2008－SE－159 (13)
2008／3／17

Webページ間の具体抽象関係の視覚化による情報獲得支援

鮫島 聡志† 砂山 渡†

†広島市立大学情報科学部 〒731-3194 広島市安佐南区大塚東 3-4-1

あらまし 本稿では Web ページ間の具体抽象関係を視覚化するインタフェースを提案する. 近年, インターネットの普及により, 個人で取得できる情報量が大幅に増加した. しかし, 取得できる情報量の増加に従い, 必要な情報を探し出すまでに多くの時間や手間が掛かるようなった. また, 必要な情報をリンクを辿って探すこともあるが, リンクだけではどちらのページが詳しくまたは簡潔に書いてあるのかといった情報を得ることは難しい. そこで, Web ページ間の具体抽象関係を定義し, 複数の Web ページをその集合中のユーザが着目する Web ページと比較し, 「具体的に記述された Web ページ」と「抽象的に記述された Web ページ」に分類し, 視覚化するインタフェースを提案する. また, 実験により提案インタフェースの有用性を確認した.

Information acquisition support by the visualization of concrete and abstract relations between Web pages

Satoshi SAMESHIMA† and Wataru SUNAYAMA†

† Faculty of Information Sciences, Hiroshima City University, 3-4-1 Ozuka-Higashi, Asa-Minami-ku,Hiroshima, 731-3194 Japan

Abstract I suggest interface to visualize concrete and abstract relations between Web pages by this report. Recently, the volume of information that is able to be acquired privately increased greatly by the spread of the Internet. However, it came to take much time till I found necessary information because information increased. Moreover, we trace the link and look for information occasionally. However, we cannot judge which page is more detailed from a link only. Then, I first compare two or more Web pages with the Web page to which the user is paying attention, classify into "Web page concretely described" and "Web page abstractly described" it, and propose the interface visualized at the end.

1. 序論

近年, インターネットの普及により, 個人で取得できる情報量が大幅に増加した. しかし, 取得できる情報量の増加に従い, 検索エンジンによる検索結果が必ずしも必要な情報とは限らず, 必要な情報を探し出すまでに多くの時間や手間が掛かるようなった. また, 必要な情報を読んでいる Web ページから張られているリンクを辿って探すこともあるが, リンク元の Web ページとリンク先の Web ページ間の関連は明確でないことが多く, リンクだけでは関係のあるページなのか, どちらのページがより詳しくまたは簡単に書いてあるのかといった情報を得ることは難しい.

そこで, 本研究では, Web ページ間の具体抽象関係を定義し, 検索キーワードについて集めた複数の Web ページをその集合中のユーザが着目する Web ページと比較し, 「具体的に記述された Web ページ」と「抽象的に記述された Web ページ」に分類し, 視覚化するインタフェースを提案する. 具体的な

記述と抽象的な記述とは以下のようなものである.

ユーザが着目しているWebページの検索キーワードに関する文章に対し,比較するWebページの検索キーワードに関する文章の名詞の種類数が多いと検索キーワードに関する詳しい内容を書いていると考えられる為,具体的に記述されていると言える.逆に,Webページの検索キーワードに関する文章の名詞の種類数が少ないと検索キーワードに関して簡単な内容しか書いていないと考えられる為,抽象的に記述されていると言える.

以下,本論文ではまず2章では,関連研究としてWebページ間の関係を測る研究について述べ,本研究の位置づけを明らかにする.3章では,本研究の目的と具体抽象の定義について述べる.4章では,本システムの詳細を述べる.5章では,本システムの有効性を検証する実験の詳細を述べる.6章では,実験結果についての考察を述べる.最後に7章で結論を述べて,本論文を締めくくる.

2. 関連研究

まず,Webページ間の関連を視覚化する研究として,2つのWebページのURLをキーワードとして得られた検索結果からリンクの共起割合を求め,Webページ間の関連性の強さを求める手法[1]がある.この手法により,複数のWebページを関連性の強さから関連の強いWebページ,関連の薄いWebページ,無関係なWebページといった具合に分類することができ,情報獲得に役立つと考えられる.また,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[2]やユーザが着目しているWebページからリンクをたどり,上位N個(任意)のWebページを関連のあるWebページとして視覚化する手法[3]がある.これらの手法により,検索キーワードが使われているWebページだけでなく,それらのWebページと関連のあるWebページも知ることができる為,情報獲得支援が期待される.

次に,Webページの関連を測る研究として,2つのWebページのURLをキーワードとして検索を行い,得られた検索結果全てのWebページからリンクが張られているWebページを関連のあるWebページとしてまとめる手法[4]がある.この手法により,2つのWebページ間の境界となるWebページを見つけることができる.

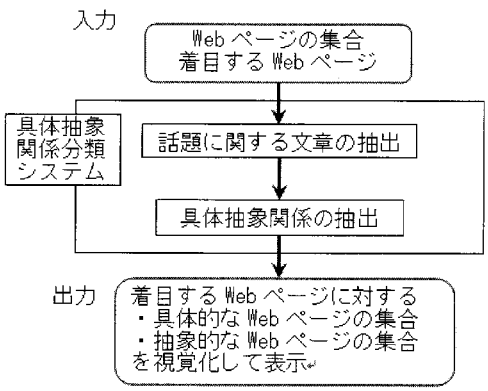

図1. システムの全体構成

最後に,Webページ間の具体抽象関係を求める研究として,Webページ間の単語の共起性から関係を求める手法[5]がある.この手法によりWebページ間でどちらがより詳しいか,どちらがより簡潔かといったことが分かり,情報獲得支援が期待されるが視覚化までは行われていない.

これらの研究に対し,本研究では,まずWebページから検索キーワードに関連のある文章を抜き出し,この抽出した文章をもとに具体抽象関係の分類を行う.次に分類結果をもとに,Webページ間の関係の視覚化を行う。これにより,ユーザは必要な情報の度合いに応じたWebページを見つけるができると考えられる.

3. 具体抽象関係の分類システム

本章ではWebページ間の具体抽象関係の分類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.入力として検索キーワードについて集めたWebページの集合とその集合中のユーザが着目するWebページを与える.着目するWebペー

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

図 2. ブロックレベル要素

表 1. Distance の値

区切りの種類	セグメント間の距離
句点	1
ブロックレベル要素	10

ジと集合中の個々の Web ページの関係を抽出し，出力として具体抽象に分類された Web ページの集合を視覚化して表示する．以下に各モジュールについての説明を述べる．

3.2. システムへの入力

本システムは入力として，ユーザが任意のキーワードで検索を行い，その検索結果から集めた Web ページの集合と集合中のユーザが着目する Web ページを与える．これはユーザが任意のキーワードで検索を行った際に集めた Web ページの集合である．

3.3. 話題に関する文章の抽出

集合中の各 Web ページに対し(ユーザが着目する Web ページを含む)，検索キーワードと繋がりの深い文をもとに各 Web ページから文章を抜き出す．

まず，各 Web ページから本文とブロックレベル要素を抜き出す．ブロックレベル要素は図 2 のようなものである．

次に，本文とブロックレベル要素を抜き出した各 Web ページから展望台システムという検索キーワードと関連のある名詞を抽出するシステム[6]を用いて，関連のある名詞 5 つとこれらの名詞の評価値を求める．

最後に，関連のある名詞の評価値をもとに，話題に関する文章抽出の方法[7]を用いて，関連のある名詞 5 つの評価値から各文に評価値を付け，評価値が閾値以上の最初の文から閾値以上の最後の文までの連続した文を抜き出す．閾値は具体抽象関係の分類の際に用いる閾値との組み合わせから

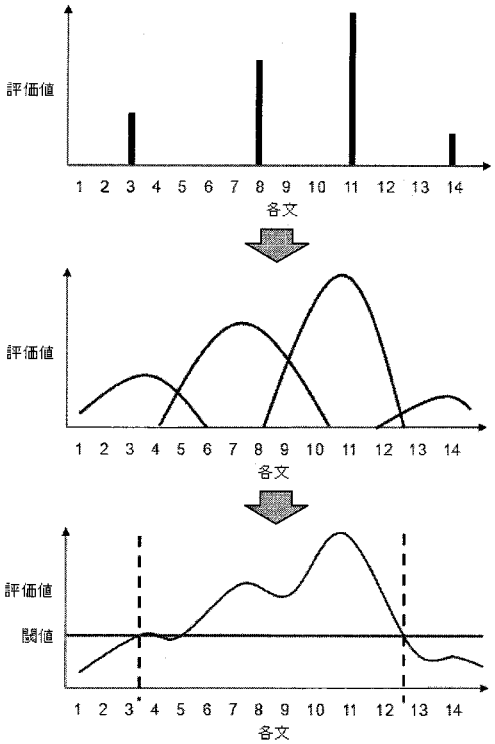

図 3. 各文につく評価値からの文章の抽出

30 とした．閾値の決定方法については 3.5 で述べる．

各文に評価値を付ける際に用いた式は(1)～(3)のようなものである．I は検索キーワードに関連する名詞，i, j は文の行番号，Si は各文，D は文章全体である．また，$valueI$ は，展望台システムから検索キーワードに関連する名詞につく評価値，$Base$ は各文が含む検索キーワードに関連する名詞の評価値からその文につく評価値，$Effect$ は周りの文が含む検索キーワードに関連する名詞の評価値からその文につく評価値である．Distance の値はセグメント(文)間の距離である．表 1 に，Distance の値を示す．各式の係数は，おおよそ文章全体に評価値がいきわたるように決定した．また，評価値のつき方を図 3 に示す．まず，各文に評価値がつき，次にその評価値を全体にいきわたらせ，最後に閾値以上の評価値を持つ最初の文から評価値以上の最後の文までの連続した文を抽出している．

$$Base(l) = 100 * valueL(l) \quad (1)$$

$$Effect(l) = 80 * valueL(l) - Distance_{i,j} \quad (2)$$

ただし Effect(l)<0 のとき Effect(l)=0

$$valueS(S_i) = \sum_{l \in S_i} Base(l) + \sum_{S_{i,j \neq i} \in D} \sum_{l \in S_j} Effect(l) \quad (3)$$

3.4. 具体抽象関係の分類

ユーザが着目する Web ページと集合中の各 Web ページを比較し，具体的な Web ページと抽象的な Web ページに分類する．

各 Web ページの検索キーワードに関連のある文章から形態素解析器 ChaSen[8]を用いて名詞を抽出し，文章中の名詞の種類数の差が閾値 A より多ければ具体的なページ，逆に名詞の種類数の差が閾値 B より少なければ抽象的なページに分類する．閾値 B は閾値 A の符号を反転した値である．本システムで抽出する名詞の種類を表 2 に示す．

3.5. 閾値の決定

話題に関する文章の抽出，具体抽象関係の分類で使用した閾値は各閾値の組み合わせにおいて，適合率が高くなる時の組み合わせを各閾値とした．この際，7 つのテーマを用意し，各テーマを検索キーワードとした検索を行い，検索結果から集めた 9 つの Web ページを用いた．用いたテーマを表 2 に示す．

話題に関する文章の抽出を行う際の閾値を 0%～100%，具体抽象関係の分類を行う際の閾値 A を 0～100 として，全ての組み合わせにおける F 値を求め，どの値が適切であるか調べた．その時の結果から，文章の抽出を行う際の閾値を 20%～60%，具体抽象関係の分類を行う際の閾値 A を 50～100 と範囲を絞り，それぞれの組み合わせで適合率と具体抽象関係の分類個数を求めた．用意したデータの具体抽象関係の組み合わせが 169 組であったため，分類個数が 100 以上且つ適合率が 7 割以上の閾値の組み合わせから，文章の抽出を行う際の閾値を 30，具体抽象関係の分類を行う際の閾値 A を 72 とした．

表 2. 用いたテーマ

Advanced W-ZERO3[es]
イベリア半島
ikonta
さよなら絶望先生
SoundHorizon
lexus
atlantis

表 3. Web ページの色による分類

Web ページ	色
ユーザが着目している Web ページ	緑色
具体的な Web ページの上位 10 件	黄色
抽象的な Web ページの上位 10 件	水色
残りの Web ページ	白色

3.6. 分類の視覚化

図 4 にインタフェースの出力画面を示す．

分類結果から表 3 のような色でそれぞれの Web ページを表示する．本稿では，色ではなく具体的な Web ページの上位 10 件を四角で囲み，抽象的な Web ページの上位 10 件にアンダーラインを引いている．また，ユーザが着目している Web ページは中心に表示している．

Web ページの表示にはタイトルの先頭 6 文字を用いている．また，各 Web ページは各 Web ページ間の名詞の種類数の差をもとにバネモデルのアルゴリズム[9]により配置している．

Web ページは◇印にカーソルを乗せると選択され，赤色で表示される．Web ページ（◇印）を右クリックするとブラウザで開くことができ，ダブルクリックで，選択中の Web ページを着目している Web ページと変更することができる．また，Web ページ(◇印)にカーソルを乗せることで，その Web ページの内容を表す語(展望台システム[6]で求めた検索キーワードと関連する名詞)を表示する．

3.7. 想定するインタフェースの使用方法

本システムは，検索キーワードについて詳しく知りたい場合や簡単に把握したい場合での使用や今見ている Web ページより詳しく書いてあるページや簡単に書いてある

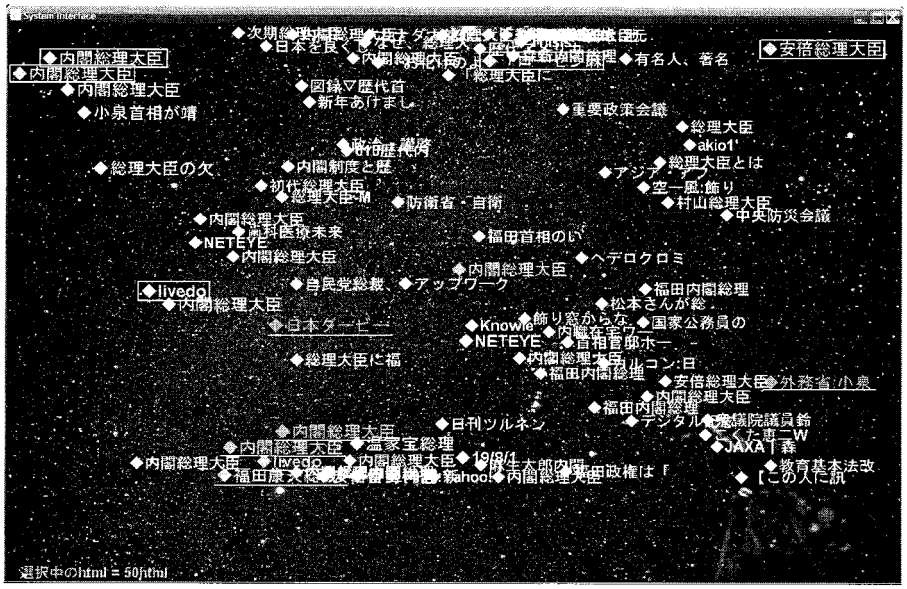

図 4. インタフェースの出力画面

ページを探すときに使用することを想定している．詳しく知りたい場合は具体的な Web ページを探すことで見つけることができ，簡単に把握したい場合は抽象的な Web ページを探すことで見つけることができる．

例えば，総理大臣に関して歴史や職務について書かれている Web ページに着目した場合，インタフェースを使用し，黄色で表示されている Web ページを開くと歴史や職務，権限や呼称について書かれた Web ページが得られ，着目している Web ページより詳しく書かれたページを見つけることができる．逆に水色で表示されている Web ページを開くと職務について書かれた Web ページが得られ，着目しているページより簡単に書かれているページを見つけることができる．

4. 評価実験

4.1. 実験目的

本実験では，検索キーワードについて集めた Web ページ集合に対してユーザが着目している Web ページと集合中の個々の Web ページ間の関係を抽出した上で，提案インタフェースがどの程度情報獲得に影響するか，また，このシステムの有効性について検証する．

表 4. 実験に用いたテーマ

総理大臣
サッカー
紅白歌合戦
センター試験
初音ミク

4.2. 実験内容

本実験では，5 種類のテーマを用意した．また，被験者を大学生及び大学院生 12 名とした．被験者には，テーマ毎に着目している Web ページ(本実験では筆者の 1 人がインタフェース上で具体抽象が同程度出力されるものを無作為に選んだ)より詳しく書かれている Web ページと簡潔に書かれている Web ページを提案インタフェースおよび比較インタフェースを用いて，Web ページ集合から探してもらった．

比較インタフェースは提案インタフェースから具体抽象関係が分かる色を取り除いたものである．

5 つのテーマの Web ページ集合は検索サイト google[10]において表 4 の各々を検索キーワードとして入力し取得を行った．5 つのテーマと 2 つのインタフェースを用い

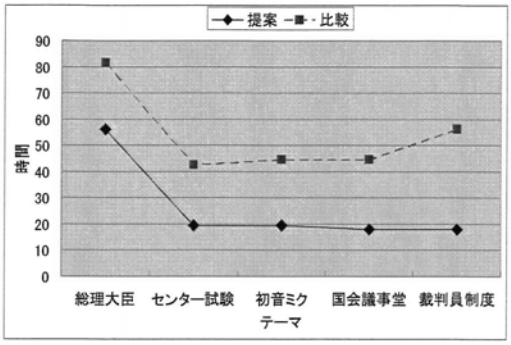

図 5. 詳しいページを見つけるまでに掛かった時間(中央値)

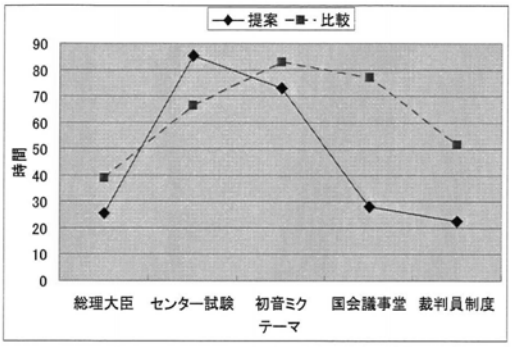

図 7. 簡潔なページを見つけるまでに掛かった時間(中央値)

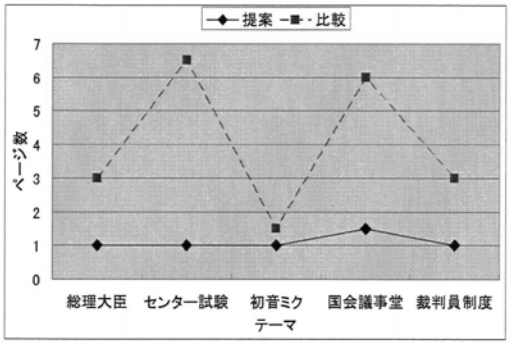

図 6. 詳しいページを見つけるまでに開いたページ数(中央値)

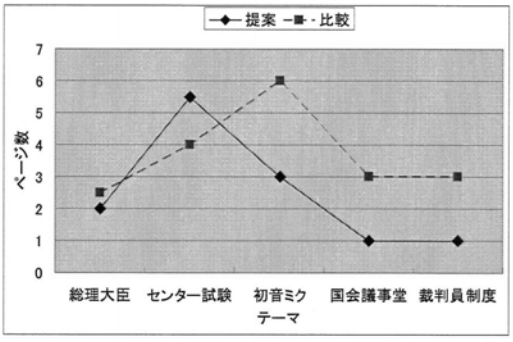

図 8. 簡潔なページを見つけるまでに開いたページ数(中央値)

て,最初に着目しているWebページより詳しく書かれているWebページおよび簡潔に書かれている Web ページを探してもらい,探し出すまでに掛かった時間とブラウザで開いたWebページ数を調べた.1つのインタフェースにつきテーマ毎に探してもらったWebページはより詳しく書かれているページか簡潔に書かれているページのどちらか一方である.

4.3. 実験結果

各テーマにおける詳しい Web ページを見つけるまでに掛かった時間の中央値を図 5 に,ブラウザで開いたページ数の中央値を図 6 に示す.また,簡潔な Web ページを見つけるまでに掛かった時間の中央値を図 7 に,ブラウザで開いた Web ページ数の中央値を図 8 に示す.

5. 考察

図 5〜8 から全体的に提案インタフェースの方が掛かった時間はおよそ半分になり,開いたページ数も少なくなっていることが分かる.これは,提案インタフェースでは着目しているWebページより具体的なWebページや抽象的なWebページに色が付いているため見るページを絞り込むことができたためであると考えられる.

しかし,「センター試験」については図 7,図 8 のように簡潔なページを見つける際に,提案インタフェースの方が掛かった時間が長く,開いた Web ページ数も多くなっている.これは,センター試験について集めた Web ページのほとんどがセンター試験に関するニュース記事やブログの記事であり,これらの記事が短い文章であったため名詞の種類数が少なく,抽象的な Web ページに分類されてしまいセンター試験について説

明しているページが見つかり難かったことが原因であると考えられる.

6. 結論

本研究では, Web ページ間の具体抽象関係を視覚化するインタフェースを作成した. 今後の課題として,

- Web ページ集合を検索キーワードから自動で取得を行う
- ばらつき具合の修正を行う

などが挙げられる為,これらの修正によりシステムの改善を行なっていきたい.

参考文献

[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,人工知能学会論文誌, Vol.16, No.3, pp.316-323, 2001

[2]福地健太郎・田中活也・池田新平・金星鏞・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, 電子情報通信学会技術研究報告.DE, Vol.103, No.191, pp.121-126, 2003

[3]豊田正史, WWW における関連コミュニティ群の発見, 情報処理学会研究報告, Vol.2000, No.69, pp307-314, 2000

[4]村田剛志, Web コミュニティの中心性, 人工知能学会全国大会(第 17 回)論文集 CD-ROM, 2003

[5]冨山祐樹・砂山渡, 複数の Web ページの単語の出現頻度による具体・抽象への分類, 電子情報通信学会技術研究報告.KBSE, Vol.105, No.652, pp.31-36, 2005

[6]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, 人工知能学会論文誌, Vol.17, No.12, pp.14-22, 2002

[7]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, 人工知能学会論文誌, Vol.19, No.6, pp.561-570, 2004

[8]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

[9]杉山公造, グラフ自動描画法とその応用, 計測自動制御, 1993

[10]「Google 検索」, http://www.google.co.jp/