情報処理学会論文誌　データベース　Vol. 4　No. 3　12–21 (Oct. 2011)

検索システムユーザの特定分野に関する専門性推定のためのクリックスルーログの利用

佐藤大祐[†1]　安田宜仁[†1]
小池義昌[†1]　片岡良治[†1]

検索システムにおいて，ユーザの多様な要求に応えることは重要である．我々は数ある要求の中でも，ユーザの知識の程度に合った情報への要求に着目し，ユーザの検索対象に対する知識度合いを専門性と定義する．本稿ではユーザの専門性を推定するための手法を提案する．提案手法では，検索システムの持つクリックスルーログを利用する．ユーザと文書をノードとし，クリック関係のあるユーザと文書のノードの間にエッジが存在する二部グラフを考える．ユーザと文書のノードには，既存手法を用いて専門性を初期スコアとして与えることができ，それぞれの初期スコアを，二部グラフのエッジを通して相互に伝播させることにより，推定精度の向上を目指す．検索語履歴を用いた既存手法との比較を行った結果，本手法が有効であることを確認した．

Estimation of the User's Domain Expertise in Information Retrieval System Using Click-through Log

Daisuke Sato,[†1] Norihito Yasuda,[†1]
Yoshimasa Koike[†1] and Ryoji Kataoka[†1]

It is important that search engines be able to meet various user needs. We focus on the need to tailor the search results to suit the user's knowledge. In this paper, we define the domain expertise of a user as the amount of knowledge he/she has about the information he/she requires. We propose a method that can estimate the domain expertise of search engine users by using click-through log entries. We view the relationship between users and selected web documents as a bipartite graph. There are edges connecting the user node and the document node clicked by the user. The scores of domain expertise given to each node by the conventional method are propagated from both sides of the bipartite graph to improve the precision of the estimation. Evaluation results show the effectiveness of our method.

1. はじめに

インターネット上の文書数が増大するにつれて，検索語と関連性のある文書が多数存在する場合が多くなってきた．検索システムが，検索語と関連性のある文書を多数提示できる場合，単に検索語との関連性の程度によって並べただけの検索結果では，多様な検索目的を持つユーザの要求を満たすことは難しい．ユーザの検索結果への要求には，より新しい情報への要求，より現在地に近い情報への要求などの多様な軸があり，検索語との関連性だけでなく，様々な観点から適切な検索結果を提示することが必要である．

現在多くの Web 検索エンジンでは検索語との関連性だけでなく，PageRank などのリンクベースの手法や，文書の更新日を利用するなど，ユーザの多様な要求に応える工夫がなされている[7]．本稿では検索に対する数ある要求の中でも，より自分の持つ知識の程度に合わせた情報への要求に着目する．たとえば，検索語としてある病名を入力した場合，ユーザが医療に関する知識を有する看護師であれば，その病気について専門的な内容の文書を提示することは適切であると考えられる．一方，ユーザが医療についてまったく知識を有さない患者であった場合，専門的な内容の文書ではなく，より平易な言葉で書かれた文書を提示することが適していると考えられる．我々は，ユーザの持つ検索対象に関する知識の度合いによって，それぞれに異なる適切な文書を検索結果として提示することが望ましいと考える．

このような考え方は，広義にはユーザに合わせた検索結果を提示する検索のパーソナライゼーション（パーソナライズド検索）ととらえることができる．本研究では，このような検索システムを実現するための，ユーザの検索対象に関する知識の度合いを推定することを目的とする．本稿ではユーザの望む情報に関する知識の程度を，ユーザが持つ専門性と定義する．また，ある文書について読者が支障なく理解できるのに必要なユーザの専門性を，文書の専門性と定義する．

我々は過去の研究において，検索語履歴に着目した，ユーザの専門性推定手法を提案した[16]．本稿では，検索システムの持つクリックスルーログに着目した手法を提案する．提案手法では，ユーザと文書をノードとし，クリック関係のあるユーザと文書ノードの間にエッジが存在する二部グラフを考える．ユーザと文書のノードには，既存手法を用いて専門性を初期スコアとして与えることができ，ユーザと文書それぞれの初期専門性スコアを，二部グ

†1 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

ラフのエッジを通して相互に伝播させることにより，検索語履歴に着目した手法よりさらに高い推定精度を目指す．

本稿の構成は以下のとおりである．まず検索システムユーザの専門性を推定するための提案法について 2 章で説明する．3 章では提案法の推定精度を評価するための実験方法および結果を述べる．4 章で本研究の関連研究を，5 章でまとめを述べる．

2. 提 案 手 法

本章では，クリックスルーログを用いることによってユーザの推定精度を向上させるための提案法について述べる．まず，クリックスルーログをユーザの専門性推定に利用するうえでの基本的な考えを述べ，続いて手法について詳細に説明する．

2.1 アプローチ

通常，検索エンジンを利用して検索が行われる場合，どのユーザがいつ，何のクエリで検索し，検索結果からどの文書をクリックしたか，というクリックスルーログを蓄積する．ユーザの検索行動にはユーザの特徴が現れており，クリックスルーログはユーザの特徴を調べるうえで有益な情報源であると考える．また，クリックスルーログは，新たに収集する必要がないという点，そして，多くのユーザに関するログが得られるという点で，利用する情報源として優れていると考える．

提案手法は，クリックスルーログを用いてユーザの専門性の推定精度の向上を図るものである．クリック関係でつながるユーザと文書が，お互いの持つ専門性に関する情報を相互に利用することによりこれを実現する．提案手法における基本的な考えは以下の 2 つである．

- 似たような文書をクリックするユーザは，同じような専門性を持つ．

ユーザに専門性が与えられている場合，同じ文書をクリックしているユーザの専門性を利用することで，もとから与えられているユーザの専門性の推定精度を向上させることができると考える．

- 専門性の高い文書を多くクリックしているユーザは，専門性の低い文書を多くクリックしているユーザより専門性が高い．

文書に専門性が与えられている場合，そのユーザがどのような専門性の文書をクリックしているかという情報を利用することで，ユーザの専門性の推定精度を向上させることができると考える．

そこで，提案手法ではクリックスルーログを利用するために，ユーザと文書とのクリック関係をエッジとする二部グラフを作成する．そして，各ノードに初期値として与えた，文書またはユーザの持つ専門性を，二部グラフのエッジを通して相互に伝播させるというモデルである．このときノードの持つ値は，特定分野に関する専門性スコアであり，高いスコアを持つノードとの間にエッジが多く存在するノードに対し，高いスコアが与えられる．スコアの伝播を繰り返すことにより，エッジでつながっているノードは，同じような専門性スコアを持つようになる．

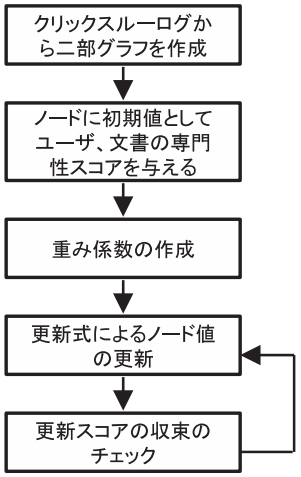

図 1　提案手法手順
Fig. 1　Procedure of proposed method.

2.2 手 法 概 要

図 1 に提案手法の手順を示す．まず，クリックスルーログに含まれる，どのユーザが，検索結果からどの文書をクリックしたかという履歴から二部グラフを作成する．次に，作成した二部グラフの各ノードに初期値としてのユーザ，文書の専門性スコアを与える．そして，ノードからノードへ情報を伝える際の重み係数を作成し，重み係数を含む更新式により更新を行う．更新はスコアの変化量が閾値を下回るまで行う．

2.3 二部グラフの作成

クリックスルーログから，二部グラフ $G = (U \cup V, E)$ を作成する．頂点集合 U はユーザ集合を表しており，$U = \{u_i\}_{i=1}^{m}$ としたとき各 u_i はそれぞれのユーザを表す．頂点集合 V は文書集合を表し，$V = \{v_j\}_{j=1}^{n}$ としたとき各 v_i はそれぞれの文書を表す．エッジ E はユーザと文書のクリック関係を表しており，ユーザが文書をクリックした場合にクリック

したユーザとクリックされた文書の間にエッジを張る．よって，エッジは集合 U と集合 V との間にのみ張られ，集合 U および集合 V 内にエッジは存在しない．ユーザ集合 U と文書集合 V 間のエッジの有無を表す行列を A^{UV} とする．A^{UV} の要素 a_{if}^{UV} は，ユーザ u_i と文書 v_j の間にエッジが存在する場合 $a_{ij}^{UV}=1$，エッジが存在しない場合 $a_{ij}^{UV}=0$ とする．また，$A^{VU}=(A^{UV})^T$ とし，$a_{ji}^{VU}=a_{ij}^{UV}$ とする．

2.4　初期値として与える専門性スコア

作成した二部グラフの各ノードに，初期値としての専門性スコアを与える．x_i をユーザ u_i の専門性スコアとし，y_j を文書 v_j の専門性スコアとする．与える専門性スコアの算出方法は本稿の議論の範囲外であるが，ユーザ，文書ノードが相互に情報を補足することにより推定精度の向上を図るという提案手法の特性上，専門性スコアの算出方法は，ユーザに関しては文書の，文書に関してはユーザの情報を含まずに算出されているものを選択する．

2.5　重 み 係 数

提案手法では，二部グラフのエッジを通して専門性スコアを伝播させることによりノードの値を更新する．ノードの値の更新は，ノード間にエッジが存在する，すべてのノードから伝播される専門性スコアを足し合わせることによって行う．このとき，単に伝播されるスコアをすべて足し合わせると，ノードの持つエッジの数によって更新スコアに大きな差を生じさせてしまう．つまり，エッジを多く持つノードは，たとえ伝播元のノードが低いスコアのものばかりであっても，更新によって高いスコアが与えられてしまう可能性がある．そこで，伝播されるスコアに対し，ノードの持つエッジ数で正規化が行われるような重み係数をかける．ノード u_i からノード v_j への伝播スコアにかかる重み係数を w_{ij}^{UV}，ノード v_j からノード u_i への伝播スコアにかかる重み係数を w_{ji}^{VU} とし，次の式によって求める．

$$w_{ij}^{UV}=\frac{a_{ij}^{UV}}{\sum_{i\in U}a_{ij}^{UV}}, \tag{1}$$

$$w_{ji}^{VU}=\frac{a_{ji}^{VU}}{\sum_{j\in V}a_{ji}^{VU}}. \tag{2}$$

図 2 に示す例を用いて，重み係数について説明をする．図 2 における左の図は，V から U へスコアを伝播させる例を示しており，エッジ上の数字は重み係数 w_{ji}^{VU} の値である．ノード v_1 に着目すると，v_1 は u_1 と u_2 との間にエッジを持ち，v_1 の持つスコアをそれぞれに伝播する．u_1 の持つエッジ数は 1 であるため，重み係数 w_{11}^{VU} は 1 となる．一方 u_2 の持つエッジ数は 4 であるため，w_{12}^{VU} は 0.25 となる．w_{22}^{VU}，w_{23}^{VU}，w_{24}^{VU} も同様に 0.25 となる．図 2 における右の図は U から V へスコアを伝播させる際の重み係数を示しており，エッジ

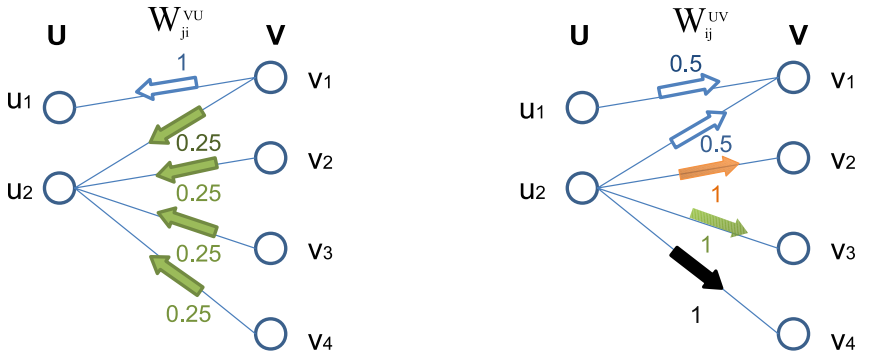

図 2　重み係数例
Fig. 2　An example of the weight coefficients.

上の数字は重み係数 w_{ij}^{UV} の値である．U から V へのスコアの伝播の場合とは逆に，v_1 はエッジで結ばれている u_1 と u_2 からスコアの伝播をうける．v_1 の持つエッジ数は 2 であり，u_1 から v_1 へ伝播するスコアにかかる重み係数 w_{11}^{UV} は 0.5 となる．同様に w_{21}^{UV} も 0.5 である．v_2，v_3，v_4 の持つエッジ数は 1 であるため，w_{22}^{UV}，w_{23}^{UV}，w_{24}^{UV} はともに 1 となる．

2.6　伝播による更新

ユーザ，文書それぞれの専門性スコアの更新を行う．更新は，あらかじめ与えた初期スコアと，伝播により更新されるスコアを足し合わせて行う．伝播による更新式は次のとおりである．

$$x_i^k=\lambda_U x_i^0+(1-\lambda_U)\sum_{j\in V}w_{ji}^{VU}y_j^{k-1}, \tag{3}$$

$$y_j^k=\lambda_V y_j^0+(1-\lambda_V)\sum_{i\in U}w_{ij}^{UV}x_i^{k-1}, \tag{4}$$

ここで，k は更新回数であり，x_i^k，y_j^k は k 回の更新後のユーザ，文書それぞれの値を表している．x_i^0，y_j^0 は初期値を表している．w_{ji}^{VU}，w_{ij}^{UV} は，重み行列の要素である．λ_U，λ_V はユーザノード，文書ノードそれぞれにあらかじめ与えた初期値と，伝播によって得られるスコアを足し合わせる割合を表し，0 から 1 の間の値である．λ_U が 1 に近い値であるほどユーザノードの持つ専門性スコア x_i が伝播によって更新する量が小さくなり，$\lambda_U=1$ であるとき，x_i は更新されない．0 に近いほど伝播によるスコアの更新量は大きくなる．同様に，λ_V が 1 に近い値であるほど文書ノードの持つ専門性スコア y_j の更新量は小さくなり，

$\lambda_V = 1$ であるとき，y_j は更新されないため初期値のままである．

式 (3)，(4) から，ユーザノードの持つ専門性スコアの更新量は次の式 (5) で表される．

$$x_i^k - x_i^{k-2} = (1-\lambda_U)(1-\lambda_V)\sum_{j\in V}\sum_{i\in U} w_{ji}^{VU} w_{ij}^{UV}(x_i^{k-2} - x_i^{k-4}). \quad (5)$$

ここで式 (5) から，更新量 $x_i^k - x_i^{k-2}$ が最も大きくなるときの λ の値は，λ_U，λ_V がともに 0 のときであることが分かる．さらに，付録に示すとおり，$\lambda_U = 0$，$\lambda_V = 0$ のとき，更新を繰り返すことによって x_i^k の値は収束する．したがって，k の値が十分に大きい場合，更新量 $x_i^k - x_i^{k-2}$ は，λ の値にかかわらず 0 に収束する．更新量が十分小さくなるまで更新を繰り返し，あらかじめ設定した閾値を下回った時点で更新を終える．最終的に得られた x_i^k の値を提案手法によるユーザの専門性の推定結果とし，y_j^k の値を提案手法による文書の専門性の推定結果とする．

3. 評価実験

提案法によって推定精度が向上するかどうかを検証するために実験を行った．提案法は，初期値として与えたユーザの専門性を向上させるものであるため，提案法による推定結果と，ユーザのクエリログを用いた推定手法[16] により与えた初期値を比較する．

3.1 実験手順

3.1.1 データセット

実験で使用する検索ログの収集を行った．まず，検索者の自己申告による特定分野の専門性を収集した．本実験では，スポーツ分野を評価対象とした．スポーツ分野は，幅広いユーザによって検索され，かつユーザによって知識の差が存在すると思われるためである．自己申告による専門性は，専門性が低い，やや低い，どちらでもない，やや高い，専門性が高い，という 5 段階から検索者が選択を行った．専門性スコアは，専門性の低い順に 1 から 5 までの整数とした．

次に，検索者がスポーツ分野に関する検索を行った．これにより，検索者の自己申告の専門性と対応がついているクエリログを収集した．**表 1** にデータセットをまとめている．

3.1.2 ユーザに与える初期値

ユーザに与える初期値は，我々が過去に提案した手法[16] によって算出した専門性スコアを与えた．以下に概要を示す．

手法では，専門性スコアを算出するために，以下に定める IQF（*Inverse Query Frequency*）を用いる．分野 q におけるクエリ t の IQF を表す $iqf_q(t)$ を求める式を次のように定義する．

$$iqf_q(t) = \log\frac{N_q}{qf(t)} + 1, \quad (6)$$

ここで N_q は，クエリログ中の対象分野 q に属する全クエリの数，$qf(t)$ は，クエリ t を入力したユニークユーザ数である．クエリ t が分野 q に対して専門性の高い語句であるほど $iqf_q(t)$ が高くなる．次に，ユーザ u の分野 q における専門性スコアを次のように定義する．

$$\text{専門性スコア} = \frac{\sum_t iqf_q(t)}{N_{uq}}, \quad (7)$$

ここで $N_u q$ は，ユーザ u が入力したクエリのうち，対象分野 q に属する語句がクエリとして入力された回数である．本実験では，スコア算出に必要となるクエリの入力者数の統計データとして，商用検索エンジンを利用した一般ユーザの，2008 年 6 月から 2010 年 1 月までの約 20 カ月分のクエリログを用いた．

表 1　データセット
Table 1　The dataset.

検索エンジン	商用検索エンジン
検索タスク	スポーツに関することであれば自由
検索被験者人数	47 人
1 人あたりの平均クリック回数	110 回
1 人あたりの平均ユニーククエリ数	66

3.1.3 文書に与える初期値

文書に与える初期値は，中谷らの手法[18] における，文書の専門性に関する部分を参考に専門性スコアを与えた．文書内に含まれるスポーツに関する専門用語を判別し，専門用語の出現頻度から文書の専門性スコアの算出を行った．専門用語の判別や，専門用語の出現頻度などを求めるために，中谷らの手法と同様に Wikipedia のデータを用いた．本実験では 2010 年 6 月 24 日時点の Wikipedia データを利用した．文書の特定分野に関する専門性スコアを算出するという目的のため，参考手法とは，専門用語の判別を行うための候補語の抽出方法，判別式におけるカテゴリに含まれる Wikipedia 記事集合，専門性スコアの算出式の一部が異なる．以下に手順を示す．

まず，被験者がクリックした文書中から，専門用語の候補となる語の抽出を行った．候補となる語を抽出するために，被験者がクリックした文書のテキストデータに対し，多言語固有表現抽出器 Namelister[15] を使用し，固有表現の抽出を行った．抽出された固有表現のう

ち，Wikipedia 記事のタイトルとなっているものを候補語とした．

次に候補語が，当該分野に関する専門用語かどうかの判別を行った．次の条件式 (8) を満たす語 t を当該分野に関する専門用語とした．

$$\frac{LF(t, W_q^+)}{|W_q^+|} > \frac{LF(t, W_q^-)}{|W_q^-|}, \tag{8}$$

ここで，当該分野に関する Wikipedia 記事集合を W_q^+，その他の記事集合を W_q^-，語 t の記事へのリンクを持っているものの数を $LF(t, W)$ としている．W_q^+ は分野に対応する Wikipedia カテゴリを基に決定し，以下に収集方法について述べる．Wikipedia 内のカテゴリは，サブカテゴリを持つ階層構造となっている場合もあるため，サブカテゴリをたどることにより当該分野に関する記事の収集を行った．まず Wikipedia のカテゴリリンクデータより，当該カテゴリに含まれる記事およびサブカテゴリの取得を行った．通常の記事はそのまま記事集合 W_q^+ に加え，サブカテゴリについては，そのサブカテゴリに含まれる記事，およびサブカテゴリを新たに取得した．サブカテゴリをたどり続けると，しだいに当該分野と関係のないページを多く含むようになるため，一定の回数で打ち切り，それまでに得られた記事を記事集合 W_q^+ とした．打ち切り回数を決定するため，各サブカテゴリごとに収集した記事集合からランダムに選択した 100 ページについて，各記事が当該分野と関係があるかどうかを人手で判定した．判定者は 1 人である．判定の結果，サブカテゴリを 3 回までたどって記事を収集した場合は，無関係と判定されたページは 100 ページ中 4 ページであったのに対し，サブカテゴリを 4 回までたどると，無関係と判定されたページは 100 ページ中 12 ページであった．このため，4 回以上たどると多くの関係のないページを含むことから，専門用語の判定のために用いる記事集合を収集する際にサブカテゴリをたどる回数を 3 回とした．収集したすべての Wikipedia 文書のうち，W_q^+ に含まれない記事集合を W_q^- とした．

判別されたスポーツに関する専門用語を用いて，次の式によって文書のスポーツ分野における専門性スコアを算出した．

$$\text{専門性スコア} = \exp\left(-\frac{1}{\log|c_i|}\sum_{t\in T(d_i,q)}\frac{1}{LF(t, W_q^+)}\right), \tag{9}$$

ここで，$T(d_i, q)$ は文書 d_i に含まれる，分野 q の専門用語集合を表し，また，$|c_i|$ は文書 d_i 中に含まれる候補語の数を表している．

表 2　λ_U，λ_V，相関係数および MAE
Table 2　λ_U, λ_V, correlation coefficient and MAE.

λ_U	λ_V	cor	MAE
0	0	0.212	2.15
0.15	0.47	0.691	0.876
1	1	0.528	1.03

3.1.4　提案手法による推定

提案手法による専門性の推定を行った．提案手法，従来手法によって算出されるスコアは，ユーザ間の相対的な値として得られる．そこで，自己申告と同じスコアの幅となるように，最小値が 1，最大値が 5 となるように正規化を行った．従来手法によって与える初期値および，提案手法による推定結果は正規化したものである．推定結果の精度は，正規化の方法によって影響を受けない，自己申告との相関をとることによって求めた．提案手法における更新を停止するための条件は，各ユーザノードの更新量の絶対値の総和を使って以下の条件式とした．

$$\sum_{i\in U}|x_i^k - x_i^{k-2}| < \theta. \tag{10}$$

本実験では，更新を停止させるための閾値 θ を 0.01 とした．また実験では，更新式中に含まれるパラメータ λ_U，λ_V を，それぞれ 0 から 1 まで 0.01 刻みで変化させた．

3.2　実 験 結 果

実験の結果を**表 2** に示す．表中の cor は，提案手法によるユーザの専門性スコアとユーザの自己申告とのピアソンの積率相関係数を表している．また，推定精度を測る別の指標として，平均絶対誤差（MAE）についても調べた．

提案手法の推定結果の精度は，λ_U，λ_V の値に大きく影響を受ける．更新式において，λ_U，λ_V はユーザと文書ノードそれぞれにあらかじめ与えた初期値と，伝播によって得られるスコアを足し合わせる割合を表す．λ_U が 1 に近い値であるほど，ユーザノードの持つ専門性スコア x_i が伝播によって更新する量が小さくなり，0 に近いほど伝播によるスコアの更新量は大きくなる．$\lambda_U = 1$ であるとき，x_i は更新されないため，cor は初期値 x_i^0 と自己申告との相関係数である．つまり，初期値には従来手法による推定結果を用いたため，従来手法の推定結果との相関係数である 0.528 となっている．一方，λ_V が 1 に近い値であるほど，文書ノードの持つ専門性スコア y_j の更新量は小さくなり，0 に近いほど伝播によるスコアの更新量は大きくなる．

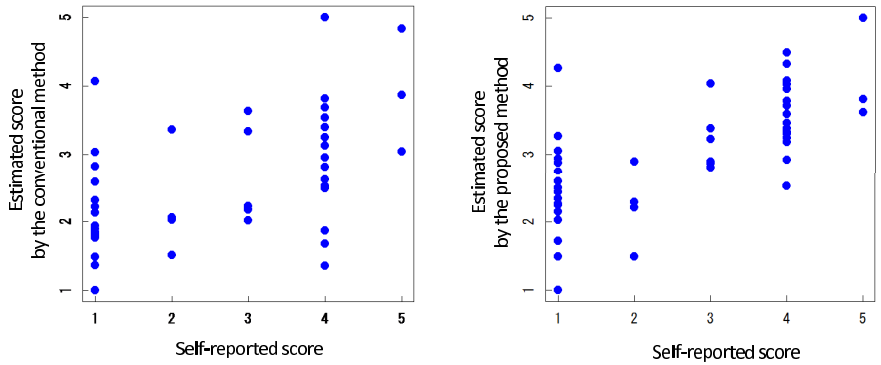

図 3　従来手法と提案手法による推定結果の散布図（$\lambda_U = 0.15$，$\lambda_V = 0.47$）
Fig. 3　Estimated score of each user node ($\lambda_U = 0.15$, $\lambda_V = 0.47$).

λ_U，λ_V を 0 から 1 の範囲で，それぞれ 0.01 刻みに変化させると，$\lambda_U = 0.15$，$\lambda_V = 0.47$ であるとき最も自己申告との相関が高くなった．最も高い相関係数は 0.691 であり，また，このときの平均絶対誤差も小さくなっていることから従来手法より精度が向上しているといえる．最も推定精度が高い場合の λ_U，λ_V を用いて，個々のユーザに関しての推定結果を従来法と比較したものが**図 3** である．

この結果より，最も良いパラメータの組合せにおいて，提案手法によりユーザの専門性推定精度が向上していることが分かる．しかし，それぞれの自己申告の値ごとに散布図を見てみると，専門性の自己申告を 1 としているユーザについて散布図からは推定結果の向上は見られない．実際に，自己申告を 1 としているユーザのみについて MAE を求めると従来手法が 1.45 なのに対し，提案法は 1.12 であり約 0.3 ポイント悪化していた．これは，我々の予想に反して，自己申告を 1 としたユーザが必ずしも専門性スコアの低い文書をクリックしていなかったためであると考えられる．実際，今回の実験においては，自己申告を 1 としたユーザのクリックした文書の専門性スコアは，自己申告を 2 としているユーザがクリックした文書よりも高かった．このようなユーザの挙動の理由として，専門性の低いユーザは，自分の持つ専門性にあった文書を選択することが難しく，様々な文書を選択してしまったからではないかと考えられる．このことは，文書の閲覧時間などの情報を収集し，短時間しか閲覧されていない文書は利用しないなどの方法によって，さらに推定精度が向上する可能性があることを示唆している．

次に，実際に λ_U，λ_V の値によって，推定精度がどのように変化をしているかを見るため

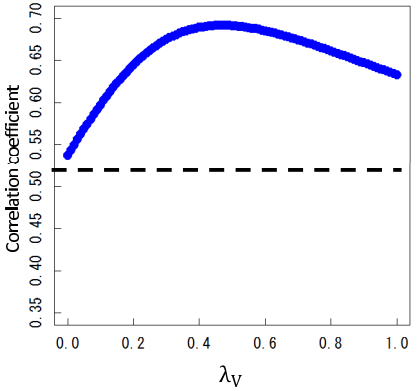

図 4　相関係数の変化（$\lambda_U = 0.15$）
Fig. 4　Correlation coefficient ($\lambda_U = 0.15$).

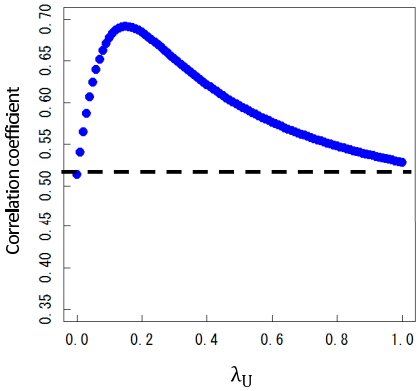

図 5　相関係数の変化（$\lambda_V = 0.47$）
Fig. 5　Correlation coefficient ($\lambda_V = 0.47$).

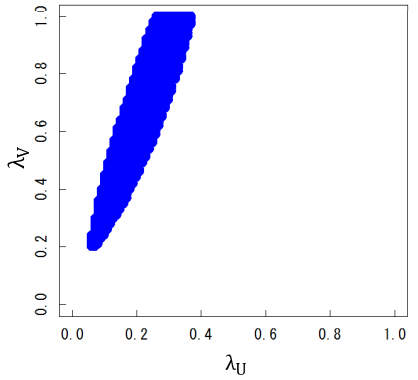

図 6　λ_U，λ_V の分布（$cor \geq 0.68$）
Fig. 6　Distribution of λ_U and λ_V ($cor \geq 0.68$).

に，一方のパラメータを固定し，もう一方のパラメータを 0 から 1 まで変化させた場合の相関係数 cor を調べた．固定する値は最も相関が高くなった $\lambda_U = 0.15$，および $\lambda_V = 0.47$ とし，それぞれの結果を**図 4**，**図 5** に示す．λ_U を動かした場合，λ_V を動かした場合ともに，1 カ所のピークをとる曲線であり，どちらの値も推定結果に影響を及ぼすことが確認できる．図中の破線は，従来手法による精度を示している．λ_U が 0 に近い値をとる場合を除

き，λ_U，λ_V どちらを動かした場合でも，全体にわたり従来法を上回る相関の高さであり，推定精度が向上していることが分かる．

次に，高い推定精度が出ているとき，λ_U，λ_V がどのような値をとっているかを調べた．**図 6** に相関係数 cor が 0.68 以上となる λ_U，λ_V の散布図を示す．図 6 を見ると，λ_U は 0.06 から 0.37 の範囲においてのみ推定精度が高くなっているのに対し，λ_V では 0.2 から 1.0 と広い範囲にわたって高い精度をとる λ_U，λ_V の組合せが存在している．今回の実験データの場合においては，ユーザの持つ専門性に関する推定を高い精度で行うためには，λ_U の値の決定が重要であるといえる．また，2 つの λ の値がともに 0 より大きい場合に高い推定精度がでていることが分かる．今回のデータでは，ユーザおよび文書ノードそれぞれに与えた初期値を利用することにより，高い推定精度を実現できたものと考えらる．

4. 関 連 研 究

文書の作成者の特定分野に関する専門性の推定を行うという研究がなされている[14]．ほかにも，ソーシャルネットワークの中において専門的な人物を特定するという試み[12]や，文書中で言及されている人物の中から，ある分野の専門的な人物を探し出すという試みもなされている[2]．これらの研究は，専門家を探し出すことを目的としており，コンテンツ提供者や文書中に登場する人物などの，一般ユーザから検索される対象人物が専門家の候補となる．

本研究では，一般検索ユーザの専門性に着目している．検索ユーザの持つ専門性については，検索行動分析において多くの研究がなされており，専門性と検索行動の間に関係があることが指摘されている．Wildemuth[11]は微生物学を専攻している学生を集め，調査時期を変えることによって専門性の違いと検索行動との関係を調べている．調査時期が後であるほど，学生の，微生物学に関する専門性が高くなっているとし，専門性の低い初期の調査時期においては，検索効率は低いという結果となっている．Duggan ら[4]は音楽分野とサッカー分野に関して，専門性と検索効率の関係について調べている．分野によって結果は異なるが，専門性のあるユーザのほうが効率的であるという結果が得られている．Zhang ら[13]はエンジニアリングの専門知識と検索行動の関係を調べている．実験では，専門性の高いユーザグループのほうが，低いグループよりも，1 つのタスクあたりに多くのクエリを入れており，さらに文字数の多いクエリをより入力しているという結果を報告している．

ユーザの専門性によって，クエリに違いが生じるという同様の指摘がいくつかの研究でなされている．Vakkari[8]は，学生を対象として，調査時期を変えることによって専門性の違いの影響を調べている．専門性が高くなるとその分野に特有の語彙をクエリとして多く入力することが報告されている．Hembrooke ら[5]は，ユーザの自己申告の専門性と，検索クエリの関係性について調べている．専門性の高いユーザは，低いユーザに比べより長く複雑なクエリを入力する傾向にあることを指摘している．また，クエリだけでなく，Bhavani[1]は，文書の選択が専門性によって大きく異なることを指摘している．

検索行動に大きな影響を与える要因の 1 つとして，ユーザの検索スキルがあげられる．専門性と検索スキルを分離し，それぞれの影響を調べる研究がいくつか行われている．Hölscher ら[6]は，被験者の検索スキルと専門性それぞれについて検索行動への影響を調べており，どちらも検索行動に影響を与えるとしている．中島ら[17]も検索スキルと専門性がユーザの検索行動に与える影響を調べている．彼らは，近年の検索エンジンの改善により，検索スキルよりも専門性が，検索効率に与える影響が大きくなってきていることを指摘している．

White ら[9),10)]は，一般の検索ユーザのデータを使うことにより，従来の研究に比べ大規模かつ，現実世界での自然な検索行動の分析を行っている．分析結果から，専門性によってクエリや文書の選択に大きな違いが生じていることが確認された．さらに，分析をふまえて，検索行動に基づいて専門知識を予測するモデルを作成している．彼らはログ収集のための専用のプラグインから得られる，ページ閲覧履歴，閲覧時間なども含む豊富なログを用い，教師あり学習を行うことで，ログを残したユーザが専門的かそうでないかという 2 値分類を行うモデルの作成を行っている．その際の教師データとして，ユーザがある特定の Web ページ（専門的な内容であると判断されたページ）を閲覧しているかどうかという情報を用いている．我々の提案手法では，このような教師データを必要としない．

本研究での提案手法は，二部グラフのリンク解析手法である Co-HITS[3]を参考にしている．Co-HITS は，各ノードに与えられた初期値を考慮したリンク解析手法である．二部グラフの 2 つの集合に与えられた初期値を，エッジを通して相互に伝播させるという手法であり，クリック関係でつながるユーザと文書が，お互いの持つ情報を相互に利用するというアプローチに適しており参考とした．Co-HITS と提案手法の大きな違いは，手法が扱う値の性質が異なるという点である．Co-HITS では，ノードに与えられる値は確率である．しかし，我々の扱いたい値は専門性スコアであり，更新によって総量は一定に保たれるという制約がない．また，確率を扱う場合は，多くのエッジを持つノードに大きな値が与えられるのに対し，専門性スコアを扱う際は，エッジの数による影響を受けないよう，エッジの数で正規化する必要があるなどの点で異なる．エッジの数で正規化することにより，提案手法での伝播による更新量は，エッジでつながっているノードの持つ値の平均値となる．付録で詳

細を述べているように，2 つの λ の値が 0 のとき，更新を繰り返すことによって，ユーザノードの値は平均化され，同じ値に収束する．2 つの λ の値が 0 でない場合は，更新の際に，初期値も加えられることになるため，各ノードの値は初期値の影響を受けた値に収束される．2 つの λ の値は初期値の影響の強さを表すパラメータであるが，Co-HITS においてはエッジ数の影響の強さを，提案手法においては，ノードの値を平均化する強さを調節するパラメータであるということもできる．

5. おわりに

本稿では，検索システムを利用するユーザの，検索対象に関する知識の程度を推定する手法の提案を行った．提案手法は，検索エンジンの持つクリックスルーログを利用し，ユーザと文書とのクリック関係をエッジをとする二部グラフとして扱う．二部グラフの初期値として与えた文書とユーザの持つ専門性が，二部グラフのエッジを通して相互に伝播するモデルである．

提案法による推定精度を評価するために実験を行った．ユーザの専門性はユーザの自己申告により取得した．実験によって，更新式中の 2 つのパラメータが提案手法の推定精度に大きな影響を与えることが分かった．我々が過去研究で提案した，ユーザのクエリログを用いた推定手法との比較の結果，2 つのパラメータのほとんどの組合せにおいて精度が向上していることが確認できた．このことより，ユーザの自己申告の専門性を高精度に推定するという点において，提案手法の有効性が確認できた．また，λ_U，λ_V の値を適切に決めることができれば，高い精度で推定ができる可能性があることを確認した．

今回の実験では 47 人の被験者による検索履歴データを用いた．データ量は多いとはいえないが，手法による推定精度の向上がみられた．提案法では，ノードの持つエッジ数が多いほど推定精度の向上が期待できる．実応用でも，多くのユーザのデータを用いることができれば十分に効果が期待できる．

謝辞　本稿の執筆において，匿名の査読者からのコメントを大変参考にさせていただきました．特に，付録は査読者のコメントによるものであり，提案手法の収束性に関して，内容を充実させることができました．ここに深く感謝の意を記します．

参考文献

1) Bhavnani, S.K.: Domain-specific search strategies for the effective retrieval of healthcare and shopping information, *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, *CHI '02*, pp.610–611, New York, NY, USA, ACM (2002).
2) Demartini, G.: Finding experts using wikipedia, *Proc. Workshop on Finding Experts on the Web with Semantics* (*FEWS2007*), *Busan, South Korea*, pp.33–41 (2007).
3) Deng, H., Lyu, M.R. and King, I.: A generalized Co-HITS algorithm and its application to bipartite graphs, *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD '09*, pp.239–248, New York, NY, USA, ACM (2009).
4) Duggan, G.B. and Payne, S.J.: Knowledge in the head and on the web: using topic expertise to aid search, *Proc. 26th Annual SIGCHI Conference on Human Factors in Computing Systems*, *CHI '08*, pp.39–48, New York, NY, USA, ACM (2008).
5) Hembrooke, H.A., Granka, L.A., Gay, G.K. and Liddy, E.D.: The effects of expertise and feedback on search term selection and subsequent learning: Research Articles, *J. Am. Soc. Inf. Sci. Technol.*, Vol.56, pp.861–871 (2005).
6) Hölscher, C. and Strube, G.: Web search behavior of Internet experts and newbies, *Comput. Netw.*, Vol.33, pp.337–346 (2000).
7) Manning, C.D., Raghavan, P. and Schütze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2008).
8) Vakkari, P.: Subject Knowledge, Source of Terms, and Term Selection in Query Expansion: An Analytical Study, *Proc. 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*, pp.110–123, London, UK, Springer-Verlag (2002).
9) White, R.W., Dumais, S. and Teevan, J.: How medical expertise influences web search interaction, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *SIGIR '08*, pp.791–792, New York, NY, USA, ACM (2008).
10) White, R.W., Dumais, S.T. and Teevan, J.: Characterizing the influence of domain expertise on web search behavior, *Proc. 2nd ACM International Conference on Web Search and Data Mining*, *WSDM '09*, pp.132–141, New York, NY, USA, ACM (2009).
11) Wildemuth, B.M.: The effects of domain knowledge on search tactic formulation, *J. Am. Soc. Inf. Sci. Technol.*, Vol.55, pp.246–258 (2004).
12) Zhang, J., Tang, J. and Li, J.: Expert finding in a social network, *Proc. Database Systems for Advanced Applications*, *DASFAA2007* (2007).
13) Zhang, X., Anghelescu, H. and Yuan, X.: Domain knowledge, search behavior, and search effectiveness of engineering and science students: An exploratory study, *Inf. Res.*, Vol.10, No.2, p.217 (2005).

14) 加藤義清，乾健太郎，黒橋禎夫：帰属文書数に基づく Web ページ情報発信者の専門性分析，Vol.2010-IFAT-99, No.3 (2010).
15) 齋藤邦子，鈴木　潤，今村賢治：CRF を用いたブログからの固有表現抽出，言語処理学会第 13 回年次大会 (2007).
16) 佐藤大祐，安田宜仁，望月崇由，鈴木智也，松浦由美子，片岡良治：検索システムユーザの分野別の知識推定，第 2 回データ工学と情報マネジメントに関するフォーラム (2010).
17) 中島　悠，土方嘉徳，西田正吾：検索経験と領域知識の WWW 情報検索行動に与える影響，情報処理学会研究報告，HI-108, pp.25–32 (2004).
18) 中谷　誠，アダムヤトフト，田中克己：理解容易性を考慮した用語説明のランキング手法，*WebDB Forum2009* (2009).

付　　録

ある行列 $\mathbf{A}$ に対して，2 つの行列 $\mathbf{W}_{UV}$，$\mathbf{W}_{VU}$ を以下のように定義する．

$$w_{ij}^{UV} = \frac{a_{ij}}{\sum_{i\in U} a_{ij}}, w_{ij}^{VU} = \frac{a_{ji}}{\sum_{j\in V} a_{ij}}. \tag{11}$$

以下の漸化式から $\mathbf{x}$ と $\mathbf{y}$ を求める．

$$\mathbf{x}^k = \lambda_U \mathbf{x}0 + (1-\lambda_U)\mathbf{W}'^{VU}\mathbf{y}^{k-1}, \mathbf{y}^k = \lambda_V \mathbf{y}0 + (1-\lambda_V)\mathbf{W}'^{UV}\mathbf{x}^{k-1}. \tag{12}$$

ただし，$\mathbf{W}'$ は $\mathbf{W}$ の転置行列である．

特に $\mathbf{x}^k$ に関して変形すると，以下の式が得られる．

$$\mathbf{x}^k = \lambda_U \mathbf{x}0 + (1-\lambda_U)\lambda_V \mathbf{W}'^{VU}\mathbf{y}0 + (1-\lambda_U)(1-\lambda_V)\mathbf{W}'^{VU}\mathbf{W}'^{UV}\mathbf{x}^{k-2} \tag{13}$$

さらに $\lambda_U = 0$，$\lambda_V = 0$ のとき，

$$\mathbf{x}^k = \mathbf{W}'^{VU}\mathbf{W}'^{UV}\mathbf{x}^{k-2}, \tag{14}$$

となる．

以下では，$\mathbf{W} \equiv \mathbf{W}'^{VU}\mathbf{W}'^{UV}$ として，$\lambda_U = 0$，$\lambda_V = 0$ のときに，十分に大きい k に対して $\mathbf{x}^k$ のすべての要素が等しくなることを示す．すなわち，$x_1^k = x_2^k = \ldots$ を示す．ただし，行列 $\mathbf{W}$ のすべての固有ベクトルの重複度が 1 であることを仮定している．

定理 1. 行列 $\mathbf{W}$ の行ベクトルの和は 1 である．すなわち，$\sum_j w_{ij} = 1$ である．

証明　$\mathbf{W} = \mathbf{W}'^{VU}\mathbf{W}'^{UV}$ より，

$$\begin{aligned}\sum_j w_{ij} &= \sum_j \sum_k w_{ki}^{VU} w_{jk}^{UV} = \sum_k \left(w_{ki}^{VU} \sum_j w_{jk}^{UV} \right) \\ &= \sum_k w_{ki}^{VU} = 1.\end{aligned} \tag{15}$$

□

定理 2. 十分に大きい k に対して $\mathbf{x}^k$ は行列 $\mathbf{W}$ の最大の固有値 λ_1 に対応する固有ベクトル $\mathbf{x}_1$ の定数倍に収束する．

証明　$|U| \times |U|$ 行列 $\mathbf{W}$ の固有値が $\lambda_1, \lambda_2, \ldots$ とすべて異なる場合で，絶対値の大きい順に並べたとする．これに対する正規直交固有ベクトルをそれぞれ $\mathbf{x}_1, \mathbf{x}_2, \ldots$ とする．任意の $|U|$ 次元ベクトル $\mathbf{x}0$ は正規直交固有ベクトル $\mathbf{x}_1, \mathbf{x}_2, \ldots$ を使って次のように表される．

$$\mathbf{x}0 = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots. \tag{16}$$

これを用いると十分大きい k に対する $\mathbf{x}^k$ は以下のように表される．

$$\begin{aligned}\mathbf{x}^k &= \mathbf{W}^{\frac{k}{2}}\mathbf{x}0 = \mathbf{W}^{\frac{k}{2}}(c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots) \\ &= \lambda_1^{\frac{k}{2}} c_1\mathbf{x}_1 + \lambda_2^{\frac{k}{2}} c_2\mathbf{x}_2 + \cdots \\ &= \lambda_1^{\frac{k}{2}} \left(c_1\mathbf{x}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^{\frac{k}{2}} c_2\mathbf{x}_2 + \cdots \right)\end{aligned} \tag{17}$$

固有値 λ_1 は固有値の中で最大，すなわち，$|\lambda_1| > |\lambda_i|$ であるため，k が十分大きいとき $\mathbf{x}_1$ 以外の項は無視できるほど小さくなる．したがって，$\mathbf{x}^k$ は下記のように収束する．

$$\mathbf{x}^k = \lambda_1^{\frac{k}{2}} c_1\mathbf{x}_1. \tag{18}$$

□

定理 3. 行列 $\mathbf{W}$ の最大の固有値は $\lambda_1 = 1$，対応するは固有ベクトルは $\mathbf{x}_1 = c(1, 1, \ldots)'$ である．

証明　Gerschgorin の定理より行列 $\mathbf{W}$ の固有値 $|\lambda_1|$ の上限は下記のように制限される．

$$|\lambda_1| \leq \max\left(\sum_j w_{ij}\right). \tag{19}$$

定理 1 より，$|\lambda_1| \leq 1$ である．$\lambda_1 = 1$ となることを仮定すると，

$$\mathbf{x}_1 = \mathbf{W}\mathbf{x}_1, \tag{20}$$

となる．このとき，行列 $\mathbf{W}$ の行ベクトルの和が 1 であることから，$\mathbf{x}_1 = c(1, 1, \ldots)'$ は上の式を満たすことが分かる．以上より，$\lambda_1 = 1$ は最大の固有値であり，対応するは固有ベクトルは $\mathbf{x}_1 = c(1, 1, \ldots)'$ である．

□

定理 4. 十分に大きい k に対して，行列 $\mathbf{W}$ から得られる $\mathbf{x}^k$ はそのすべての要素が等しくなる．

証明　定理 2 と定理 3 より，十分に大きい k に対して $\mathbf{x}^k$ は行列 $\mathbf{W}$ の最大の固有値 $\lambda_1 = 1$ に対応する固有ベクトル $\mathbf{x}_1 = c(1, 1, \ldots)'$ の定数倍に収束する．したがって，十分に大きい k に対し $\mathbf{x}^k$ は，$x_1^k = x_2^k = \ldots$，そのすべての要素が等しくなる．

□

(平成 23 年 3 月 18 日受付)
(平成 23 年 7 月 5 日採録)

(担当編集委員　中島 伸介)

佐藤　大祐（正会員）

2007 年早稲田大学理工学部機械工学科卒業．2009 年同大学大学院創造理工学研究科総合機械工学専攻修士課程修了．同年日本電信電話（株）入社．現在，NTT サイバーソリューション研究所勤務．情報検索の研究に従事．

安田　宜仁

1997 年京都大学総合人間学部基礎科学科卒業．1999 年同大学大学院人間・環境学研究科修士課程修了．同年日本電信電話（株）入社．2009 年東京工業大学大学院総合理工学研究科単位取得満期退学．現在，NTT サイバーソリューション研究所勤務．音声対話システム，自然言語処理，情報検索の研究に従事．博士（工学）．

小池　義昌

日本電信電話株式会社サイバーソリューション研究所所属．1989 年東北大学大学院材料化学専攻修士課程修了後，日本電信電話株式会社に入社．以来，パターン認識の研究，遠隔教育システムのの研究開発，情報検索サービスの研究に従事．

片岡　良治（正会員）

日本電信電話株式会社サイバーソリューション研究所所属．1987 年千葉大学大学院電子工学専攻修士課程修了後，日本電信電話株式会社に入社．以来，トランザクションの並行処理制御方式の研究，マルチメディア情報システムの研究，ポータルサービスシステムの研究開発に従事．