D-023

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

A Facet-Based Legwork Mechanism for Website Effective Suggestion in Web People Search

佐藤 慶三† 中島 誠† 伊藤 哲郎†
Keizo Sato Makoto Nakashima Tetsuro Ito

1. はじめに

近年，特許や学術論文の投稿機会の増加に伴い，人物情報をウェブ上で収集することも多くなっている．検索エンジンを利用した情報収集では，ユーザの要求に合わないサイトも多く出力されてしまう．人物情報収集では人名の曖昧性（同姓同名）問題[14]があり，人名での質問だけでは，収集したい人物情報がうまく取り出せない．検索結果から適合サイトを探してゆく作業は労力を要するため，ユーザは多くの場面で潜在的な適合サイト獲得の機会を逃してしまっている．

様々なサイトが混在する検索結果に適当なキーワードで絞り込み検索を行うことで，ユーザにとって真に所望のサイトを見つけることが容易になる．得られた適合サイトの情報を利用する[7][13]ことで，未参照の適合サイトも効率的に見つけ出してゆける．

本稿では，絞り込み検索に適当なキーワードを，ファセット表の形でユーザに提示し適合サイト獲得を支援するファセット型聞き込み機構を，未参照のサイトの適合可能性示唆する機構[13]に組み入れた人物情報収集について提案する．また，この人物情報収集の仕組みについて，計算機実験を行った．実験を通して，ファセット表利用の実用性および検索効率の面での有効性について確認できた．

2. 関連研究

Google[3]などの商用検索エンジンでは，近年多数の検索結果を絞り込むために関連語を提示したり，複数のキーワードでの検索で質問中のキーワードとは異なるキーワードを示唆したりすることで検索結果の質を向上しようという取り組みがなされている．しかしながら，示唆されるキーワードの選定は検索エンジンが自動的に行うもので，ユーザの検索要求が十分に反映されておらず，実際に調べたい事項と関連のない情報の提示に陥ってしまう．

ユーザの判断基準を，検索履歴などを解析して得られるユーザプロファイルとして表現し，検索を支援する仕組みについても研究されている[9][16]．この仕組みは，長期的にユーザの嗜好を捉えようとするため，興味や要求の変化に対応できない．そのため，新たな視点や興味での情報収集には不向きである．

検索結果を分類することで参照コストを低減する試みもなされている[17]．[2][6]ではファセット分類の考え方にもとづいた検索支援を行っている．しかしながら，多面的なユーザの検索要求を正確に捉えることは困難なため，検索結果に対する視認性が悪化することで情報収集は非効率的なものになってしまう．

適合フィードバック[7]では，ユーザの適合判断結果を学習することでランキングの精度を向上させている．ユーザの適合判断を逐次的に反映させることで，効果的にユーザの興味を捉えられているが，精度を高めるのに頻繁に検索結果を再ランキングするため，参照済みの適合サイトを見失うなど，ユーザの情報収集活動に対する進捗把握を混乱させてしまう恐れがある．

3. 適合可能性示唆機構による人物情報収集

ここでは，適合可能性示唆機構[13]に適合情報を効率的に集めるための聞き込み処理を組み入れた情報収集の仕組みについて述べる．聞き込み処理とは，ファセット型聞き込み機構による，適合サイト収集処理である．

図1に聞き込み処理を組み入れた，適合可能性示唆機構による情報収集の流れを示す．ユーザの操作を実線，各機構の役割を破線で示してある．

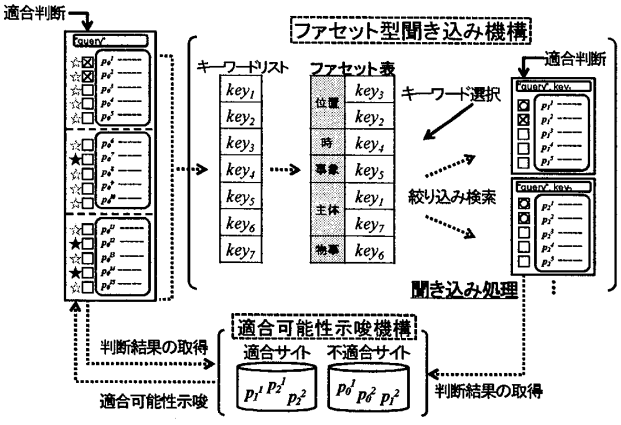

図1. 聞き込み処理を組み入れた適合可能性示唆機構の働き
Fig.1 A suggesting mechanism incorporating legwork process.

†大分大学工学部知能情報システム工学科

図中，左はユーザが初めに入力する検索質問（以下，初期質問と記す）の結果で，上位2件を参照して，適合サイトが得られなかったことを示している．図中右上のファセット型聞き込み機構は，検索結果から得られるキーワードをファセットに分類している．ユーザが選択したキーワードで絞り込んだ検索結果（図中右）の参照を通して得られたサイトが，図中下の適合可能性示唆機構に格納されている．これにより，初期質問の検索結果中の未参照サイトについて適合可能性示唆が提示されるようになっている．

聞き込み処理は，絞り込まれた検索結果の少数の参照で完了する．また，絞り込みにより検索結果自体の適合サイトの比率が高くなるため，この処理は時間的にも小さいコストで済む．聞き込み処理で得られた適合サイトを，初期質問の検索結果の参照に利用することで，より多くの適合サイトを集められるようになる．

適合可能性示唆機構は，参照済みの適合あるいは不適合サイトを学習することで未参照サイトの適合可能性を示唆する仕組みである．適合可能性の判定は以下の式で行う．

$$\max\left(s(q,p_i),\ \max_{p'\in P'} s(p',p_i)\right) \geq \max\left(\tau,\ \max_{p''\in P''} s(p'',p_i)\right)$$

式中，$p'(p'')$は適合判断済みの適合サイト（不適合サイト）を表し，$P'(P'')$は$p'(p'')$の集合を表す．閾値τは，不適合サイトの学習不足時に不適切な適合可能性示唆を回避するために取り入れている．適合判断のたびに，P'あるいはP''が更新される．

この適合可能性示唆機構を検索結果の参照に適用するのに，以下の手順を行う．

(S1) 検索結果上位数件を参照し，適合判断する
(S2) 適合サイトが十分に見つからなければ，聞き込み処理を行う
(S3) ランキング順に適合可能性を示唆されたサイトを参照する
(S4) 未参照サイトをランキング順に参照する

(S1)では，適合可能性の示唆によらず数件のサイトを参照し，適合判断結果を入力する．ここで得られた適合判断結果をその先の検索結果を参照する際の適合可能性示唆に用いる．一般に検索結果ページの上位数件については特に支援なしに参照しているというユーザの検索行動[4]にもとづいている．(S1)で適合サイトが得られなければ(S2)で聞き込み処理を行う．(S1)，(S2)での適合判断を通して得られた情報をもとに，(S3)で適合可能性示唆にもとづいてサイトの参照を行う．ここでも適合判断を継続的に行うことで，未参照サイトの適合可能性をより正確に判定できる．(S4)は，参照漏れを防ぐための処理であり，(S3)までで十分な適合サイトが得られていれば省略してもよい．

キーワードの抽出には検索結果ページに表示される要約（スニペット）を利用する．これは，展示会における出展企業一覧や学会発表のプログラムなどを掲載しているサイトでは，調べたい特定の企業や発表者以外に関連するキーワードが多数抽出されてしまうためである．

4. ファセット型聞き込み機構の働き

ここでは，聞き込み処理を行うためのファセット表の構築方法及び利用方法について詳細を述べる．

4.1 検索結果をもとにしたファセット表構築

検索結果から得られるキーワードによれば，そのキーワードを含むサイトの存在が保証されるため，検索結果中の適合サイトを取り出すのに有効である．しかしながら，多数のキーワードの中から，適当なキーワードを選ぶのは容易ではない．これに対し，キーワードの属性に着目し，それに沿ってあらかじめ分類し，キーワード選択を支援する．このような，分類にもとづく情報提示は図書館情報学の分野でも長く有効な手段として利用されている．また，この考え方を画像注釈に取り入れた研究もなされている[8].

検索エンジンを利用して得られるサイトは様々な分野の情報が混在するため，出現するキーワードを体系づける基準がない．ここでは，[8]でも用いられているEDR概念辞書[4]の概念体系を利用する．この辞書のROOT直下にある最上位概念ついて，各キーワードを対応づけることでキーワードの属性に着目したファセット表を構築する．下位概念を有さない概念や未分類の概念を除く，「位置」，「時」，「事象」，「主体」，「物事」の５つ上位概念をファセットとして扱う．

キーワードをつづりの似た見出し語と置き換え，その見出し語を概念体系にしたがって分類することで，自動的にファセット表を構築し，ユーザに提示する．各ファセットのキーワードは文字列長の昇順でソートしておくことで視認性が向上する．見出し語に置き換えることで，サイトに多くみられる表記ゆれも吸収され，キーワード数を縮退させられる．サイト中の実際の表記と異なっても，検索エンジン自体が同義語などに対処してきているため，異なる表記のサイトが検索されなくなる，という問題はない．

つづりの類似性を数量化[1]し，一定以上の値を示す場合をつづりが似ているとみなす．これは，わずかな表記のちがいがあってもうまくファセットに対応付けるためである．キーワードの文字数をn，1文字単位の合致数を$s1$，2文字単位の合致数を$s2$とすると，以下の式で数量化できる．

$$(s1+2\times s2)/(n+2\times(n-1))$$

ただし，EDR 概念辞書の総概念数は，中間概念を含めて

380,944で，各概念の見出し語の異なり語数は148,986となっている．約15万語の見出し語すべてと照合すると時間がかかる．上の式によれば隣接2文字組みで合致する組のないキーワード間では最大でも，

$$n/(n+2\times(n-1))$$

となり，$n \geqq 2$のもとでは0.5以下の値しか示さない．この性質を利用し，見出し語中に出現する隣接2文字組みと見出し語を対応づけるインデックスを作成しておく．これにより，高速な対応づけが可能になり，ファセット表を短時間で構築できる．

固有名詞の多くはEDR概念辞書に含まれていないため，形態素解析器[15]の品詞情報を利用する．得られる種々の品詞のうち「名詞-固有名詞-人名」はEDR概念辞書中の「人間」概念，「名詞-固有名詞-組織」は「組織」概念にそれぞれ置き換える．

4.2 人物情報に対応したファセット表の利用

ファセット表で整理されたキーワードについて，いずれかのファセットだけで必要なキーワードが特定できれば，適合情報の収集をさらに効率化できる．サイト中に出現する人物情報については，その人物を特定するのに有効な16の属性[14]の中でも，「職業」，「組織名」はキーワードとして扱える．これらは，EDR概念辞書によれば「主体」概念に属するため，このファセットのキーワードからだけでも，特定の人物についての情報を十分に取り出せる．

注目するファセットが限定されることで候補となるキーワードが絞り込まれていれば，そこから特定のキーワードを選び出すのは容易である．これは，キーワードの選択プロセスではサイトを参照する際のように「文章を読解する」必要がないためで，このプロセスが一種の画像認識とみなせるためである．ヒトの画像認識能力が高いことはよく知られており，[8]によれば，実際の計測でキーワードの選択に要する時間は，キーワード１つあたり0.4秒程度で済むことが確認されている．

5. 実験的考察

3.，**4.**で述べた，情報収集の仕組みについて有効性を調べた．実験データには，大分大学工学部の職歴を有する教員および研究員120名の氏名を質問とし，各質問で上位100件の検索結果をテストデータに用いた．各サイトの，適合性判定には[10][11][12]に掲載されている研究者情報を利用した．7名については，検索結果に不適合サイトが含まれなかったため除外し，113名分のデータを実験データとして採用した．適合可能性判定の類似度計算にはコサイン測度を用いた．

ファセット型聞き込み機構の有効性を検証するために，質問を２つのグループに分けた．**3.**の(S1)での適合判断を検索結果ページ1ページ分（上位10件）としたうえで，適合サイトが2件以下しかない質問を聞き込み処理の必要な質問のグループ（グループL）とした．また，グループLに属さない質問のグループをグループHとした．グループLは11名，グループHは102名となった．

5.1 聞き込み処理のコスト評価

ここでは，**3.**で述べたファセット表による聞き込み処理の，時間的コストへの影響について考察する．対象として，グループLに属する質問を取り上げる．まず，ファセット表構築の所要時間について調べたところ，各質問で平均1101.6得られたキーワードでファセット表を構築するのにかかった時間は平均0.9秒であった．この所要時間であれば，ユーザが質問を入力してから検索結果を参照できるまでの待ち時間は短く，情報収集の作業の妨げになることはないといえる．

ファセット表によるキーワードの選択をシミュレートするために，各質問について聞き込み処理で利用するキーワードを選出した．人物特定情報として有効[14]な，組織を表すキーワードおよび同僚や上司，部下など人名を表すキーワードのうち，適合サイトに出現するものを選んだ．質問１件当たり20.2個のキーワードが得られた．

[8]の観察結果をもとに，1キーワードあたりの目視時間を0.4秒として，得られたキーワードをファセット表上で見つけるのに係る時間をシミュレートした．時間は，

ファセットまたはリスト中での出現ランク×0.4秒

として求めた．その結果，ファセット表を用い，主体ファセットを参照した場合は69.3秒となった．ファセット表を用いないリスト形式での表示では平均247.0秒かかっていた．このことから，キーワードの選択時間が大幅に短縮できたことが分かった．

聞き込み処理のうち，最も大きな時間がかかるのは絞り込み検索の結果に対する適合判断である．この時間について，サイト１件あたりの参照時間23.9秒[18]をもとに計算すると，上位10件分の適合判断で239.0秒（=4分）となることが分かった．

5.2 検索効率評価

検索効率について，**4.**での手順（Sugと記す）を調べた．比較のために，全てをランキング順に参照した場合（Rnkと記す）と，適合フィードバックによる方法（RFと記す）も調べた．グループごとに'L'，'H'を付して記す．

図2にグループLでの実験結果を示す．グループLについて(S2)の聞き込み処理を行わなかったときの結果をSugL-として記した．聞き込み処理には，**5.1**で選出したキーワードを利用し，各質問での平均をもとにグループLでの平均を計算した．図から，聞き込み処理により，SugL-やRnkLよりも高い適合率が得られたことが分かる．SugLやRFLがRnkLに対し統計的な有意差がみられたのに対し，SugL-はRnkLに対し統計的に有意な差はみられなかった．

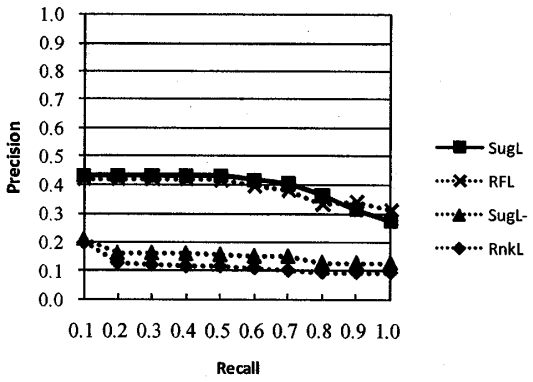

図2. 再現率-適合率グラフ(GroupL)
Fig.2 Recall-precision curves for Group L.

聞き込み処理について，平均的には検索効率が改善されていたが，汎用的であるために十分な改善が得られなかったキーワードもみられた．実際の運用では，ユーザの検索対象人物に対する断片的な知識が利用できるため，これらのキーワードは選択されないものと思われる．

図3に，グループHでの検索効率を示す．RnkHでも高い適合率を示していた．また，SugH-，RFHいずれもRnkHより統計的に優れていた．聞き込み処理を経ずに高い適合率が得られたことから，(S1)での適合サイトの獲得状況に応じて，聞き込み処理を行えばよいことが分かった．

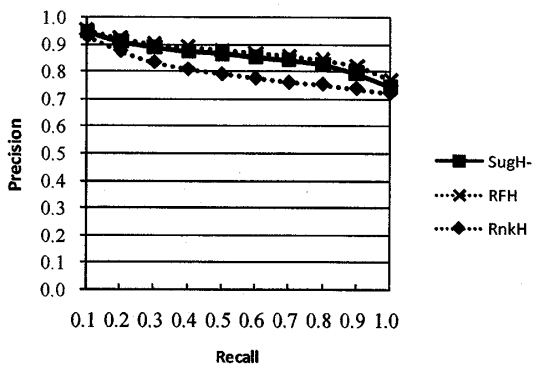

図3. 再現率-適合率グラフ(GroupH)
Fig.3 Recall-precision curves for Group H.

グループL，Hの検索効率の実験を通して，SugはRFと同様に高い適合率を示し，両者間での有意差はみられなかった．ただし，RFでは頻繁に検索結果の再ランキングが発生していた．このことから，聞き込み処理を適合可能性示唆機構に組み込むことが有効であると言える．

6. おわりに

本稿では，ファセット型聞き込み機構の実装により，適合情報を効率的に集める手法について述べた．ファセット表によるキーワードの提示で，適当な関連語を短時間で見つけ出せることが確認できた．また，検索効率の比較により，聞き込み処理により収集した適合情報が有効に働くことも検証した．

今後の取り組みとしては，聞き込み機構と示唆機構の連携による情報収集の仕組みを人物情報収集以外のケースに適用した場合の有効性評価も行う予定である．特に，本稿では検索対象を人物情報に限定したため，ファセットの参照範囲を「主体」概念に限定できたが，今後は検索内容に応じたファセットの示唆基準についても，検討してゆく．

参考文献

[1] 安部ほか: 構造マッチングによる文献の知的検索と結果の色空間表示，情報処理学会研究報告，人文科学とコンピュータ，Vol.29, No.5, pp.25-30, Jan. 1996.
[2] Dash, D., et al.: Dynamic Faceted Search for Discovery-driven Analysis, Proc. CIKM'08, Napa Valley, California, USA, October 26-30, 2008.
[3] Google.: http://www.google.com
[4] iProspect: Search Engine User Behavior Study, April 2006.
[5] Japan Electronic Dictionary Research Institute, Ltd.: EDR Electronic Dictionary Version 2.0. 1998.
[6] Koren, J., Zhang, Y., Liu, X.: Personalized Interactive Faceted Search, Proc. WWW2008, Beijing, China, pp.477-485, April 21-25, 2008.
[7] Limbu, D.K., et al.: Contextual Relevance Feedback in Web Information Retrieval, Proc. IIiX'2006, Copenhagen, Denmark, pp.138-143, Oct. 18-20 2006.
[8] Nakashima, M., et al.: Proceeding with Keyword-based Web-Image Annotation Conceptually in Folksonomy, Proc. VENOA2009, Fukuoka, Japan, pp.995-1000, Mar. 2009.
[9] Peter, I.H.: Web Personalisation Through Incremental Individual Profiling and Support-based User Segmentation, Proc. WI'07, Silicon Valley, USA, pp.213-220, Nov. 2007.
[10] 大分大学研究者一覧トップページ.: http://bunsyo1.ad.oita-u.ac.jp:8080/kentop.asp
[11] ReaD研究開発支援総合ディレクトリ.: http://read.jst.go.jp
[12] 産学プラザ研究者データベース.: http://www.sangakuplaza.jp/category/researcher
[13] Sato, K., et al.: The Effect of a Website Directory When Employed in Browsing the Results of a Search Engine, IJWIS, vol.1, no.1, pp.43-51, Mar. 2005.
[14] 関根聡: Web検索における人名の曖昧性解消技術の動向-同姓同名のクラスタリング-, 情報処理, vol.49, no.5, pp.573-578, May 2008.
[15] 日本語形態素解析システムSen.: https://sen.dev.java.net
[16] Sieg, A., et al.: Web Search Personalization with Ontological User Profiles, Proc. CIKM'07, Lisbon, Portugal, pp.525-534, Nov. 6-8 2007.
[17] 杉山ほか: Web検索結果における人名の曖昧性解消への半教師有りクラスタリングの適用，情処研報，vol.2007, no.94, pp.15-20, Sept. 2007.
[18] Terai, H., et al.: Differences between Informational and Transactional Tasks in Information Seeking on the Web, Proc. IIiX'08, London, UK., pp.152-159, Oct. 14-17 2008.