Google [9] に与え，それらの検索結果の上位から，各検索語についての説明を含まないページを除いて 10 ページずつ用意し，得られた 10 ページ内の任意の 2 ページ間の相対的な具体抽象関係（各検索語ごとに 45 ペア）について評価を行った．被験者は，情報科学を専攻する大学生及び大学院生 12 名とし，すべての検索語について，印刷した 10 ページを「検索語について詳しく書かれてある」順に，制限時間を 5 分以上 10 分未満として，一列に並べてもらった．ただし同程度に詳しいと判断した場合には，同順位として並列に並べること，また検索語に関連する内容ではないと判断したページは，順位付けの列から除くように指示を与えた．最終的に，任意の 2 ページ A，B のペアに対して，ページ A がページ B よりも詳しいと判断した被験者が，逆の判断をした被験者よりも 6 人以上多かった場合に，ページ A はページ B よりも詳しいと同定した．この結果を，表 4 の具体抽象ペア数に示す．

表 5　各検索語で，F 値が最大になったときの適合率，再現率，式 (2) と式 (5) のしきい値
Table 5　Precisions, recalls, and thresholds for Eq. (2) and Eq. (5) where F-measures were maximum.

Keywords	式 (2)	式 (5)	F-measure	Prec.	Rec.
イベリア半島	73	35	0.65	0.58	0.75
国会議事堂	29	70	0.56	0.54	0.58
センター試験	29	62	0.72	0.78	0.68
総理大臣	61	47	0.86	0.87	0.85
裁判員制度	97	1	0.79	0.73	0.86
初音ミク	73	93	0.76	0.87	0.68

この表から，検索ヒット件数が少ない，より具体性の高い単語ほど具体抽象関係が認定されにくく，逆にヒット件数が多い一般的な単語ほど，具体抽象関係が認定されやすかったことが分かる．これは，一般的な（オントロジーの概念階層の上位に位置する）単語ほど，具体性の幅が広くなる（多くの下位概念をもつ）ことに準じていると考えられる．しかし「初音ミク」においては，ヒット件数とは逆にペア数が少なくなった．これは,「初音ミク」が固有名詞であること，またヒット件数がネット上で大きく話題になったことに起因し，単語の一般性を反映しているとは考えにくいことから，固有名詞は，具体性の幅が広くなりにくいと考えられる．

4.2　具体抽象ペアの抽出精度評価

3.3.3 で述べた話題を表す文章範囲を特定するために設けた式 (2) のしきい値，及び具体抽象関係の判断に用いる名詞種類数の差，式 (5) のしきい値を，それぞれ 0 から 100 の範囲で 1 刻みで網羅的に調べ，システム出力と，先の被験者による具体抽象関係とを比較した．適合率（Prec.）と再現率（Rec.）の調和平均である F 値が最大となったときの，各々の値としきい値を表 5 に示す．

結果から，適合率，再現率ともに 5 割から 8 割の値となっており，提案手法により一定の具体抽象関係を

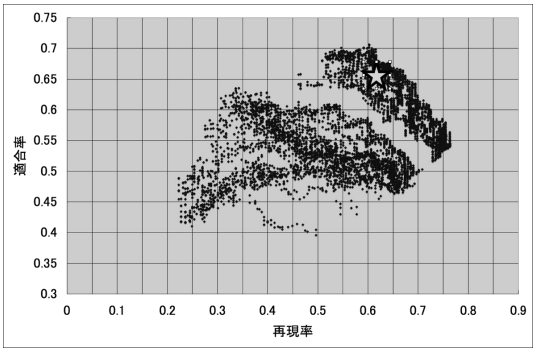

図 6　全しきい値による適合率と再現率の分布（星印は本論文で設定したしきい値による点を表す）
Fig. 6　Distribution of precision and recalls by thresholds. (The star denotes a point by the thresholds defined in this paper)

抽出することができることが分かった．しかし，しきい値の値にはばらつきがあり，動的なしきい値決めは現時点では困難と考えられる．式 (2) のしきい値は，検索語の違いよりも集められた Web ページに含まれる，検索語と無関係なテキスト量に依存すること，また，式 (5) のしきい値は，具体抽象の判断基準となることから，これらのしきい値は，一般名詞や固有名詞といった名詞の種類によって決められる値ではないと考えられる．

そのため，図 6 に示す全しきい値による適合率と再現率（いずれも六つの検索語の平均）の分布をもとに，今回は，式 (2) のしきい値を 30，式 (5) のしきい値を 70 として定めた．このときの，F-measure，適合率，再現率の平均は，それぞれ 0.64, 0.65, 0.63（図中の星印）となり，各検索語ごとの適合率，再現率もすべて 0.5 を超えている．

最後に，二つのしきい値と適合率，再現率との関係を述べる．今回定めた二つのしきい値について，それぞれの値からプラスマイナス 20 の範囲においても，F-measure の平均値は 6 割を超えていた．そこで，実際の運用の際には，以下をもとに適合率と再現率のいずれを重視するかによって，調整することも可能と考えられる．

- 式 (2) の本文抽出のしきい値は，大きくなるほど用いられるテキスト範囲が限られるようになり，範囲内の名詞種類数が少なくなり，名詞種類数の差によるシステムの出力数が減る．そのため，しきい値を上げることで，適合率が上がり，再現率が減る傾向がある．

表 6　視覚化の評価実験に用いた検索語
Table 6　Themes (search keywords) for experiments.

1	総理大臣
2	センター試験
3	初音ミク
4	国会議事堂
5	裁判員制度

- 式 (5) の名詞種類数の差のしきい値は，大きいほど具体抽象とみなされる条件が厳しくなるため，システムの出力数が減る．そのため，しきい値を上げることで，適合率が上がり，再現率が減る傾向がある．

5. 相対的な具体抽象関係の視覚化による情報収集支援実験

本章では，Web ページ間の相対的な具体抽象関係を視覚化するシステムが，ユーザの情報収集に役立てられることを確認するために行った実験について述べる．

5.1 実験内容

実験は，表 6 に示す五つの検索語を，検索エンジン Google [9] に与えて得られる検索結果 100 件を用いて，その中から選んだ基準ページに比べて，より詳しい情報が含まれるページ，または簡潔に情報が書かれているページを探してもらうことで行った．基準ページは，検索結果の中から，検索結果全体の中で，具体，抽象のどちらにも偏っていないページを無作為に選んだ．

比較インタフェースとして，図 2 において，具体と抽象の色付けと印付けを行わないインタフェースを用意した．被験者は，情報科学を専攻する大学生及び大学院生 20 名とし，検索語とインタフェースの組合せ，及び検索語の順序とインタフェースの使用順序には偏りが生じないようにした．すなわち一人の被験者には，五つの検索語について提案インタフェースと比較インタフェースを用いて，それぞれ詳しいページ及び簡潔なページを探してもらった．

評価は，基準ページとその基準ページをもとにした検索結果のインタフェースを提示した状態から，詳しい情報を含むページ，または簡潔に情報が書かれているページを見つけるまでに，かかった時間と，内容を確認するためにブラウザ上で内容を確認した Web ページの数を測ることで行った．1 回の制限時間は 5 分として実験を行った(注2)．本実験においては，被験者に「詳しさ」や「簡潔さ」に関する明確な基準が与

(注2)：1 名の被験者が，制限時間の 5 分を 1 回だけ使い切った．

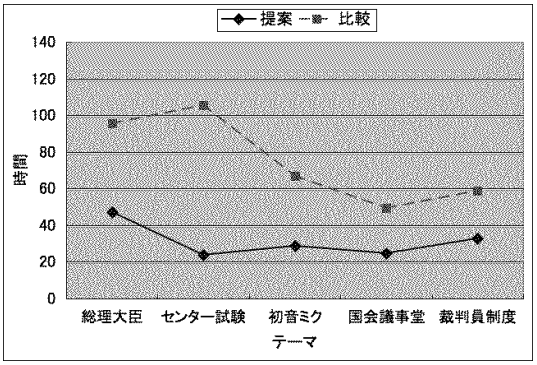

図 7　詳しいページを見つけるまでにかかった時間
Fig. 7　Times for finding a concrete Web page.

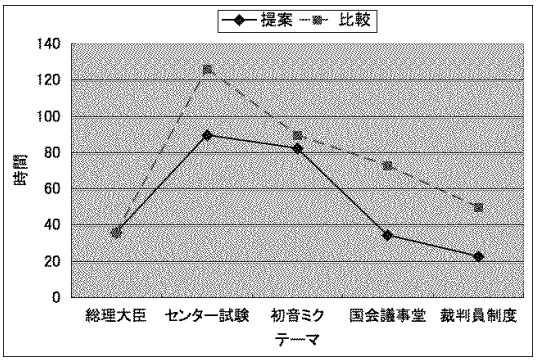

図 8　簡潔なページを見つけるまでにかかった時間
Fig. 8　Times for finding an abstract Web page.

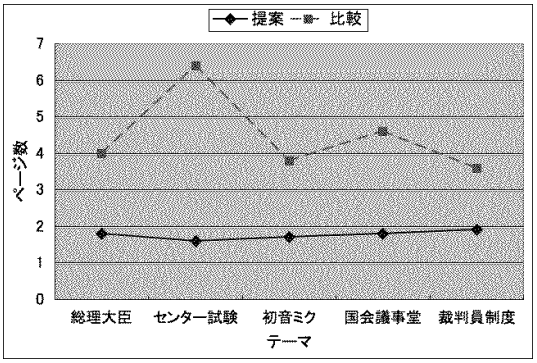

図 9　詳しいページを見つけるまでに開いたページ数
Fig. 9　Number of opened Web pages for finding a concrete one.

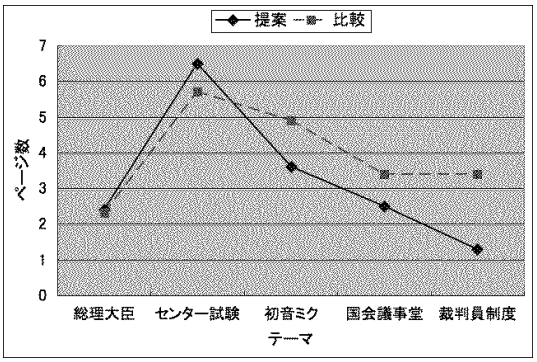

図 10　簡潔なページを見つけるまでに開いたページ数
Fig. 10　Number of opened Web pages for finding a abstract one.

えられないことから，被験者の回答の質を確保するために，Web ページを回答する際には，そのページのどのような点が，詳しいまたは簡潔であるかも併せて回答してもらった．

5.2 実験結果

まず，被験者が Web ページを回答した際の根拠として，「詳しい情報が含まれる」の理由には，基準ページにはないが，見つけたページには載っている情報を具体的に挙げた回答がほとんどであった．また「簡潔に情報が書かれている」の理由には，文章が短くまとめられていて，見やすい，読みやすい，という回答が多かった．これらのことから，被験者に Web ページを探してもらう作業は適切に行われたと考えられる．

各検索語について，詳しい Web ページを見つけるまでに掛かった時間の平均値を図 7 に，簡潔な Web ページを見つけるまでに掛かった時間の平均値を図 8 に示す(注3)．

図 7 の詳しいページを見つけるタスクでは，すべての検索語において提案インタフェースの時間が短く，比較インタフェースに比べて，およそ半分の時間で済んでいる．このことから，提案システムが詳しいページを見つけるためのナビゲーションの役目を果たしたことが分かる．

図 8 の簡潔なページを見つけるタスクでは，全体的に提案インタフェースの時間の方が短く済んでいるものの，「総理大臣」「初音ミク」においては，比較インタフェースとの時間差はほとんどなく，「センター試験」「初音ミク」においては，提案インタフェースの他のテーマに比べて時間がかかる結果となった．「センター試験」で提案インタフェースの時間がかかった理由は，検索結果として集められた Web ページ集合には，ニュース記事やブログ記事が多く含まれており，比較的簡潔なページが集まっていたため，それらの中

(注3)：グラフ上の各点は，10 名の平均値となっている．

から，より簡潔なページを特定することが難しくなったためと考えられる．すなわち，具体抽象の差が少ないページ集合が集められたときには，システムがそれ以上の支援を行うことは難しい．

次に，詳しいページを見つけるまでにブラウザで開いたページ数の平均値を図 9 に，簡潔なページを見つけるまでにブラウザで開いた Web ページ数の平均値を図 10 に示す．

開いたページ数においても，簡潔なページを探す際の「総理大臣」「センター試験」を除いて，提案インタフェースの方が，実際にブラウザ上で確認した Web ページの数が減少しており，より少ない労力によって，情報を収集できることが分かる．情報を確認する回数を減らせるということは，単に労力が減るというだけでなく，検索の必要を生じさせた人間の活動に，余分な情報を入れることなく集中して行うための支援となる．

本システムでは，一部の検索語では，他の課題に比べて時間がかかっており，効果が十分ではない点があった．すなわち，Web ページ集合に，具体抽象の偏りが少ない Web ページ集合が与えられた場合や，複数の意味で使われる単語が与えられた場合などにおいては，十分な情報ナビゲーションが行えない可能性があると考えられる．

前者の点については，今回定めた，具体抽象のしきい値を超える Web ページが全くない場合については，動的にしきい値を変更することで，やや具体的，やや抽象的なページについての情報も出力する方法も考えられる．しかしその精度は下がるため，あくまでしきい値を超える Web ページがない場合の方策として位置づけられると考えている．

後者の点については，現在は Web ページの単語種類数でのみ具体抽象関係を測っているため，検索語のどのような点で具体的であるかまでの情報が得られない．そのため，具体的な情報を探す際には，具体的と評価された Web ページをある程度網羅的に探す必要が生じるが，より具体的な Web ページほど，より多くの情報を含んでいるとも考えられるため，検索語の多義性について，現在でも部分的には対応できると考えている．しかし，より十分な支援とするためには，今後どのような点で具体的なのかをキーワードによって示すなどの改良を行っていきたい．

以上のことから，本提案システムを用いることで，閲覧中の Web ページに比べて，より詳しい具体的な Web ページ，また簡潔で抽象的な Web ページを，本システムを用いない場合に比べて，少ない時間と労力によって収集できることが分かった．

5.3 システムの実行時間

ユーザが検索語を与えてから，検索結果 100 件の具体抽象関係がインタフェース上に出力されるまでの，テキスト収集，話題に関する文章の抽出，テキスト間の具体抽象度計算，の三つの処理時間（CPU：Core2duo 2.33 GHz，メモリ 2 GByte）について述べる．

テキスト収集に関しては，現在のシステムでは，他の検索エンジンの検索結果を流用しているために，それらをダウンロードする必要が生じており，100 件のテキストをダウンロードするために，およそ 90 秒の時間を要している．しかしこの時間は，データベースをシステム側で構築する，または検索エンジン側と提携して構築した場合には，考慮する必要がなくなる．

次に，本システムの一つ目のモジュール「話題に関する文章の抽出」においては，展望台システム [5] を用いたキーワード抽出と，話題に関するテキストセグメンテーションを行っているが，100 件のテキスト（約 1 MByte）に対して，約 20 秒の時間がかかっている．

また具体抽象度の計算モジュールについて，インタフェース上で各 Web ページを配置するために，すべての Web ページの組合せについて，名詞種類数の差を計算しており，時間計算量のオーダは，Web ページ数 n に対して $O(n^2)$ となるものの，出力 Web ページ数を 100 件としたときには，こちらも 20 秒程度の時間で処理している．より多くの Web ページの出力を想定する場合においても，基準ページとの具体抽象関係の計算は $O(n)$ の時間計算量で終わるため，インタフェース上での各 Web ページの配置方法を工夫することや，各モジュール間でのデータの受渡し方法を改善することで，処理時間の短縮が可能になると考えている．

本システムの各モジュールの処理は，プログラムの改良による計算時間の短縮が可能であり，数秒程度ですべての処理を実行，実用化が可能になると考えている．

6. むすび

本論文では，Web ページ間の相対的な具体抽象関係を抽出して視覚化するシステムを提案し，本システムが情報収集支援に有効であることを実験により確認した．

今後は，基準ページに比べて，単純に具体的，抽象的というだけではなく，どのような点で具体的，抽象的なのかについての情報を提示するシステムの改良を行うこと．また，各 Web ページの話題の重なり具合や，特定のサイトからの情報などをインタフェース上に明示することで，具体的な話題の情報獲得に役立つシステムの構築を目指したいと考えている．

文　献

[1] 村田剛志，“参照の共起性に基づく Web コミュニティの発見，” 人工知能誌，vol.16, no.3, pp.316–323, 2001.

[2] 豊田正史，“WWW における関連コミュニティ群の発見，” 情処学研報，vol.2000, no.69, pp.307–314, 2000.

[3] J.J. Rocchio, Jr., “Relevance feedback in information retrieval,” in The SMART Retrieval System: Experiments in Automatic Document Processing, ed. G. Salton, pp.313–323, Prentice-Hall, Englewood Cliffs, NJ, USA, 1971.

[4] 是津耕司，田中浩也，池田新平，KIM Sungyong，田中克己，“Web 上での散策行動を支援する周辺情報提示機構，” 信学技報，DE2003-52, 2003.

[5] 砂山　渡，谷内田正彦，“観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装，” 人工知能誌，vol.17, no.12, pp.14–22, 2002.

[6] 井上晃洋，砂山　渡，谷内田正彦，“多角的な話題の収集を目的とした話題の独自性に基づく Web ページの分類システム，” 人工知能誌，vol.19, no.6, pp.561–570, 2004.

[7] 松本裕治，北本　啓，山下達雄，平野善隆，松田　寛，高岡一馬，浅原正幸，日本語形態素解析システム [茶筌]version.2.3.3，使用説明書，2004.

[8] 杉山公造，グラフ自動描画法とその応用，計測自動制御学会，1993.

[9] 検索エンジン Google:（URL）
http://www.google.co.jp/

（平成 20 年 6 月 9 日受付，10 月 3 日再受付）

砂山　渡

1995 阪大・基礎工・制御卒．1997 同大大学院博士前期課程了．1999 同大学院博士後期課程中退．同年同大学院助手，2003 広島市立大学助教授，2007 同准教授，現在に至る．博士（工学）．人間の創造活動を支援する研究に興味をもつ．人工知能学会，言語処理学会，IEEE 各会員．

鮫島　聡志

2008 広島市大・情報機械システム卒．現在，同大大学院博士前期課程在学中．

西原　陽子　（正員）

2003 阪大・基礎工・システム科学卒．2005 同大大学院基礎工学研究科博士前期課程了．2007 同研究科博士後期課程了．同年より日本学術振興会特別研究員，PhD．2008 東京大学工学系研究科助教，現在に至る．博士（工学）．人のコミュニケーション支援に興味をもつ．人工知能学会，情報処理学会各会員．