比喩的素描手法精緻化に向けた *descriptor* 分類手法の検討

渡邊雄太* 桝井文人 Ptaszynski Michal 新田大征
(北見工業大学)†

1 はじめに

新しい語やうろ覚えの語の正確な表記，意味，特徴などを調べるために WWW 検索を用いることが多くなっている．しかしながら，WWW 検索は「ユーザの求める知識」への経路が最適化されていないため，必要な情報を得ることが困難な場合がある．このような問題に対して，桝井ら [1, 2] は WWW から情報要求 (*query*) を素描する比喩的素描表現 (*descriptor*) を抽出し，*query* 語のイメージ知識である *descriptor* を提示するシステム「Murasaki」を構築している (Fig.1)．

Fig. 1 Murasaki の出力画面

Murasaki は，*query* 語を入力し取得した *descriptor* の一覧を提示する．そこに *query* 語と *descriptor* の AND 検索結果のリンクが貼られている．これは，単語だけでは理解しきれない *query* 語と *descriptor* の関係を調べるためである．例えば，*query* 語「初音ミク」において *descriptor*「無理」は，*query* 語に対して何が無理であるかという点が曖昧であり，それだけでは関係を理解できない．しかしながら，ユーザが全 *descriptor* の検索結果を確認するには膨大なコストがかかる．また，*descriptor* が異なっていても，検索結果として同じ文書が表示される場合があるため，ユーザには重複ページを確認するコストもかかる．

そこで，取得した *descriptor* を分類したり，検索結果を分類するなどして，その結果をユーザに提示することができれば，ユーザは自身が興味のある知識を効率よく参照できるようになり，かつ不必要な情報の確認コストを削減することも可能となる．

本研究では，*query* 語と *descriptor* の AND 検索結果を分類したのち，複数文書要約を施すことで用例を提示し，ユーザにかかる検索結果を確認するコストの削減とユーザの *query* 語理解促進支援を目指す．例えば，*query* 語「初音ミク」，*descriptor*「vocaloid」であれば,「初音ミクはボーカル・アンドロイド=vocaloid である」のように用例を提示することでユーザの検索結果を確認するコストを抑えることが可能になり，*query* 語の理解促進に繋がると考えられる．

本稿では，文書要約効率化のために *descriptor* を分類する手法について検討する．具体的には，*descriptor* の分類手法として EDR 概念辞書 [3] を利用した分類，URL 情報を利用した分類の 2 種類の手法の調査および取得した文書に対して LSA（潜在的意味解析）を施した場合の有効性について議論する．

2 基本的なアイデア

前章で述べた通り，我々は *descriptor* 集合を分類し，これらを介して複数文書要約を行うことを目指している．複数文書要約は，1 つの文書を短く要約するのではなく，複数の文書の特徴や関連を判断して要約を行うため，文書の内容があまりにかけ離れていると適切な要約が困難となる．そのため，文書の内容，文章の特徴などにより文書を分類し，その分類結果を利用して複数文書要約を施す必要がある．しかし，*query* 語と取得した *descriptor* のみを使った分類では適切な分類は難しいことが報告されている [1]．

この問題を解決するアイデアとして，*query* 語と *descriptor* の AND 検索で取得した文書を解析し，そのトピックによって分類する方法が挙げられる．本稿で扱うトピックは，対象とする文章が何を取り上げているか，何について論じているかによって決まる．よって，取得文書をトピックによって分類できれば，*descriptor* の分類も可能になるはずである（Fig.2）．

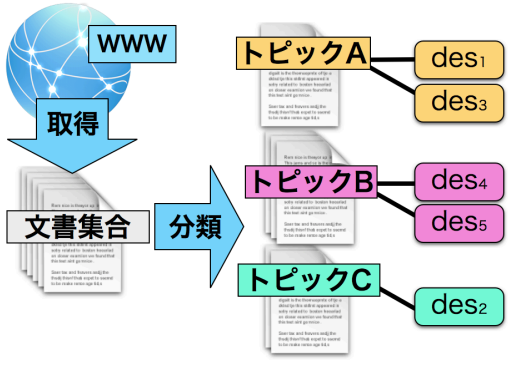

Fig. 2 トピックによる分類と *descriptor* 分類

* yw167@ialab.cs.kitami-it.ac.jp
† 北海道北見市公園町 165 番地

トピックを扱う際にはいくつかの問題が生じる．例えば，同義語の異表記のような表層的には異なるが同一概念を示す語や，それとは逆に表層的には同一のようでも概念が異なる語の扱いの問題がある．表層的に異なるが同一の概念である場合，それら *descriptor* は同様のトピックが書かれた文書に出現する可能性が高い．この問題には，概念辞書を利用した概念一致判断あるいは文脈を考慮した分類が効果的であると考えられる．

また，表層的には同一でも概念自体が異なる語は区別して扱う必要があり，取得文書のトピックにも違いが見られるはずである．

さらに，書き手の違いによって，書かれる文章の特徴が異なるという可能性がある．例えば，個人サイトで書かれる文章と企業や公的機関のサイトで書かれる文章は大きく異なる．個人サイトで書かれる文章では口語表現を用いたり，感情表現として顔文字を利用するなど自由度が非常に高いが，企業や公的機関のサイトではこういった表現は使われない．よって，書き手の違いに基づいて，文書の内容を判断できると考えられる．例えば，個人サイトやブログなどには主観的意見が書かれやすく，公的機関やニュースサイトなどには事実が明確に書かれるはずである．

書き手の違いを判断する手がかりとして URL 情報を利用する手法が挙げられる．URL 情報を用いた分類が可能であれば，書き手による内容の違いや表現の違いを判断できるはずである．

3　EDR 概念辞書を利用した分類

本章では，EDR 概念辞書を利用した分類手法を説明する．

3.1　仮説

descriptor の類似性を判断するため EDR 概念辞書を利用した *descriptor* の分類を試みた．EDR 概念辞書には，単語の上位下位関係や多義性を持つ単語の概念を識別する概念識別子などが登録されている．概念識別子を利用することで，違う単語だが概念が一致するものや，「ほっぺ」と「ほっぺた」のような表記のゆらぎを丸め込むことができる [4]．同一概念と判断された *descriptor* は，同様のトピックに出現する可能性が高く，分類の手がかりとなるはずである．

例えば,「ゲーム」という単語には概念識別子が「0efb5e:0efb5f:0efb60:3d05f5」のように登録されている．これは「ゲーム」という単語が複数の概念を持っていることを示し，それぞれで使い方が異なる．この中の「0efb5e」という概念識別子であれば,「ふたり以上で勝ち負けを争うあそび」という概念が表示される．

次に,「競技」という単語を EDR 概念辞書で調べてみると「3d05f5」という概念識別子が得られる．この概念識別子は「ゲーム」の概念識別子としても登録されているため,「ゲーム」と「競技」は同一の概念が存在するということである．このような特性を利用して *descriptor* の類似性を判断する．

3.2　調査

調査で使用する *descriptor* は *query* 語「PS3」で取得された 41 個,「カーリング」で取得された 28 個,「リンゴ」で取得された 46 個である．調査方法は，*query* 語ごとの *descriptor* の概念識別子を取得し，*descriptor* 同士での概念識別子の一致数を数える．

調査結果を Table1 に示す．調査の結果，*descriptor* 同士で概念識別子が一致していたのは *query* 語「カーリング」の *descriptor*「ゲーム」,「競技」,「スポーツ」の 3 個と *query* 語「リンゴ」の *descriptor*「ほっぺ」,「ほっぺた」の 2 個のみであり，*query* 語「PS3」では概念識別子が一致している *descriptor* はなかった．

また，*descriptor* には単語でなく名詞句となるものがあり，それらのほとんどが EDR 概念辞書には登録されていなかった（Fig.3）．

Table 1　概念識別子が一致した *descriptor*

query	*descriptor*
カーリング	ゲーム，競技，スポーツ
リンゴ	ほっぺ，ほっぺた
PS3	該当なし

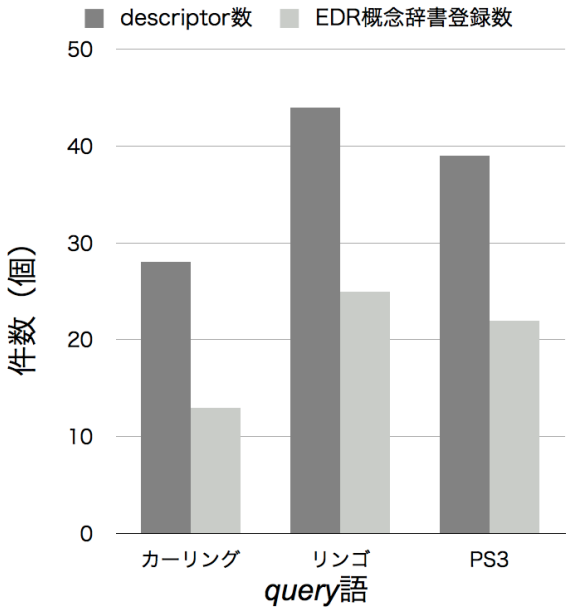

Fig. 3　*descriptor* 数と EDR 概念辞書登録数の比較

上で述べた調査結果の考察を行う．概念識別子の一致が *query* 語「カーリング」で 3 個,「リンゴ」で 2 個という結果から，EDR 概念辞書を利用した類似性判断は効果的ではなかった．原因としてまず，EDR 概念辞書に登録

されているのは厳密な表記の単語のみであり，曖昧性が許されないことが考えられる．例えば，*query* 語「カーリング」の *descriptor* として「きょうぎ」が取得されていた．これは「競技」の異表記であり類似語と判断できるが，EDR 概念辞書には登録されていないため対応できない．

また,「スポーツ」,「マイナー」,「ニュー」などの単語は単体では登録があっても，これらを組み合わせた「マイナースポーツ」,「ニュースポーツ」などは登録されていなかった．Murasaki は，*descriptor* として名詞句や複合語も取得できるという特徴がある．しかし，それらは EDR 概念辞書にはほとんど登録されていないため，この特徴を活かすことができなかった．例として，*query* 語「カーリング」では 53.5%（28 個中 15 個）,「PS3」では 43.5%（39 個中 17 個）,「リンゴ」では 43.1%（44 個中 19 個）の *descriptor* が EDR 概念辞書に登録がなかった．

以上のような結果から，Murasaki が取得した *descriptor* の分類に EDR 概念辞書を活用するには更なる工夫が必要である．具体的な工夫の一つとして，複合語に対応できれば，*query* 語「カーリング」の「ニュースポーツ」や「マイナースポーツ」,「氷上競技」といったカーリングを特徴づけるような *descriptor* を分類対象にできる．しかし，ひらがなと漢字での異表記は場合によっては取得文書のトピックに違いが出る可能性もあるので注意が必要である．

4　URL 情報を利用した分類

本章では URL 情報を利用した分類手法を説明する．

4.1　仮説

URL 情報によって *query* 語に関する知識がどのようなサイトに記述されているかを判断できると考えられる．これにより，知識源の文書ごとに分類が可能となる．

そこで，*query* 語と *descriptor* の AND 検索により取得された文書の URL 情報を利用した分類手法について調査した．

例えば「blog」が URL 情報に含まれていた場合，そのページは weblog である可能性が高く，主観的な意見が書かれやすいと推測できる．また,「youtube」が含まれている場合は動画サイトであると考えられるため動画として分類することが可能である．

4.2　調査

調査には *query* 語 { カーリング，リンゴ，PS3} を使用した．各 *query* 語ごとに *descriptor* 上位 20 個，AND 検索結果は上位 100 件，合計 2000 件を使用した．調査文字列は「blog」,「news」,「nicovideo」,「wikipedia」,「youtube」の 5 つである．

調査結果を Fig.4 に示す．このグラフは各 *query* 語の取得文書において，今回調査した 5 つの文字列を URL 情報を含む文書数を表している．

全ての *query* 語で「blog」を URL 情報として含む文書が多かった.「カーリング」では,「news」,「wikipedia」,他の 2 つでは「news」,「youtube」が多かった．各 *query* 語を比較すると,「blog」,「news」,「nicovideo」では *query* 語「PS3」が最も多く,「wikipedia」では「カーリング」,「youtube」では「リンゴ」が最も多かった．

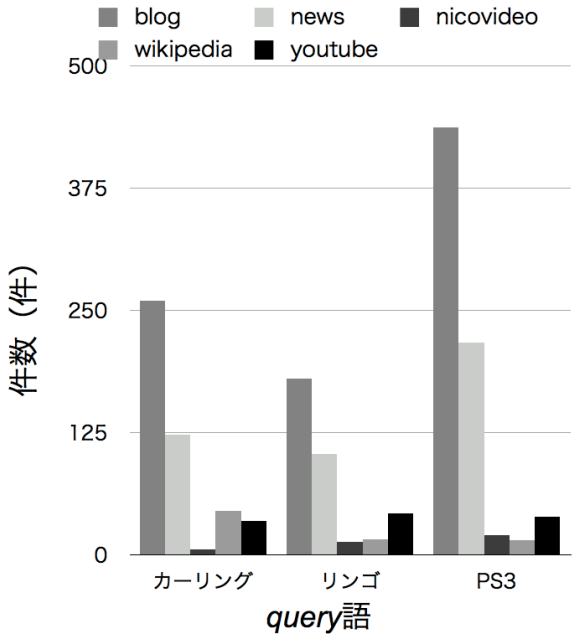

Fig. 4　取得文書の URL 情報における文字列調査結果

上で述べた調査結果の考察を行う．URL 情報による分類は，調査対象の 5 つの文字列では,「news」,「nicovideo」以外の文字列を含むものは分類へ利用可能と考えられる．

文字列「news」を含む文書を調査した結果，*query* 語に関するニュースや新情報が取得できた．実際に意図したウェブサイトは，新聞社や「google ニュース」,「yahoo! ニュース」といった時事を取り扱うサイトであったが，これら以外にゲームの新作情報を紹介するブログや，掲示板のまとめサイト[1] なども多く取得されていた．特に *query* 語「PS3」ではゲームや周辺機器の新作情報や使用した感想などを紹介するブログが多く取得されていた．これらは，本来の意図とは違っているが「news トピック」としては正しいといえる．

また，URL 情報からは各新聞社やニュースサイトを判断できるような文字列は発見できなかった．実際に使用されている文字列を調査し辞書登録して利用できれば,「新聞記事」や「ニュースサイト」という分類も可能になるであろう．

「nicovideo」というフレーズは,「ニコニコ動画」以外にも「ニコニコ大百科」や「ニコニコ静画」といった関連したウェブサイトにも使用されるため，単体では判

[1] 2 ちゃんねるなどの掲示板のスレッドを読みやすくまとめ，それらを記事として紹介するブログやサイトのこと．

別できない．今回の調査では，「nicovideo」を URL 情報に含む文書全 39 件中，10 件は動画以外のページであった．この問題には，URL 情報をより詳細に利用することで解決できると考えられる．例えば，「ニコニコ大百科」であれば「dic.nicovideo」，「ニコニコ静画」であれば「seiga.nicovideo」のように利用文字列を拡張することで対応できる．

文字列「blog」，「youtube」，「wikipedia」では，それぞれの取得文書が「主観意見」，「動画」，「wikipedia による定義文」として分類が可能であろう．

5　LSA による分析

本章では，*descriptor* が記述されている web ページを利用し，LSA による分析を行う．

5.1　仮説

LSA（潜在的意味解析）[5, 6, 7] は，文書に出現する単語に対し統計的計算を施すことで語の潜在的な意味を考慮する手法である．語の潜在的な意味とは文書の表層からは判断できない単語の意味である．例えば，「VOCALOID の曲」と「VOCALOID の技術」では，同じ語でも前者はソフトウェアを指し，後者は音声合成技術を指す．このように同一単語でも異なる意味になってしまう．

この手法ではまず，文書中の出現単語の頻度情報を利用して文書と単語の行列を作成する．しかし，このままでは文書情報が高次元で表現されるため扱いにくいので，特異値分解により次元圧縮を行う．ここでは 3 次元ベクトル空間として表現するため，3 次元に圧縮する．圧縮された行列は元の文書単語行列の射影であり，文書内容の意味的特徴を数値化した特徴量が近似している．これにより，単語の潜在的な意味，関係を推論することが可能になる．

例えば，*query* 語「初音ミク」における文書で，「ボーカロイド」と「vocaloid」，「ボカロ」はほとんどの文書で意味的には同じように使用されている．しかし，処理を行う際には別物として認識されてしまうため，単純な頻度情報などにより処理を行うと，実際の文書トピックとはかけ離れた結果が出力される恐れがある．これに対し，LSA では上記 3 語を同一の意味を持つ概念として扱い，処理できるようになる．その結果として，文脈的には関連していながら表層的な判断では分類できなかった文書の分類が期待できる．

5.2　分析

分析には *query* 語 { 初音ミク，twitter，リンゴ，カーリング } とそれぞれの *descriptor*15 個による AND 検索での取得文書全 3,000 件を使用した．

上記検索結果に LSA を適用した後，解析結果として得られた特徴量を 3 次元空間にベクトルとして視覚化し，分析を行った．Fig.5 は *query* 語「リンゴ」の特徴量を 3 次元ベクトル空間で視覚化した結果である．ここからは，*query* 語「リンゴ」を例に説明する．まず，大きく遠隔している *descriptor* が幾つか見られた．例えば，*descriptor*「実」や「丸み」である．また，近接している *descriptor* には，共通した特徴が見て取れる．例えば，*descriptor*「酸味」，「甘み」，「食感」といったものはリンゴの食べ物としての共通した特徴である．

しかし，*descriptor*「味」のように共通した特徴であるにも関わらず，遠隔しているものも存在した．他にも，「爽やか」は漢字表記とひらがな表記で同じ *descriptor* が 2 つ取得されたが，それぞれの出現位置はあまり近接していなかった．

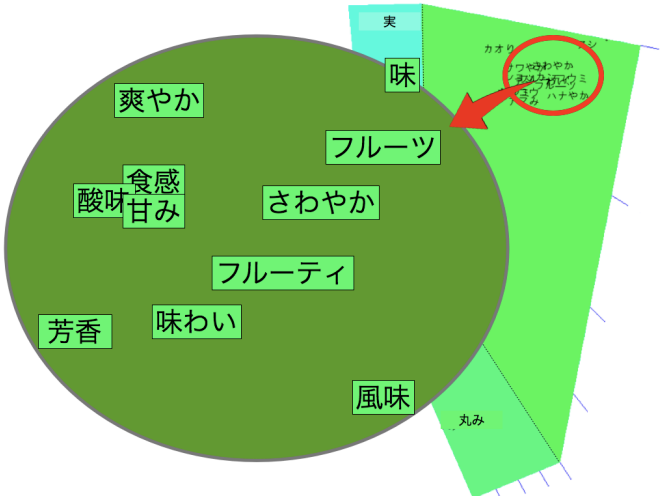

Fig. 5　*query* 語「リンゴ」の分析結果

上で述べた分析結果の考察を行う．全体としては，表層上の文字列は異なるが意味的に近い *descriptor* は近接し，表層上は同一でも意味的に遠い *descriptor* は互いに遠隔した．

各 *query* 語の分析結果において，遠隔した *descriptor* は，取得文書のトピックがその他の *descriptor* とは類似していないと考えられる．例えば，*query* 語「リンゴ」で，近接した *descriptor* とそれらから遠隔したものを見てみると，近接した *descriptor* は「甘み」，「食感」，「酸味」などであり，遠隔した *descriptor* は「実」，「丸み」であった．まず，「甘み」，「酸味」，「食感」などの特徴は，リンゴの食べ物としての共通した特徴といえる．

これに対し，「実」，「丸み」などはリンゴの見た目や果物としての特徴であり，このような *descriptor* が現れる文書には，食べ物としての特徴は少なかった．例えば，*descriptor*「実」では，りんごの栽培に関する文書が取得されており，栽培方法や，栽培の際に気をつけることなどが主な内容であった．「甘み」，「酸味」，「食感」は，類似した文書が取得されていると考えられ，それらが近接しているということは，トピック分類できているといえる．

しかし,「甘み」や「食感」といった *descriptor* と近いと考えられる *descriptor*「味」は比較的離れた位置に現れている．取得した文書内容を見ると，他の食品が「リンゴ味である」という表現が多い．文書のトピックが「リンゴ自体の味」ではなく,「他食品のリンゴ味」であるため，上記 *descriptor* とは遠隔していると考えられる．

descriptor「爽やか」と「さわやか」は異表記である．それぞれで取得された文書を見ると，ひらがな表記では「信州高山さわやかりんご」や「さわやかはちみつりんご水」のような固有名詞が取得されている．一方，漢字表記では「自家製のヨーグルトをミックスしました。爽やかな青リンゴ風味です」のような，紹介文中などに出現していた．このような傾向から，*query* 語「リンゴ」では *descriptor*「爽やか」,「さわやか」は近接しないという結果になったと考えられる．

これらの分析結果より，各文書のトピックが重要視されているといえ，LSA による調査・分析は有効であったと考えられる．

6　考察

本章では前章までの調査結果に対する考察を行う．

まず，EDR 概念辞書の概念識別子を利用して概念の一致を判断できると考えたが，実際には複合語，名詞句といった単語以外の *descriptor* も多く，活用には更なる工夫が必要であることが明らかとなった．この問題を解決するには，複数形態素と EDR 概念辞書に登録されている単語の照合が必要である．例えば，岩城ら [8] による上位下位関係精緻化の手法を取り入れることで複合語へ対応が可能になると考えられる．

URL 情報による書き手や情報源の違いを考慮した分類では，文字列「blog」や「youtube」などは想定した分類が可能であった．また，より詳細な URL 情報を調査したり，新聞社などの URL 情報を辞書登録するといった工夫をすればさらに多くの文書が分類可能と考えられる．

LSA による分析では，表層上の文字列が異なる場合でも意味的に近い *descriptor* は近接し，表層上の文字列が同一でも意味的に遠い *descriptor* は遠隔した．よって，*descriptor* 同士の関連を反映できたといえる．

しかし，今回の LSA による分析では *descriptor* ごとに分類を行ったが，実際に取得された文書内容を見ると *descriptor* によってはトピックが一意に決まらないということもある．例えば，*query* 語「初音ミク」では *descriptor*「無理」や「ツール」といったものが挙げられる．これらは *descriptor* のみでは *query* 語との関係が理解しづらい上，取得された文書集合を見ても複数の話題に対しての文書が現れる．今回の分析ではこのようなマルチトピックに対しての手立てを何もしていないため，単語の潜在的意味を効果的に利用できていない可能性がある．

この問題は，*descriptor* ごとに LSA を適用するのではなく，例えば URL 情報によって分類した『主観意見』や『定義文』といった集合に適用したり，タイトルに含まれる *query* 語と *descriptor* の情報を利用するなどといった対策が必要である．

URL 情報を利用した分類を施したのち，LSA を利用した分類を行ったり，LSA によって複合語をある程度丸め込み，その後 EDR 概念辞書を活用するなど，今回調査した手法を組み合わせることでさらなる分類精度向上に繋がるはずである．

また，新たな *descriptor* 分類の手法として拡張 PMI-IR 手法を応用することで，web 上での出現傾向などを考慮した分類が可能と考えられる [9]．この手法は，本来掲示板での有害書き込みを発見するための手法であるが，使用する単語やフレーズを *descriptor* や *query* 語に置き換えることで *descriptor* の連想度を算出できると考えている．

7　おわりに

今回，比喩的素描手法により取得した *descriptor* 分類のため，EDR 概念辞書を利用した分類手法と *descriptor* の URL 情報を利用した分類手法の調査，LSA を利用した分析を行い，その結果から *descriptor* 分類手法の検討を行った．

EDR 概念辞書を利用した分類では，類似した *descriptor* は少なく，また複合語や名詞句に対応できないため EDR 概念辞書の利用には工夫が必要であることが明らかとなった．

次に，*descriptor* が記述されている web ページの URL 情報を利用した調査では，調査した 5 個の文字列のうち，3 個は分類に利用が可能であり，より詳細な調査によってさらに利用できる可能性があることを示した．

LSA による分析では，近接した *descriptor* には特徴の類似が見られた．また，大きく遠隔した *descriptor* にはトピックの違いが見られたため，LSA が有効な手法であることが示唆された．

今後の展望として，複数手法の組み合わせ方や各手法の課題点について調査を進めるとともに，拡張 PMI-IR 手法を用いた調査・分析を行う予定である．

参考文献

[1] 桝井文人, ジェプカ・ラファウ, 木村泰知, 福本淳一, 荒木健治:“WWW 活用による語の比喩的素描手法”, 知能と情報, Vol.22, No.6, pp.707-719(2010).

[2] Fumito MASUI, Rafal RZEPKA, Yasutomo KIMURA, Junichi FUKUMOTO and Kenji ARAKI: “ Acquisition of Japanese Word Descriptions from World Wide Web”, In Proceedings of IWMST201,pp.153-158(2010).

[3] 日本語電子化辞書研究所, EDR 概念辞書, 日本電子化辞書研究所, 1995.

[4] 梶川健, 桝井文人, 河合敦夫, 井須尚紀:“電子化辞書を利用した概念特徴の自動分類 (福祉と知能・認知障害/一般)”, 電子情報通信学会技術研究報告. WIT, 福祉情報工学 105(508), 31-36, (2006).

[5] Scott Deerwester; Susan T Dumais; George W Furnas; Thomas K Landauer; Richard:“Indexing by Latent Semantic Analysis”, Journal of the American Society for Information Science (1986-1998); Sep 1990; 41, 6; ABI/INFORM Global pg. 391.

[6] 北研二, 津田和彦, 獅々堀正幹:“情報検索アルゴリズム”, 共立出版 (2002).

[7] 石田基広:“R によるテキストマイニング入門”, 森北出版 (2008).

[8] 岩城秀則, 桝井文人:“WWW から獲得した語の比喩的素描表現の上位下位関係精緻化に関する一考察”, 言語処理学会第 18 回年次大会発表論文集, pp.1180-1183(2012).

[9] 新田大征, 桝井文人, Ptaszynski Michal, 木村泰知, Rzepka Rafal, 荒木健治:“カテゴリ別関連度最大化手法に基づく学校非公式サイトの有害書込み検出”, 第 27 回人工知能学会全国大会, インタラクティブセッション 203-9in(2013).