様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 5 月 11 日現在

研究種目：基盤研究(C)
研究期間：2007～2008
課題番号：19500086
研究課題名（和文）パーソナル概念辞書と適合・不適合ウェブページの組入による概念的検索エンジンの実現
研究課題名（英文）Realizing Conceptual Information Retrieval by Conjoining a Web Search Engine with Relevant/Non-Relevant information and Personal Concept Dictionaries

研究代表者
伊藤　哲郎（ITO TETSURO）
大分大学・工学部・教授
研究者番号：30029558

研究成果の概要：
改良した概念的検索機構 Gce で市場の検索エンジンをラップし，高速の探索性を有しながら高い検索効率を出せる概念的検索エンジンを実現する。キーワードベースの検索は，もともと，(i)語彙問題（キーワードは同義性と多義性を有する）と(ii)逆転問題（ユーザの要求は結果を見てはっきりする）を内包している。改良版 Gce では，ウェブディレクトリ辞書に加えて，適合・不適合判断辞書とパーソナル概念辞書を編集して利用し，両問題の解決に当る。実現した概念的検索エンジンンの性能についても記す。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,000,000	300,000	1,300,000
2008 年度	500,000	150,000	650,000
総　計	1,500,000	450,000	1,950,000

研究分野：　総合領域
科研費の分科・細目：情報学，メディア情報学・データベース
キーワード：概念的検索エンジン，ウェブ探索，概念辞書，適合・不適合判断，聞き込み，画像検索

１．研究開始当初の背景

キーワードでの質問を受け付ける検索エンジンの質問処理法は，大きく分け次の２つになる。

タイプ A：質問キーワードが現れるウェブページを取り出す

タイプ B:質問キーワードと概念的に関連したキーワードが現れるウェブページで，ユーザが読みたい内容が記されているものを取り出す

ユーザが暗黙のうちに想定しているタイプは B である。ところが，処理が効率よく進むことを念頭に作られた Google，Yahoo!，MSN などの市場の検索エンジンは，論理検索モデルを採用しており，タイプ A の質問を扱うのは得意であるが，タイプ B についてはそうではない。もちろん若干の対処はなされている。例えば検査エンジン Google では，リンク情報に基づくページランク方式を採用して，他のページから張られているリンクの数が多いものを上位にランクするように

している。しかしながら，望んだ内容が記されているページを複数取り出そうとすると，ユーザは依然として多くの不要なものを参照しなければならなくなる。また，ユーザは検索に先立って質問キーワードをうまく作れない場合があるが，市場の検索エンジンはこれへの対応も十分でない。今後開発する検索エンジンでは，これら欠点の解消が必要である。

２．研究の目的

キーワードでの質問を受け付ける検索には，もともと，(i)語彙問題（ウェブページに質問キーワードと意味は同じでも綴りが違うキーワードや綴りは同じでも意味が違うキーワードが現れている）と(ii)逆転問題（ユーザの要求は結果中のウェブページを参照した時点ではっきりする）が内包されている。上記の欠点もこれらが原因となっている。キーワードをベースにした画像検索でも，同様の問題が生じる。

語彙問題については，当該研究者らが定式化してきた概念的検索機構 Gce を使えば，かなりの程度緩和できる。Gce は，質問のキーワードベクトルがウェブページのキーワードベクトルとどれくらい概念的にマッチしているかを，ウェブディレクトリ概念辞書を参照しながら測り，マッチ度の高さの順にウェブページをランクする。

ここでは，逆転問題も解決できるよう，実用性を重視しながら，Gce を改良する。そして，改良した Gce で市場の検索エンジンをラップすることで，高速の探索性を有しながら高い検索効率を出せる概念的検索エンジンを作り上げる。

３．研究の方法

ウェブディレクトリ辞書に加えて新しく2種の辞書を編集する方法を定式化し Gce で扱えるようにすることで，逆転問題も解決する。2 種のうちの 1 つは，ユーザが質問に対する検索結果のウェブページに下した適合判断ならびに不適合判断状況を，それぞれ，正例と負例のキーワードベクトル集合として逐次学習した，適合・不適合判断辞書である。残りの 1 つは，検索結果のウェブページ中に含まれるキーワードの集まり対して，それらの間の概念的関係を明示した，汎用の概念辞書の一部としてのパーソナル概念辞書である。

(1) 適合・不適合判断辞書の編集

当該研究者らによる研究によれば，適合・不適合判断辞書を使えば，ユーザごとの検索意向が一般ユーザのそれと一致しない場合とか短期的に変わる場合でも，逆転問題にうまく対処できることがわかっている。使い方は，以下の通り。辞書中の正例（質問自体も含む）および負例で入力質問を置換し，正例で置き換えられた質問とのマッチ度が負例で置き換えられた質問とのマッチ度より大きなウェブページで未読のものを，読みたい内容が記されているウェブページの候補としてユーザに示す。そして，ユーザによる適合判断に従い辞書を更新する。ここでは，適合・不適合判断辞書の更新作業を実時間で行う方法を求める。

(2) パーソナル概念辞書の編集

人物検索で同姓同名の人物のウェブページが多数見受けられる場合など，ユーザ自身にとって，初期に入力した質問で検索意向を明確に表現しているつもりでも現実的には曖昧さが残っているような，逆転問題が生じることがある。これには，パーソナル概念辞書を編集して対処する。ここでは，パーソナル概念辞書の効率的な編集方法と，ユーザが簡便に利用できるようにする方法を定式化する。

４．研究成果

語彙問題に対処するため，近年，多くの検索エンジンでは，質問キーワードの同義語を補足したり，関連語の候補をユーザに示して選択してもらったりする方式が採用されるようになってきている。すなわち，キーワードの概念的な扱いに力を入れるようになってきている。逆転問題への対処として，Google は誕生時から，ページランク方式を採用してきている。他には，ユーザごとの検索履歴をユーザプロファイルとして記録しておき，これを参照しながら検索を進める試みもある。

上記の対処法にも欠点がある。語彙問題への対処については，限定された同義語や関連語しか扱えなかったり，ユーザの負担が増えたりする。逆転問題への対処については，特定ユーザの検索意向が一般ユーザのそれと一致しない場合とか，短期的に変わる場合には，それをうまく捕まえることはできない。また，ユーザ自身が検索意向を明確に表現しているつもりでも現実的にはそうなっていない場合にも，対応できていない。

ここでは，語彙問題を緩和できる Gce を利用しながら，そこに新たな 2 種の辞書を組み込み，改良された Gce で市場の検索エンジンをラップすることで，逆転問題を解決する。その際，市場の検索エンジンの高速探索性とページランク情報を生かしながら，かつ，より高い検索効率を出せるよう，実用性に重点を置いた辞書の編集法・利用法を求める。

(1) 適合・不適合判断辞書を利用した適合可能性示唆 [学会発表④，⑥]

適合・不適合判断辞書の基本的な編集方法と使い方は上に述べた。この辞書を，コストをかけずにかつそれを利用した検索時の効率が落ちないよう編集するために，検索結果のウェブページのスニペット（要約）中からキーワードを抽出する方法を定式化した。人物情報の検索実験により，この方法が，検索結果のウェブページの本文中からキーワードを抽出する場合より再現率の高い時に優れているのを確かめた。大学教員の氏名を検索エンジン goo に質問（総数 7）として与え，検索結果の上位 100 件のウェブページについて，質問した教員についての内容かどうか判断した。上位 10 件を適合・不適合判断辞書の初期データとして学習し，11 件目からの閲覧で辞書を更新しながら適合可能性示唆の効果を検証した。そのときの再現率-適合率グラフを図 1 に示す。概念表現とあるのは，キーワードとあるウェブディレクトリ辞書を利用しない時のグラフである。従来手法とあるのは，検索結果をランク順に見ていった時のグラフである。

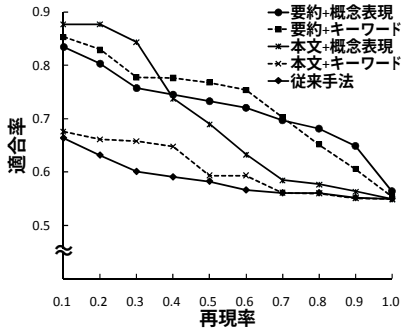

図 1 再現率-適合率グラフ

(2) パーソナル概念辞書を利用した聞き込み処理 [学会発表①，②]

入力された質問 Q について，ユーザが明確に検索意向を表わしていると思っても，検索エンジン側では多数の解釈が可能である場合には，初期の質問 Q だけでは適合・不適合判断辞書をうまく作れない。これには，パーソナル概念辞書を援用する。パーソナル概念辞書とは，汎用の概念辞書の一部分で，質問に対する検索結果としてのウェブページ中から抽出されたキーワードの集合 K に対し，各キーワードの綴りを概念名に含む概念だけに注目したものを指す。汎用概念辞書として EDR 概念辞書等がある。

キーワード集合 K を，コストをかけずに実時間で収集するため，適合・不適合判断辞書の編集時に使ったのと同様，スニペット（要約）中からキーワードを抽出するようにする。そして，ユーザの使い勝手を考え，パーソナル概念辞書そのものの提示を避け，辞書の構造を反映させたファセット表の形で提示する。ユーザは検索意向に関連する特定のファセットに属するキーワードでその意向を反映する 1〜2 を選び，初期の質問 Q に追加する形で利用する。新しい質問に対する検索（聞き込みと呼ぶ）で正例を 1〜2 見つけ，これらを Q に対する適合・不適合判断辞書の核として扱い，もとの処理を続行する。このようにして，検索エンジン側が多数の解釈をしてしまう場合の逆転問題に対処する。

人物情報の検索実験で，聞き込みを必要としない質問（総数 102）で再現率-適合率グラフを図 2(a)に，聞き込みを必要とする質問（総数 11）でのグラフを図 2(b)に示す。実験方法は上記と同様であるが，再現率-適合率グラフには，適合・不適合判断辞書の初期学習データとして使った，上位 10 件のウェブページの適合判断状況も含まれている。

図 2(a)で，RnkH は従来手法での結果，SugH-は Gce で適合・不適合判断辞書を利用した場合の結果のグラフである。SugH-は RnkH より統計的に見ても優れていた。図 2(b)で，RnkL は従来手法での結果，SugL は Gce で適合・不適合判断辞書とパーソナル概念辞書を利用した場合の結果，SugL-は適合・不適合判断辞書のみを利用した場合の結果のグラフである。SugL は RnkL より統計的に見て優れていた。以上より，逆転問題を緩和するのに，2 種の辞書が有効に働いたことがわかる。

両辞書を組み入れた Gce の仕組みを図 3 に示す。この Gce と市場の検索エンジンを連動させれば，概念的検索エンジンができあがる。

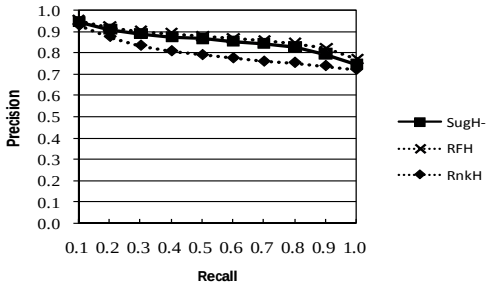

図2(a) 再現率-適合率グラフ（聞き込みを必要としない場合）

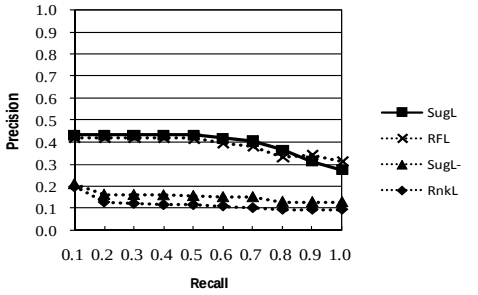

図2(b) 再現率-適合率グラフ（聞き込みを必要とする場合）

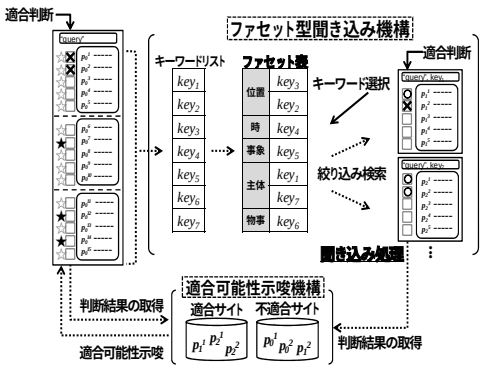

図 3 聞き込みを組み入れた適合可能性示唆

概念的検索エンジンの特徴は，市場の検索エンジンの利点を保存しながら，ユーザの適合判断情報と整理された概念関係情報を利用して，高速かつ高効率の検索を遂行するという点である。適合・不適合判断情報を利用した検索法として，従来からの関連性フィードバック法があるが（図 2 では RFH，RFL グラフでこれを利用した場合の結果を示している），これを市場の検索エンジンに組み込むと，ページランク情報が台無しになってしまう。ここでの方法では，ページランク情報

も有効利用する。概念関係情報の利用に関しては，一般の方法では，ユーザの質問作成支援に，あらかじめ編集しておいた概念辞書を提示する。これでは，ユーザにとって使いづらい上，使えても逆転問題の解決手段にはならない。ここでの方法では，現実の検索の状況に応じて小規模の概念辞書を適宜編集し，扱い易いファセット表の形で提示してユーザを支援する。逆転問題の解決にも使える。

改良した Gce を検索エンジン goo に組み込んだ。現在，ベータ版であるが，図 4 に示すようなインタフェースで利用できる（goo の外枠にあるツールバーや広告記事などは省いている）。今後，goo 以外の検索エンジンに組み込めるよう開発を進める。

図4 概念的検索エンジンのインタフェース

(3) パーソナル概念辞書の利用による画像へのキーワード付加支援 [雑誌論文①，学会発表③，⑤，⑦]

パーソナル概念辞書はキーワードをベースにした画像検索においても有効に働く。この種の検索では，各ウェブ画像に前もってその内容を示すキーワードを付加しておく必要がある。画像へのキーワード付加の仕方は 2 つある。1 つは，ウェブページの場合と同様，画像の周りにある説明文中からキーワードを自動抽出して与える仕方，もう 1 つはユーザが人手で与える仕方である。後者による方がより良い検索結果を生むのに繋がる。画像検索でも逆転問題や語彙問題が生じる。加えて，画像に人手でキーワードを付加するとなると，その際のコストが問題となる。

コスト問題の解決のために，通常は，汎用の概念辞書を用意するが，上に述べたような問題が生じる。ここでは新しく，パーソナル概念辞書の考え方を組み入れた画像へのキーワード付加支援機構を定式化する。具体的には，キーワード付加対象のウェブ画像 g について，(i) g の周りにあるウェブページの説明文，(ii) g と似た画像の周りにある説明文，(iii) g を見て思いつくキーワードで検索された画像でユーザが g と似ている判断した画像の周りにある説明文から取り出したキーワード，をもとにパーソナル概念辞書を編集する。この辞書を使う際には，その構造を反映したファセット表を，g のタイプ（オブジェクト，風景，風景にオブジェクト，雰囲気の 4 つのいずれか）に見合う形に整形して提示する。ユーザは整形されたファセット表から，付加すべきキーワードを選べばよい。付加すべきキーワードを思いつく必要はない。キーワード付加支援の仕組みを図 5 に示す。

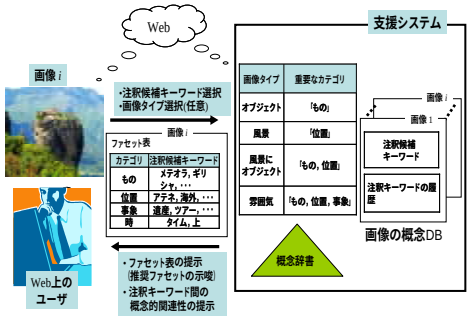

図 5 ウェブ画像へのキーワード付加支援

少数のユーザによるキーワード付加では結果に偏りが生じる危険性があるため，1 つの画像に対し複数のユーザがキーワードを付加することが肝要である。このために，別なコスト問題が生じる。必要最小限の数のユーザとするにはどうすれば良いか。解決策として，Gce の概念的にマッチ度を測る仕組み

を使う。今，あるユーザ u がキーワード付加をしようとするとき，先のユーザによって付加されているキーワードの集まりと，それより先の（複数の）ユーザによって付加されているキーワードの集まりとの概念的なマッチ度の増加具合をグラフ化して提示する。このグラフの増加具合が飽和していると，ユーザ u は，これ以上のキーワード付加は不要との判断できる。パーソナル辞書を用いた画像へのキーワード付加支援のためのインタフェースを図 6 に示す。

図 6 キーワード付加支援インタフェース

今後は，このキーワード付加支援法をウェブで公開し，多数のユーザに使ってもらえるようにする。また，市場のキーワード付加支援システムとの連携を図れるようにして，キーワード付加をより容易にしてゆく。画像の検索に関しては，ウェブページの検索と同様な仕組みを市場の検索エンジンに組み込み，概念的画像検索が可能となるようにする。

５．主な発表論文等

〔雑誌論文〕（計 1 件）

①M. Nakashima, T. Hiyama, K. Sato, and T. Ito: Proceeding with Keyword-based Web-Image Annotation Conceptually in Folksonomy, Proc. of CISIS 2009, Fukuoka, Japan, 995-1000 (2009 年 3 月 19 日) (査読有).

〔学会発表〕（計 7 件）

①佐藤慶三: ファセット型聞き込み機構を組み入れた適合可能性示唆機構の人物情報検索への適用，FIT 2009，全 4 ページ (2009 年 9 月（予定）).

②佐藤慶三: ファセット型聞き込み機構と適合可能性示唆機構の連携によるウェブ上での人物情報収集，火の国情報シンポジウム 2009，全 8 ページ(2009 年 3 月 14 日), 九州産業大学

③肥山高大: フォルクソノミーによる概念的画像注釈付加支援のための注釈候補キーワードの取得，火の国情報シンポジウム 2009，全 8 ページ (2009 年 3 月 14 日), 九州産業大学

④栗本大資: Web 検索の結果を広く閲覧できるユーザインタフェース, 2008 年度電子情報通信学会九州支部学生会講演会 D-13, 全 1 ページ (2008 年 9 月 26 日) , 大分大学

⑤平野拓也: ファセット表を用いた画像注釈付加支援のためのキーワード抽出, 2008 年度電子情報通信学会九州支部学生会講演会，D-13,全 1 ページ (2008 年 9 月 26 日), 大分大学

⑥佐藤慶三: ウェブディレクトリを元に構築した概念辞書を組入れた概念的検索エンジン，火の国情報シンポジウム 2008，全 8 ページ (2008 年 3 月 7 日), 長崎大学

⑦肥山高大: ファセット表を利用したキーワードの概念的扱いによる画像注釈付加支援環境の設計，情報処理学会第 70 回全国大会講演論文集，第 1 分冊，475-476 (2008 年 3 月 13 日). 筑波大学

６．研究組織

(1) 研究代表者
伊藤 哲郎 (ITO TETSURO)
大分大学・工学部・教授
研究者番号：30029558

(2) 研究分担者
中島 誠 (NAKASHIMA MAKOTO)
大分大学・工学部・准教授
研究者番号：00253774