情報処理学会論文誌　データベース　Vol.7　No.4　13–28 (Dec. 2014)

検索専門性と事前知識に着目した検索行動とタスク満足度の関係性分析

梅本 和俊[1,2,a)]　山本 岳洋[1,b)]　田中 克己[1,c)]

受付日 2014年6月20日, 採録日 2014年10月7日

概要：本稿では，検索タスク実行時のユーザの検索行動と，終了時のタスクに対する満足度との関係性を調査する．従来の情報検索に関する研究では，適合ページを多く提示することがユーザの満足度の向上につながるという前提の下で，検索結果のランキング手法が考案されてきた．しかし，それぞれの適合ページで記述されているタスクの答えが異なる場合は，かえってユーザの不満を引き起こす可能性がある．また，情報検索に対する専門性やタスクに関する事前知識といった属性の有無についても，ユーザの検索行動や満足度の評価基準に影響を与えることが予想される．そこで我々は，事実発見型タスクの検索ログに対して被験者が発見した答えを抽出することでデータセットを作成し，これらの 2 種類のユーザ属性が両者の関係性に与える影響を分析した．その結果，(1) 情報検索の専門知識を持つユーザについては，発見された答えの一貫性と満足度との間に負の相関関係が存在する可能性がある，(2) 情報検索の専門知識を持つユーザは，答えの発見以後も長い時間をかけてタスクに取り組む，および (3) 情報検索の専門知識を持たないユーザは，タスク開始から一定時間が経過した後も，特定の答えに絞り込んだ検索を行わない，という傾向が見られた．

キーワード：検索行動，タスク満足度，検索専門性，事前知識

Analysis of Relationship between Search Behavior and Task Satisfaction Focused on Search Expertise and Prior Knowledge

Kazutoshi Umemoto[1,2,a)]　Takehiro Yamamoto[1,b)]　Katsumi Tanaka[1,c)]

Received: June 20, 2014, Accepted: October 7, 2014

Abstract: In this paper, we investigate the relationship between user behavior observed in search tasks and satisfaction perceived by users in the tasks. Various kinds of methods for ranking search results have been proposed by existing work on information retrieval under the assumption that providing many relevant pages would lead to user satisfaction. As for search tasks where inconsistent answers are found, however, users may feel dissatisfied about the information obtained from the relevant results. As well as the type of search tasks, user attributes such as expertise in information retrieval and prior knowledge on the task could affect search behavior of users and their satisfaction perception. To analyze the effect of the two attributes on the relationship, we extracted answers from each page in search logs of fact-finding tasks. As a result of analysis of this dataset, we found the different tendencies in accordance with the presence or absence of these attributes: (1) finding inconsistent answers may cause dissatisfaction of search experts, (2) search experts still continue to search after finding some answer candidates, and (3) users without search expertise try to search for any answers even in the closing stage of search sessions.

Keywords: search behavior, task satisfaction, search expertise, prior knowledge

[1] 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[2] 日本学術振興会特別研究員（DC1）
JSPS Research Fellow (DC1)
[a)] umemoto@dl.kuis.kyoto-u.ac.jp
[b)] tyamamot@dl.kuis.kyoto-u.ac.jp
[c)] tanaka@dl.kuis.kyoto-u.ac.jp

1. はじめに

情報検索に関する研究分野では，検索結果リスト中の各文書の適合度に長らく重点が置かれてきた．ここで，検索クエリに対する文書の適合度とは，クエリによって表現されたユーザの情報要求が，その文書の閲覧によってどの程度満

たされるかを表す概念である [6]．適合度は，BM25 [26] をはじめとするさまざまなランキングアルゴリズムにおいて中心的位置を占めると同時に，検索結果のランキング評価においても，重要な構成要素と考えられている．たとえば，現在広く用いられている評価尺度の 1 つである normalized Discounted Cumulative Gain（nDCG）[18] は，適合度の高い文書が上位に多く出現しているランキングに対して高いスコアを与える．

しかし，適合度の高い文書を多く閲覧することが，ユーザの最良な検索体験につながるとは必ずしも限らない．その例として，ユーザが「アメリカ史上で最悪の干ばつが発生した年を知りたい」と思い，Web 検索を行うという状況を考える．入力クエリに対する検索結果ページにおいて各結果で記述されている答えが異なる場合，このユーザは得られた答えの信頼性に対して疑問をいだく可能性がある．この場合，各結果の Web ページは，干ばつの発生年に関する記述を含むという意味で適合度が高いにもかかわらず，ユーザは不満を感じたまま検索を終えてしまうかもしれない．

こうした検索結果リストの各文書に対する適合度とユーザの検索体験とのずれの存在から，近年では検索システムをユーザ指向な尺度で評価することに研究者の注目が集まっている．代表的なユーザ指向の評価尺度として満足度 [10] やフラストレーション [9] が存在し，こうした尺度とユーザの検索行動との関係性を詳細に分析している研究 [3] もある．これらの尺度の中でも特に満足度について，その度合いをユーザの検索行動から予測するという研究課題は近年さかんに取り組まれている [10], [13]．その理由として，ユーザの満足度の形成過程を理解することが，検索体験の最大化という検索エンジンの目的に密接に関連していることが考えられる．

ここで，ユーザの満足度には複数の観点が存在する．その 1 つとして「検索を通して情報要求がどの程度満たされたか」という観点があげられる [9]．この観点での満足度は，ユーザの情報要求およびその下で獲得された情報によって決定される．そのため，内容に対する満足度ととらえることができる．一方で，「検索時に利用したシステム自体に対する印象」という別の観点も存在する [32]．前者の満足度と対比した場合，後者はシステムに対する満足度と考えられる．後者の満足度は，検索エンジンの応答速度や信頼性，使いやすさといった側面から構成される．ユーザの検索体験には以上の 2 種類の満足度が深く関与する．先の干ばつに関する検索タスクの例では，最悪の干ばつの発生年を特定できないという理由で内容に対する満足度が低くなり，提示された検索結果が信頼できないといった理由でシステムに対する満足度が低くなる．そのため，検索結果ページ自体はユーザの情報要求に適合しているが，ユーザの検索体験は低くなることが予想される．

適合度と満足度は一見すると似ているが，上述の例のように一方の向上が必ずしも他方の向上を引き起こすとは限らない．これまでに，両者の関係性の理解を目的としたいくつかの研究がなされてきた．文献 [15] や文献 [28] では，Text REtrieval Conference（TREC）の interactive track を利用したユーザ実験によって，再現率重視のタスクにおいては適合度とユーザ実験指標の間に相関がないことが報告されている．その一方で，文献 [2], [17] では，検索結果リスト上位の文書集合の適合度とユーザの満足度との関係性が分析されており，特に検索クエリが navigational な意図 [5] を表す際に両者の間に強い正の相関関係が成立することが判明している [17]．

このように，満足度の形成過程はタスクの種類によって大きく異なることが予想される．同様に，ユーザの検索戦略についてもタスクの種類に影響を受けることが知られている [34]．また，ユーザが備え持つ属性が検索行動に影響を及ぼすという報告も存在する．たとえば文献 [30] によると，検索ドメインに対する知識を持つユーザはそうでないユーザに比べて当該ドメインでの検索に成功しやすい傾向にある．また別の文献 [29], [33] では，検索トピックに対する知識量や先入観が，得られた情報に対する判断基準に影響を与えると述べられている．こうした既存研究の知見をふまえると，検索の上手さや情報の受け止め方に関するユーザの属性は，最終的なタスクの満足度に影響を与える可能性がある．

本研究では，情報検索に対する専門性とタスクに関する事前知識という 2 種類のユーザ属性を考慮し，それらがユーザの検索行動や満足度の評価基準に与える影響を分析する．上述の干ばつに関する検索タスク例のように，ユーザがタスク実行中に多種多様な答えに遭遇する場合，適合度と満足度との間には単純な相関関係が成り立たない可能性がある．そこで本稿では，事実発見型 [21] の検索タスクのうち，こうした矛盾する答えが複数存在するものを分析の対象とする．我々はこの種類の検索タスクにおいて，(1) タスク実行中に発見された答えの数，(2) タスクの達成に要する時間，および (3) 発見される答えの時間経過による変化，の 3 点に着目して，両者の関係性を分析する．Web 上で入手可能な事実発見型タスクの検索ログに対して，閲覧ページに含まれるタスクの答えを抽出し，上記観点に基づいた分析を行った結果，検索専門性と事前知識の有無に応じて，ユーザの検索行動や満足度の評価基準は以下のように異なるという傾向が確認された．

- 情報検索の専門知識を持つユーザについては，発見された答えの一貫性と満足度との間に負の相関関係が存在する可能性がある．
- 情報検索の専門知識を持つユーザは，答えの発見以後も長い時間をかけてタスクに取り組む．
- 情報検索の専門知識を持たないユーザは，タスク開始

から一定時間が経過した後も，特定の答えに絞り込んだ検索を行わない．

本稿の構成を以下に記す．2 章では関連研究について述べる．3 章では分析に利用するデータセットについて説明する．このデータセットに対して，検索行動と満足度の関係性を 4 章で分析する．そこで得られた結果をもとに 5 章では，検索支援のあり方と分析の限界点について考察する．最後に 6 章で本研究のまとめと今後の課題を述べる．

2. 関連研究

2.1 検索結果の適合度評価

検索結果リストの定量的な評価のために，nDCG や Expected Reciprocal Rank（ERR）[8] など，さまざまな指標がこれまでに提案されてきた．これらの評価指標の多くは，次式による解釈が可能といわれている [7], [27]．

$$\frac{1}{\mathcal{N}}\sum_{k=1}^{\infty} g(k)d(k).$$

ここで $g(k)$ は検索結果リストの k 位の文書の閲覧によって得られる利得，$d(k)$ は k 位の文書の閲覧に対するディスカウント比を意味する．また $\mathcal{N}$ は評価尺度の正規化のための項である．検索結果リストの下位文書の閲覧は，ユーザにとってコストの高い行為といえる．そのため，$d(k)$ は k に関する単調減少関数となるように定められることが多い．たとえば nDCG では $1/\log_2(1+k)$，ERR では $1/k$ という項が $d(k)$ に相当する．利得に関しては，尺度によって多少の違いはあるものの，k 位の文書の適合度がその値に大きく関係する．これらの尺度は，ユーザが検索結果リストを上位から順番に眺めることを暗黙的に仮定しており，上位に多くの適合文書が出現するランキングを高く評価する．

タスクの中には複数回の検索が必要なものも存在する．こうした複数クエリに対する検索結果リスト集合を含むセッションに対して，検索システムの有用性を評価しようという試みも存在する．Järvelin ら [19] はその評価尺度として session-based DCG（sDCG）を提唱している．名前から分かるように，sDCG は評価対象がセッションとなるよう DCG を拡張したものであり，セッション序盤での検索結果リストの上位に適合文書が多く出現するほど高い評価値を与える．Kanoulas ら [20] は sDCG と同様の考えの下で，ユーザの文書閲覧行動およびクエリ修正行動を幾何分布によってモデル化したセッション評価指標を提案している．

上述の評価指標はすべて，ユーザが検索の早い段階で閲覧した文書の適合度を重視して，検索システムの有用性を計算する．しかし，セッション序盤に閲覧された適合文書が最終的なタスクの満足度に大きな影響を及ぼすとは限らない．この点に関して本稿 4.3 節では，ユーザが適合文書を閲覧した時期とタスク満足度との関係性を分析する．

2.2 ユーザ指向な評価尺度

上述の nDCG や ERR は，検索結果リストの有用性を適合度という観点で定量的に評価するためのものであり，システム指向な評価尺度といえる．これに対して近年では，満足度 [12], [23] や成功度 [11]，フラストレーション [9] といったユーザ指向な評価尺度に注目が集まっている．その中で，ユーザの検索行動をセッション単位で評価することに取り組んでいる研究を以下で述べる．Hassan ら [11] は，検索時のユーザ行動をマルコフモデルで表現し，検索タスクに対する成功/失敗を予測する手法を提案している．Guo ら [10] の提案手法は，検索クエリやページ閲覧時間，カーソル操作といったさまざまな特徴量から構成されたロジスティック回帰モデルによって，タスクに対する満足度をユーザの検索行動から予測する．

ユーザ指向な評価尺度を検索行動から予測する研究が近年さかんに行われている一方で，その尺度とシステム指向な尺度との関係性を詳細に分析した研究 [2], [17] は少ない．本研究では，満足度やフラストレーションの予測に利用された検索ログ [9] に対して，そこに含まれる各閲覧ページの適合度を判定し，満足度と適合度の関係性を分析する．

2.3 ユーザ属性が検索行動に与える影響

ユーザの備え持つ属性である専門性が彼らの検索行動に影響を及ぼすことが過去の研究で示されてきた．ここで専門性には，（1）検索を行う対象に関するユーザの習熟度や専門知識（ドメイン専門性），および（2）検索エンジンの仕組みに関する理解や，検索戦略の立て方の上手さ（検索専門性），の 2 種類が存在すると考えられる．

Hembrooke ら [14] は，検索ドメインに対する知識が検索クエリ生成に及ぼす影響を調査している．彼らはユーザ実験を通して，ドメイン知識のあるユーザによって生成されるクエリは語彙が豊富であることを発見している．Hölscher ら [16] は，Web 上での検索に習熟しているユーザが生成する検索クエリの特徴を分析しており，Web 検索への習熟度が高いユーザは一般ユーザに比べ，クエリ生成時にブール演算子などの高度な検索オプションを使う割合が高いと報告している．Yamamoto ら [33] は大規模アンケートを行うことで，検索ドメインに対する知識の有無によって情報の信憑性判断時にユーザが重視する評価基準が異なることを示している．

このように，検索時にユーザがとる戦略やその下で得られた情報に対する受け止め方は，ユーザの属性によって異なる．同様に，ユーザ属性の有無は検索タスク終了時の満足度にも影響を及ぼす可能性がある．そこで本研究では，検索専門性として情報検索分野に関する専門知識を，ドメイン専門性としてタスクに関する事前知識を考慮し，これら 2 種類のユーザ属性が満足度と適合度の関係性に与える影響を調査する．

3. データセット

3.1 ユーザ実験

検索タスクに対する最終的な満足度と，そのタスクの実行中に発見された答えの関係性を分析するために，本稿では Feild ら [9] が実施したユーザ実験における検索ログを利用する．この実験は，タスク実行時のユーザのフラストレーションを検索行動から予測するために実施されたものである．そのため実験用に用意された 12 個のタスクはすべて，その答えが単一のページからは見つかりにくく，簡単には達成できないものとなっている．

この実験には，大学に所属する 30 名が被験者として参加しており，各被験者の専攻は計算機科学，工学，運動生理学，経済学，文学などと多岐にわたっている．各被験者の属性として，情報検索に関する大学院レベルの研究について専門的な知識があるか（*is_ir*）という情報がブール値で記録されている．各被験者には上述の 12 個のタスクの中から 7 個がラテン方格法によって割り当てられており，各タスクの開始前および終了後に次の質問に答えることが求められている．タスク開始前の質問は，今から行うタスクの答えをどの程度事前に知っていたか（*knowledge*），終了後の質問は，タスク中でのすべての検索行動を通して元々の情報要求がどの程度満足されたか（*satisfaction*）というものであり，それぞれの質問に対して被験者は 5 段階のリッカート尺度で回答している．これらのフィードバックに加えて，検索行動を通して被験者が最終的に判断したタスクの答え（*reported_answers*）に関するメモも残されている．そのほかにも，タスク実行中に入力された検索クエリや，閲覧されたページの内容（URL および HTML）などといった情報が検索行動ログとして残されており，それぞれのイベントの発生時刻も記録されている．このユーザ実験に関する詳細な情報およびデータセットは，Web 上で取得可能である[*1]．

3.2 対象タスク

上述のユーザ実験のために準備されたタスクは，与えられた条件を満たす事実発見型 [21] に設計されており，その多くは次の 2 種類に分類することができる．一方は，「2007 年以降に崩壊したアメリカの橋を 3 つあげなさい」のような複数の答えを要求するタスクである．他方は，「2008 年に最も売れたテレビのブランド名とモデル名は何か？」といった 1 つの答えを要求するタスクである．前者のタスクについては，その説明から複数の答えの存在が予想可能なのに対して，後者については，ユーザに唯一の解を期待させうる内容となっている．そのため，発見された答えの内容が最終的な満足度に大きな影響を与える可能性がある．

[*1] http://hank.feild.net/downloads.html

表 1　ユーザ実験 [9] の被験者に提示されたタスクの説明文の日本語訳

Table 1　Task descriptions given to subjects of user study [9] (translated into Japanese).

タスク	説明文
Drought	アメリカ史上最悪の干ばつが起こったのは何年か？また，その年の同国の平均降水量は？
Pixels	Apple が MacBook の交換に応じるには最低何個のドット落ちが必要か？ただし，MacBook は保証期間中と仮定する．
TV	2008 年に最も売れたテレビのブランド名とモデル名は何か？
Verizon	マサチューセッツ州の Verizon Wireless の電話相談サービスの番号は？

表 2　各タスクから抽出された答えの一部，ページの適合度判定の一致度，および各タスクのセッション数と平均ページ数

Table 2　Examples of found answers, agreement on page relevance between two assessors, and size of each task.

	Drought	**Pixels**	**TV**	**Verizon**
抽出された答えの例	1930 to 1931	Any pixels	Samsung	800-922-0204
	1950s	5 pixels	LN32B460	800-899-4249
	1988 to 1989	Case by case	LN52A650	800-256-4646
	2001 to 2003	No public policy	Sony	1-800-VERIZON
答えの総種類数	33	17	5	7
一致度	0.75	0.52	0.78	1.00
セッション数	16	14	19	17
平均ページ数	10.25	5.86	8.63	7.29

そこで本研究では，上述のユーザ実験に関する検索ログの中から後者のタスクに属するものを選択し，満足度と発見された答えとの関係性の分析対象とした．

上述の検索ログに含まれる 12 個のタスクのうち，この条件を満たすものは 4 個存在した．これらの各タスクを実行する際に被験者が提示された説明文を**表 1** に示す．本稿では説明の便宜上，これらのタスクをそれぞれ *Drought*，*Pixels*，*TV*，および *Verizon* と名付ける．これらのタスクについて，後述の処理によって抽出された答えの具体例とその総種類数[*2]を**表 2** の 2 行目と 3 行目に示す．これらを見れば分かるように，一見すると一意な解が存在すると思われる各タスクに対して，一貫性のない複数の答えを被験者が実際に発見している．同表には，各タスクのセッション数やセッション中に閲覧された平均ページ数といった指標に関する値も示されている．なお本稿では，1 人のユーザが 1 つのタスクにおいて行った一連の検索行動を 1 セッションとして扱う．分析対象のデータには合計で 66 個のセッションが存在しているため，各被験者は平均で 2 個以上のタスクを実行したことになる．また，セッション中の平均閲覧ページ数が 5 ページを超えていることから，タス

[*2] 答えの総種類数については，3.4 節の処理を事前に適用し，表記は異なるが同一の内容を指す記述を 1 種類の答えと見なして，その値を算出した．

クの達成が容易ではないことが予想される．

3.3　発見された答えの抽出

分析対象の検索ログは，被験者が最終的に判断したタスクの答えに関する情報を含んでいる一方で，彼らがタスク実行中に閲覧ページから発見したすべての答えに関する情報までは記録されていない．そこで各被験者が発見したすべての答えを得るために，本稿の著者のうちの2名がそれぞれ独立に，各被験者がセッション中に閲覧した各ページからタスクの答えに該当する記述の抽出を行った．なお，答えに関する記述がページ中に複数存在する場合には，それらすべての記述を抽出した．ただし検索結果ページについては，必ずしもすべてのタイトルやスニペットをユーザが見るとは限らず，またそこからのみでタスクの答えを決定することは難しいと思われるため，答えの抽出対象からは除外した．また，ページ中の記述が表1中のタスクの情報要求を直接満たすことを，答えの抽出基準として設定した．

表2に，上述の手続きによるタスクの答えの抽出の評価者間での一致度を示す．評価の一致度のための指標にはカッパ係数を利用し，同一のページから2名の評価者がともに何らかの答えを抽出した（あるいは何も抽出しなかった）場合を評価が一致したものとして扱った．全タスクの閲覧ページ集合を対象とした際の答え抽出の一致度は0.78であり，文献[24]によると，これはsubstantialな一致といえる．また，Pixelsタスクを除く3つの各タスクについても同等あるいはそれ以上の一致となった．一方で，Pixelsタスクに関する一致度は0.52であり，同文献によるとmoderateな一致となった．

一致度の低いタスクが存在した要因として，評価者の母語（日本語）と評価対象のページ言語（英語）の違いがあげられる．表2から分かるように，TVタスクやVerizonタスクの答えに関する記述には，テレビの型番や電話番号といった特徴的なパターンが含まれるため，母語と異なる言語であっても比較的目に付きやすい．しかし，Pixelsタスクの答えの場合は，記述方法に規則性がない，もしくは他の内容と区別がしにくいといった理由で，たとえページ中に答えに関する記述が存在していても気付かれない可能性がある．そのため，どちらか一方の評価者が答えに関する記述を見落とすという事例が少なからず発生し，同タスクに関する一致度の低下を引き起こしたと考えられる．

こうした見落としによる答え抽出の欠如に対応するため，各ページについて2名の評価者のうち少なくとも一方が何らかの記述を抽出している場合は，その記述を同ページに存在する答えに関する記述の正解値として用いることにする．

3.4　表記揺れへの対応

同じ答えに対する記述であっても，ページによってその表記方法は異なることがある．たとえば，Verizonタスクにおける評価者の抽出結果には，“(800) 922-0204”と“800-922-0204”という記述が含まれていた．これらは表記は異なるものの，同じ電話番号を指しているため，同一の答えを表す1つのエンティティに関する記述として扱われるべきである．他にも，答えに関する記述として前後の文脈も含めた文章が抽出されている事例も見受けられた．そこで，前節の適合度判定によって得られた答えに関する記述集合に対して，そこに含まれる答えのエンティティを手作業で抽出した．以降では，この抽出作業によって得られたタスクの答えのエンティティ集合を*found_answers*と表記する．

3.5　分析に用いる指標と属性

上述の各処理によって得られた次のデータを4章での分析に利用する．各セッションの評価指標として，本研究では次の3種類に着目する．タスク満足度（*satisfaction*），および報告されたタスクの答え（*reported_answers*）は，ユーザ実験の各被験者からのフィードバックによるものである．残りの指標である発見されたタスクの答え（*found_answers*）は，各被験者が閲覧したページ集合から，実験関係者ではない2名の評価者が抽出したものである．なお，本研究で対象とするタスク満足度は，1章で述べた複数の概念のうち，内容に対する満足度に相当する．これは，3.1節で述べた定義からも明らかである．

ユーザが備え持つ属性は検索の仕方や得られた答えに対する考え方に大きな影響を与える可能性がある．そこで情報検索に対する専門知識（*is_ir*），およびタスクに関する事前知識（*knowledge*）という2種類のユーザ属性を，各セッション評価指標およびそれらの関係性に影響を与えうる要因として分析時に考慮する．前者は2.3節で取り上げた検索専門性のうち，検索エンジンの仕組みに関する理解に該当する属性である．一方で後者の属性は，検索対象への習熟度および専門知識を表すドメイン専門性に深く関連するといえる．

以降では，$is_ir = \mathrm{TRUE}$である場合を検索専門性のあるユーザ，そうでない場合（$= \mathrm{FALSE}$）を検索専門性のないユーザ，と表現する．また*knowledge*に関しては，タスクの答えをある程度知っている場合（> 1）を事前知識のあるユーザ，そうでない場合（$= 1$）を事前知識のないユーザ，として表現を統一する．

4.　満足度と適合度の関係性分析

ユーザ属性の各値に対するセッション数とセッション間の平均満足度（および標準偏差）を**表3**に示す．分析対象の66セッションのうち，検索専門性のあるユーザ（$is\_ir = \mathrm{TRUE}$）のセッションは17個（$\simeq 26\%$），タスクに関する事前知識のあるユーザ（$knowledge > 1$）のセッ

表 3　各ユーザ属性値に対応するセッション数と平均満足度（および標準偏差）

Table 3　Number of sessions and mean satisfaction (with standard deviation) for each user attribute.

	is_ir		*knowledge*		全体
	TRUE	FALSE	>1	$=1$	
セッション数	17	49	9	57	66
satisfaction	3.65 (0.79)	3.47 (1.16)	4.11 (0.78)	3.42 (1.08)	3.52 (1.07)

ションは 9 個（$\simeq 14\%$）存在する．また，事前知識のあるユーザの満足度が他のユーザと比べて高い値をとる傾向にあることも，表より分かる．その要因の 1 つとして，事前知識を利用したクエリ生成やページ選択によって，少ない労力で答えを含むページにたどり着けることが考えられる．このようなユーザ属性が満足度と適合度の関係性に与える影響について，4.1 節では発見された答えの数，4.2 節では答えの発見に費やした時間，4.3 節では答えを発見した時期に着目し，詳細な分析を行う．

以降の分析では，得られた結果に対して検定法を適用することで，仮説の正当性を検証する．検定時の有意水準 α には通常 .05 あるいは .01 といった値がとられることが多い．しかし，本研究の分析対象は個人差の生じうるユーザであり，表 3 に示したように，一部の属性については対応するユーザ数がきわめて少ないといった特徴が存在する．そのため本稿では，$\alpha = .10$ という比較的大きな値を有意水準として採用することにする．有意水準を高くすることの危険性については，得られた結果に対する解釈とともに 5 章で議論する．

4.1　発見された答えの数

検索行動分析に関する既存研究によって，タスクに対する満足度はその達成に要した検索コストに影響を受けることが示されてきた [10]．たとえば，複数回の検索クエリ修正を経てタスクの答えが得られる場合は，1 回の検索で答えが発見できる場合に比べて，ユーザはその検索タスクに対して不満を感じやすくなると報告されている．また，タスクの実行中に閲覧したページ数も満足度に影響することが知られている．本節では，ユーザの専門性の有無がこれらの検索コストと満足度との関係性に与える影響について調査する．

また，本研究で対象とするタスクには，検索の過程で複数の一貫しない答えが見つかるという特徴が存在する．そのため，発見された答えの内容がタスクに対する満足度に影響を与える可能性がある．そこで本節では，検索コストに関する特徴量，および答えの数に関する特徴量とタスク満足度との関係性について，2 種類の専門性の有無が与える影響を分析する．我々は，既存研究 [10] で報告された検索コストとタスク満足度の関係性は，ユーザの属性によらず成立すると推測する．一方で，発見された答えの数がタスク満足度に与える影響については，ユーザ依存であるという予想を立てる．それぞれの予想に対応する具体的な仮説を以下に示す．

H1 すべてのユーザに共通して，検索コストが増加するとタスク満足度は低下する．

H2 ユーザに検索専門性あるいは事前知識がある場合は，発見された答えの数もタスク満足度に影響を与える．

本稿ではユーザが発見した答えの数え方として，(1) 答えの総数，(2) 答えの種類数，および (3) 答えのエントロピー，の 3 種類を考える．各特徴量の計算方法を説明するために，ユーザがセッション内で発見した答えの集合を A，そのセッション内での閲覧ページ集合から 3.3 節の手続きによって答え $a \in A$ が抽出された回数を $m(a)$ と表記する．このとき上記の 3 種類の特徴量はそれぞれ以下の式によって計算される．

$$\sum_{a \in A} m(a), \tag{1}$$

$$|A|, \tag{2}$$

$$-\sum_{a \in A} \frac{m(a)}{\sum_{a' \in A} m(a')} \log_2 \frac{m(a)}{\sum_{a' \in A} m(a')}. \tag{3}$$

特徴量 (1) はユーザによって発見された答えの合計数であるため，1 種類のみの答えが多数発見された場合でもその値は高くなる．一方，特徴量 (2) は発見された答えの種類数を数えるため，このような事例では低い値をとる．特徴量 (3) は，答えの種類数だけでなくその偏りも考慮しており，多種類の答えがそれぞれ同程度の割合で発見されたときに，その値が高くなる．

3.5 節に示したユーザ属性の各値について，該当するセッションに対して上記の特徴量を計算した結果を**表 4** に示す．ここで，表中の各セルの左段は該当する特徴量の平均値（および標準偏差）を，右段はその特徴量とタスク満足度のピアソンの積率相関係数（および無相関検定時の p 値）を表している．

4.1.1　ユーザ属性間で共通の傾向

同表から，すべての属性に共通して，検索コストに関する特徴量とタスク満足度との間の相関係数が負になっていることが分かる．無相関検定の結果，検索専門性のあるユーザを除くすべての場合において，入力クエリ数とタスク満足度との間の相関係数に有意性が確認された（検索専門性なし：$r = -.546$，$p < .001$，事前知識あり：$r = -.764$，$p = .017$，事前知識なし：$r = -.413$，$p = .001$）．一方で閲覧ページ数については，検索専門性のあるユーザについてのみ，タスク満足度との間に有意な相関関係が認められた（$r = -.631$，$p = .007$）．

これらの結果から，検索コストとタスク満足度との間の負の相関関係はすべてのユーザに共通する特徴であるとい

表 4　検索コストおよび発見された答えの数に関する特徴量．各セルの左段は特徴量の平均（および標準偏差）を，右段は特徴量と満足度のピアソンの相関係数（および p 値）を表す

Table 4　Features related to search costs and number of found answers. Left value of each cell is mean (with standard deviation) of feature values, and right one is Pearson's r (with p-value) between feature and task satisfaction.

ユーザ属性		入力クエリ数		閲覧ページ数		答えの総数		答えの種類数		答えのエントロピー	
		平均	相関係数	平均	相関係数	平均	相関係数	平均	相関係数	平均	相関係数
is_ir	TRUE	3.06 (1.39)	−.266 ($p = .303$)	4.88 (2.23)	−.631 ($p = .007$)	4.18 (3.13)	−.380 ($p = .132$)	3.71 (2.76)	−.339 ($p = .183$)	1.69 (0.85)	−.458 ($p = .074$)
	FALSE	2.53 (1.23)	−.546 ($p < .001$)	4.02 (2.38)	−.072 ($p = .625$)	3.53 (2.34)	.114 ($p = .435$)	3.02 (1.84)	.181 ($p = .213$)	1.59 (0.73)	−.021 ($p = .896$)
knowledge	> 1	2.22 (1.09)	−.764 ($p = .017$)	3.44 (2.65)	−.210 ($p = .592$)	3.11 (1.76)	.443 ($p = .232$)	2.78 (1.56)	.432 ($p = .246$)	1.48 (0.72)	−.010 ($p = .981$)
	$= 1$	2.74 (1.30)	−.413 ($p = .001$)	4.37 (2.31)	−.120 ($p = .374$)	3.79 (2.66)	.000 ($p = .998$)	3.26 (2.19)	.050 ($p = .711$)	1.64 (0.77)	−.120 ($p = .400$)

える．このことは仮説 H1 を支持する裏付けと見なせる．しかし，相関関係は因果関係と異なり，変数間の方向性を示すものではない．そのため，「検索コストの高いタスクでは満足度が低くなりやすい」ではなく，「満足度の低いタスクに共通して高い検索コストがかかる」という解釈も可能である．今回の分析からは，どちらの解釈が適切かまでは判断できない．また，本稿で扱う満足度が内容指向であることを考慮すると，「情報要求に対する満足度は，高い検索コストをかけて答えの十分な検証を行うことで向上する」という H1 とは異なる仮説も考えられる．この仮説の支持につながる結果が得られなかった理由の 1 つとして，対象とするタスクでは複数の答えが存在し，それらの真偽判定が容易ではない，ということがあげられる．検索コストの増加は，検証が困難な答えの発見数の増加につながり，そのことが原因で満足度が低くなったものと予想される．

表 4 の結果は，検索コストと満足度の関係性という点では既存研究 [10] と一貫性がある一方で，検索コストの特徴量のうち満足度への影響度が高いものは，ユーザの属性によって異なるという別の知見も示唆している．3.5 節で述べたように，検索専門性のあるユーザは検索エンジンの仕組みを熟知しており，その中には検索クエリの処理に関する知識も含まれる．彼らの満足度と入力クエリ数との間の相関係数に有意性が確認されなかったことに対する 1 つの解釈として，クエリ処理に関する知識の存在により効果的なクエリの生成が可能になり，その結果，多くのクエリを入力しても満足度が低下しなかったという説明が考えられる．

検索コストのユーザ属性間での共通の傾向とは対照的に，答えの数に関する特徴量は，ユーザの属性ごとにそれぞれ異なる結果が得られた．そこで以降では，検索専門性，および事前知識のそれぞれが答えの数とタスク満足度との関係性に与える影響を分析し，仮説 H2 が成立するかについて議論する．

4.1.2　検索専門性の影響

検索専門性に関する属性（*is_ir*）の影響として，専門性のあるユーザ（*is_ir* = TRUE）については，答えの数に関する特徴量とタスク満足度との間の相関係数が負であることが確認された（総数：$r = -0.380$，$p = .132$，種類数：$r = -0.339$，$p = .183$，エントロピー：$r = -0.458$，$p = .074$）．これらの特徴量のうち，答えのエントロピーについては，タスク満足度との間の相関係数に有意な傾向が見られた．この傾向は，他の属性値を持つユーザに対しては見られないため，検索専門性のあるユーザに特有の性質である可能性が高い．

一方で，検索専門性のないユーザの場合は，答えの総数，答えの種類数，答えのエントロピー，のいずれの特徴量についても，タスク満足度との間に有意な相関関係は確認されなかった．全 5 種類の特徴量のうち，これらのユーザの満足度との間の相関係数が有意であったのは，入力クエリ数のみであった（$r = -.546$，$p < .001$）．

本項で得られた分析結果は，検索専門性の存在が答えの一貫性とタスク満足度との間の関係性に影響を与えることを示唆しており，これは仮説 H2 の裏付けと見なせる．この結果から，検索専門性のないユーザについては，上述の特徴量のうち検索コストのみが満足度評価に影響を与えるものと予想される．一方で，検索専門性のあるユーザは満足度を評価する際に，検索コストの大小だけでなく，得られた答えの一貫性についても考慮していると考えられる．

両者の間で満足度の評価基準に差異が生じることの説明として，検索専門性の中でも，検索結果のランキングの影響があげられる．Nakamura らが行った大規模アンケート調査 [25] によると，多くの一般ユーザは，検索エンジンをある程度信用している一方で，検索結果のランキングの仕組みに関しては正確に理解していないといわれている．そのため，検索専門性のないユーザは，検索エンジンが提示した検索結果の順位を過度に信用してしまう可能性がある．対照的に，検索専門性のあるユーザは，検索結果のランキングの要因について一定の知識を持っている．そのため，彼らは得られた答えを評価する際に慎重を期している可能性が考えられる．

ここで，上述の結果に対するタスクの影響について考察する．3.5 節で述べたように，本研究では内容に対する満足度を対象としている．そのため，ユーザが明確な基準で答えの正しさを判断できる場合，発見された答えの数は，

タスクの満足度に影響を与えることはないはずである．それにもかかわらず，満足度と答えのエントロピーの間に負の相関が確認された要因として，本稿で分析対象とした検索タスクには，答えの正しさを客観的に判断することが難しいという特徴が存在することがあげられる．そのため，たとえ複数の答えが存在しても，その真偽が容易に判定できるタスクの場合には，本項で述べた議論は成立しない可能性がある．

4.1.3　事前知識の影響

次に，タスクの事前知識に関する属性（*knowledge*）の影響に注目する．事前知識のあるユーザ（$knowledge > 1$）に対してのみ観測された傾向として，エントロピーを除く答えの数に関する特徴量とタスク満足度との間の相関係数が比較的大きな正の値になったことがあげられる（総数：$r = .443$，$p = .232$，種類数：$r = .432$，$p = .246$）．しかし，無相関検定の結果，これらの相関係数については有意性を確認することはできなかった．そのため，仮説 H2 の事前知識に関する部分については，その裏付けとなる結果が得られなかった．

事前知識のないユーザ（$knowledge = 1$）の場合は，5 種類の特徴量のうち入力クエリ数についてのみ，タスク満足度との間に有意な相関関係が存在することが確認された（$r = -.413$，$p = .001$）．この結果は，前項における検索専門性のないユーザ（$is\_ir = \mathrm{FALSE}$）の傾向と類似している．そのため，これらのユーザは，タスクに対する満足度を，その答えを得るまでに費やしたコストから評価するという共通の特徴を持っていることが予想される．

4.2　タスクの実行に費やす時間

前節では，タスクに対する満足度に影響を与える可能性のある要因として，発見された答えの数に着目し，ユーザの属性ごとに両者の関係性の分析を行った．本節では，ユーザがタスクを終えるまでに費やす時間に着目する．タスク終了に要する時間とユーザの満足度との間には負の相関関係が存在することが既存研究 [32] により示されてきた．しかし，ユーザの属性の有無によって両者の関係性は異なる可能性がある．そこで我々は，ユーザ属性とタスク終了までの時間との関係性に関する以下の仮説を検証することで，ユーザ属性の有無によるタスク中の時間の使われ方の違いを分析する．

H3 検索専門性のあるユーザは，タスクの終了までに長い時間を費やす．

H4 事前知識のあるユーザは，タスクの終了までに費やす時間が短い．

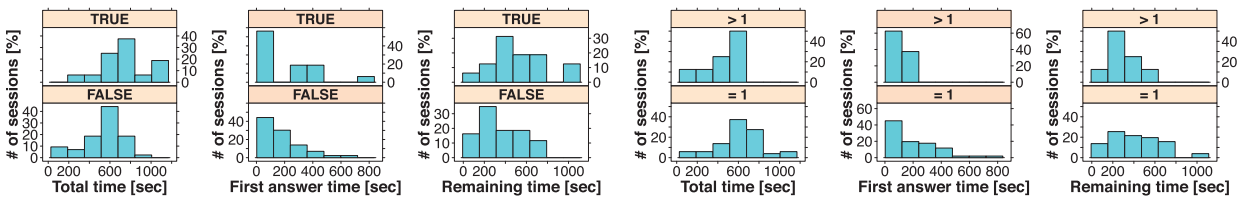
検索クエリの修正過程を分析した既存研究 [4] では，ユーザはまず広範な検索クエリを用いてタスクの概観を把握し，その後で対象を絞り込むためにクエリを詳細化する傾向にあると報告されている．この知見をふまえると，複数の答えが発見されるという特徴を持つ本研究の対象タスクの場合，ユーザは最初の答えを発見する前後で検索の方針を変更する可能性がある．そこで我々は，ユーザがセッション中で最初の答えを発見する時点を区切りとした，以下の 3 種類の区間に関してユーザが費やす時間の傾向を分析する．

- セッション開始から終了までの経過時間
- セッション開始から最初の答えの発見までの経過時間
- 最初の答えの発見からセッション終了までの経過時間

上記の各経過時間に関するセッション数の分布をヒストグラムで表現したものを**図 1** に示す．同図において，左部（図 1 (a)）は検索専門性（*is_ir*）の有無で場合分けしたセッション分布を，右部（図 1 (b)）はタスクに関する事前知識（*knowledge*）の有無で場合分けしたセッション分布を表している．

4.2.1　検索専門性の影響

検索専門性が最初の答えの発見に要する時間，およびタスクの終了までに費やされる時間に与える影響について述べる．検索専門性の値ごとのタスク達成時間の分布（図 1 (a) の最左部）に注目すると，検索専門性のあるユーザ（$is\_ir = \mathrm{TRUE}$）はそれ以外のユーザ（$is\_ir = \mathrm{FALSE}$）に比べ，長い時間をかけてタスクを実行する傾向にあることが分かる．より具体的には，前者のユーザに関するセッションの約半数においてタスク達成時間が 10 分を超えているのに対し，後者のユーザに対するそのようなセッションの割合は全体の 4 分の 1 に満たない．両者の時間分布に対して Welch の t 検定を適用した結果，検索専門性のあるユーザはそれ以外のユーザに比べてタスク達成時間が有意に長いことが分かった（$t(24) = 3.29$，$p = .003$）．これは仮説 H3 を支持する結果といえる．そこで以下では，最初の答え発見までの時間，およびそれ以降の時間を調べるこ



(a) 検索専門性（*is_ir*）による影響　　(b) 事前知識（*knowledge*）による影響

図 1　答えの発見に費やした時間に関するセッション数のヒストグラム

Fig. 1　Histograms of number of sessions regarding elapsed time for finding answers.

とで，検索専門性のあるユーザがタスクのどの部分に多くの時間を費やしているかを分析する．

検索専門性のあるユーザが最初の答えを発見するまでの時間を調べたところ，彼らの半数以上がセッション開始から 100 秒以内に最初の答えを発見していることが判明した（図 1 (a) の中央部）．その一方で，検索専門性のないユーザに関する当該セッションの割合は 4 割程度であった．タスク達成時間に関する分析と同様に両者の時間の検定を行ったが，最初の答えの発見に要する時間に関しては，検索専門性の有無によって有意な差は見られなかった（$t(20) = -.985$，$p = .337$）．そのため，最初の答え発見までの時間に関しては，検索専門性の有無による影響はないものと考えられる．

図 1 (a) の最右部のセッション分布は，検索専門性のあるユーザの多くが最初の答えの発見以降も長時間にわたってタスクを継続していることを示している．検索専門性のあるユーザの平均的なタスク継続時間は約 520 秒，検索専門性のないユーザの当該時間は約 350 秒であり，両者には有意な差が存在することが確認された（$t(21) = 2.24$，$p = .036$）．そのため，最初の答えの発見以降の時間については，検索専門性の影響が存在すると考えられる．

本項での分析によって，検索専門性のあるユーザはタスクの終了までに長い時間を費やしており，その長さは最初の答え発見以降の検索行動に起因するものであることが判明した．この結果から，検索専門性のあるユーザは検索の過程で発見した答えに対して慎重な態度を示す傾向があり，答えの候補が 1 つ見つかっただけではタスクを終えない，という姿勢がうかがえる．さらに，答えのエントロピーと満足度との間に負の相関関係が存在するという 4.1 節の分析結果もふまえると，検索専門性のあるユーザは現在得られている答えの信頼性を確認するために，残りのセッションにおいて証拠となる他の情報源を探している可能性がある．ただし，本項における議論は，内容に対する満足度を前提としたものである．そのため，別の種類の満足度に対しては，タスク達成時間との間に別の関係性が成り立つことも考えられる．たとえば，システムに対する満足度の立場の下では，達成までに長い時間を要するシステムに対して，検索専門性のあるユーザは不満を覚えるかもしれない．

本節冒頭で述べたように，タスク達成時間は満足度と負の相関関係にあることが既存研究によって報告されてきた [32]．しかし，本項での分析を通じて，検索専門性の有無によってタスク達成時間の長さの意味合いが異なる可能性が高いことが分かった．ユーザの属性を考慮せずに検索行動の分析を行った場合，検索専門性のある少数のユーザのデータは，それ以外の多数のユーザデータに埋もれてしまい，彼らに特有の検索行動を把握することが困難になる．本項の分析結果は，ユーザ属性の違いを考慮したうえで検索行動を分析することの重要性を示唆している．

4.2.2　事前知識の影響

タスクに関する事前知識についても，最初の答えの発見時間やタスク達成時間に影響を与えている．図 1 (b) の最左部からは，事前知識のあるユーザがそうでないユーザに比べ短時間でタスクを終えていることが分かる．事前知識のあるユーザは全員 10 分以内にタスクを達成しており，その時間は事前知識のないユーザと比べると有意に短い（$t(11) = -2.52$，$p = .029$）．この結果は，仮説 H4 の成立を支持している．そこで以降では，最初の答え発見までの時間とそれ以降の時間のうちのどちらが，全体の時間の短縮に寄与しているかを分析する．

図 1 (b) の中央上部のヒストグラムから，事前知識のあるすべてのユーザは，タスク開始から 200 秒以内に最初の答えを発見していることが分かる．一方で事前知識のないユーザの約 4 割は，最初の答えを発見するまでに 200 秒以上を費やしており，両者の間に有意差が存在することが確認された（$t(23) = -2.73$，$p = .012$）．タスクの答えに関する事前知識が存在すると，その発見に有用な語を含む検索クエリをセッション開始時点から生成しやすくなると考えられる．そのため，この結果は納得のいくものといえる．また，文献 [22] で述べられている，ドメイン専門性の存在による閲覧時間の短縮に関する効果についても，事前知識のあるユーザの答えの早期発見に寄与した可能性がある．

一方で図 1 (b) の最右部は，最初の答えを発見してからのタスク継続時間の最頻値が，事前知識の有無にかかわらず 200 秒前後であることを示している．実際に，事前知識のあるユーザとそれ以外のユーザとの間には，タスク継続時間に有意差が確認されなかった（$t(13) = -1.45$，$p = .172$）．そのため，タスク継続時間に関しては，事前知識の有無による影響はないものと考えられる．

本項での分析結果から，タスクに関する事前知識の存在によって，タスクの開始から終了までに費やされる時間が短くなるという傾向が確認された．また，事前知識の存在による最初の答え発見までの時間の短縮が，その主要因となっていることが判明した．短時間でのタスクの達成は，事前知識のあるユーザの満足度が他のユーザに比べ相対的に高くなったこと（表 3）に影響を与えていると考えられる．この傾向は，検索コストと満足度との間に負の相関関係が成り立つという前節の分析結果とも一貫性がある．

4.3　時間経過による答えの変化

前節の分析によって，検索専門性のあるユーザは，最初の答えの発見以降もセッション終了までに長い時間をかけることが確認された．彼らはセッションの残りの時間を使って，最終的に報告する答えに関する情報を中心に検索している可能性がある．そこで本節では，ユーザがセッション中で発見する答えの時間的な変化に着目した分析を行う．具体的には，時間経過にともなう検索対象の答えの

変化に関して，以下の仮説を立てる．

H5 検索専門性のあるユーザは，時間の経過に従って検索対象の答えを絞り込む．

分析の前準備として，我々はユーザがセッション中で発見した答え（*found_answers*）を，(1) 報告された答え（*reported_answers*），および (2) 報告されなかった答え（*other_answers*），の 2 種類に分類した．セッションの序盤と終盤のそれぞれにおいて，これらの各種類の答えが閲覧ページ中にどの程度現れているかを計算することで，時間経過による答えの出現傾向の変化を調べる．

セッションの序盤と終盤の設定の仕方としては，最初の答えの発見時刻や，タスク達成時間の中間値による分割など，いくつかの方法が考えられる．しかし，表 2 に示したように，対象タスクの中には平均閲覧ページ数が 6 を下回るものも存在する．こうしたタスクに対して，上記の基準を用いた分割を行うと，セッションの各段階に相当するページ数がきわめて小さくなるため，分割点の選択によって結果に大きな変化が生じる可能性がある．そこで今回は，明示的に分割点を定めるのではなく，(検索結果ページを除く）ページ集合を閲覧時刻の昇順で並べた系列をセッション序盤，降順で並べた系列を終盤と見なすことで，この問題に対処することとした．

これらの各段階におけるページ系列に対して，その中で出現する答えの傾向を計算するために本稿では nDCG [18] を用いる．nDCG を計算する際には，対象としている種類の答えを含むページを適合，それ以外を非適合ページとして扱う．nDCG は上位に多くの適合ページを含む系列を高く評価するため，その値が高いほど対象の答えが対象の段階において頻繁に出現していることを表す．

セッションの各段階における閲覧ページ系列に対して，答えの種類ごとにその出現傾向を計算した結果を**表 5** に示す．表中の各セルの左段はセッション序盤に対する出現傾向の平均値（および標準偏差）を，右段はセッション終盤に対する結果を表している．また，各セルの中段は，セッション序盤と終盤との間での出現傾向の増減（および対応ありの t 検定時の p 値）を示す．増加した結果には $\nearrow$，減少したものには $\searrow$ が付与されている．

4.3.1 検索専門性の影響

検索専門性のないユーザ（$is\_ir = \mathrm{FALSE}$）については，*reported_answers* および *other_answers* のどちらについても，セッションの時間経過に従って出現傾向が有意に増加するという結果が得られた（*reported_answers*：$t(48) = 3.50$，$p = .001$，*other_answers*：$t(48) = 2.07$，$p = .044$）．この結果は，検索専門性のないユーザはセッション序盤に比べて終盤に多くの答えを発見しており，それは答えの種類によらないことを示唆している．

一方で，検索専門性のあるユーザ（$is\_ir = \mathrm{TRUE}$）に対しては別の傾向が存在する．専門性のあるユーザの場合，時間の経過に従い *reported_answers* の出現傾向は有意に増加している（$t(16) = 1.75$，$p = .099$）．また，*other_answers* については減少傾向にあることが同表から分かる．ただし，後者の変化については有意差は確認されなかった（$t(16) = -1.26$，$p = .225$）．そのため，検索専門性のあるユーザの場合，最終的に報告される答えのみが時間の経過に従って発見頻度が増加するといえる．これは，仮説 H5 の裏付けと見なすことができる．

本項での分析結果から，検索専門性のないユーザはセッションの終盤にさしかかっても特定の答えに絞り込んだ検索を行っていないことが予想される．対照的に，検索専門性のあるユーザは最終的に報告する答えに関するページをセッションの終盤で集中的に検索していると考えられる．これらの結果は 4.2.1 項での考察とも整合するものであり，彼らの妥当性検証のプロセスが専門性のないユーザとは異なるということを示唆している．

4.3.2 事前知識の影響

タスクに関する事前知識（*knowledge*）についてはその

表 5 セッションの各段階において発見された答えの種類の傾向．各セルの左段/右段はセッションの序盤/終盤での閲覧ページ系列の nDCG 値の平均（および標準偏差）を，中段は序盤と終盤との間での nDCG 値の増減（および p 値）を示す

Table 5 The tendency of types of found answers in each stage of sessions. Left/Right value of each cell is mean (with standard deviation) in early/closing stage of sessions. The change (with p-value) in nDCG scores between two stage is shown in the middle.

ユーザ属性		*reported_answers*			*other_answers*		
		序盤	⟷	終盤	序盤	⟷	終盤
is_ir	TRUE	.514 (.33)	$\nearrow$ ($p = .099$)	.652 (.39)	.646 (.41)	$\searrow$ ($p = .225$)	.548 (.37)
	FALSE	.444 (.39)	$\nearrow$ ($p = .001$)	.591 (.45)	.504 (.39)	$\nearrow$ ($p = .044$)	.609 (.41)
knowledge	> 1	.549 (.45)	$\nearrow$ ($p = .671$)	.590 (.47)	.550 (.47)	$\nearrow$ ($p = .459$)	.626 (.48)
	$= 1$	.448 (.36)	$\nearrow$ ($p < .001$)	.610 (.44)	.539 (.39)	$\nearrow$ ($p = .317$)	.588 (.39)

有無にかかわらず，検索専門性のないユーザと同様の結果（どちらも答えの種類も増加傾向）となった．しかし，出現傾向の変化に有意性が確認されたのは，事前知識のないユーザ（$knowledge = 1$）の *reported_answers* に関する結果のみであり（$t(56) = 4.03$，$p < .001$），全属性のうち事前知識あり（$knowledge > 1$）の場合についてのみ，時間経過にともなう *reported_answers* の増加に有意性が認められなかった．この理由の 1 つとして，事前知識のあるユーザのセッション序盤における *reported_answers* の発見頻度が高いということがあげられる．事前知識が存在することで，答えに関する情報をタスク開始直後から高精度で検索することが可能になったものと思われる．

5. 考察

本章では，4 章の分析から得られた知見を整理し，その結果から考えられる検索支援について考察する．また，分析に利用した変数間での交絡の可能性について検証する．さらに，個々のタスクを区別せずに分析を行ったことによる影響について考察する．最後に今回行った分析の限界点を述べるとともに，今後の課題を整理する．

5.1 可能な検索支援

4.1 節の分析から，検索専門性のあるユーザは，閲覧ページ数と答えのエントロピーの 2 つの特徴量がそれぞれタスクの満足度と負の相関関係にあることが確認された．検索専門性のないユーザの場合は，入力クエリ数が満足度と同様の関係を示した一方で，答えの数に関する特徴量に関しては満足度との相関関係が認められなかった．この 2 つの結果から，検索専門性のないユーザが満足度を評価する際には，タスク終了までに費やした検索コストが中心的な要因となっており，タスク実行時に発見した答えの整合性については考慮していないことが予想される．

次に 4.2 節の分析からは，検索専門性のあるユーザが最初の答えを発見以後もタスクを長時間継続するという傾向が確認された．その一方で，専門性のないユーザのタスク継続時間は，専門性のあるユーザに比べて有意に短いことが判明した．この結果から推測可能なユーザ心理として，検索専門性のないユーザは答えが見つかったことに満足し，その正確性や信頼性の検証に注意を向けていないという可能性が考えられる．

最後に 4.3 節の分析では，検索専門性のあるユーザは，タスク終了時に報告する答えをセッション終盤において多く閲覧しているのに対し，専門性のないユーザはそれ以外の答えについても，時間の経過に従ってその閲覧量が増加するという結果が得られた．この結果から，専門性のないユーザは特定の答えに絞り込んだ検索を行えていないことが示唆される．

以上の知見をふまえると，検索専門性や事前知識のないユーザに対する検索支援としては，現在調べている答え以外にも多くの候補が存在することを認識させることが重要になると考えられる．そのための手段として，非検索対象の答えに関する情報をクエリ推薦などを用いて提示することは，当該ユーザの注意を喚起するという意味でも有用と思われる．

しかし，4.1 節の分析を通じて得られた「検索専門性のあるユーザについては，タスク満足度と答えのエントロピーとの間に負の相関関係が存在する」という結果をふまえると，彼らは多くの答えに遭遇した場合，そうでない場合に比べてタスクに満足しにくくなることが予想される．そのため，タスクの答えとして可能性のある複数の候補を列挙するという単純なアプローチでは，かえって当該ユーザの満足度の低下を引き起こすおそれがある．

その解決策として，それぞれの答えに対する Web 上での言及数を示すことで，答えの多数派/少数派を意識した検索を支援できる可能性がある．さらに，それぞれの答えの典型性や答えの言及元の信頼性といった情報の判断基準 [33] を提示することで，ユーザがタスクに関する知識を持たない場合であっても，事前知識のあるユーザと同等の情報精査が可能になると予想される．また，文献 [1] で提案されている意見の多様化手法をタスクの答えに関する情報に対して適用できれば，専門性のないユーザであっても答えに対する賛成/反対意見を低い検索コストで収集可能になると考えられる．

5.2 変数間の交絡

4.1.2 項における検索専門性の影響分析の結果，検索専門性のあるユーザに特有の傾向として，答えのエントロピーとタスク満足度との間の負の相関係数が有意な傾向にあることが確認された（$r = -.458$，$p = .074$）．この結果に基づき同項では，検索専門性の存在がタスク満足度に負の影響を与えるという主張を行った．

ここで注意すべき点として，タスク満足度と答えの発見数のそれぞれが検索コストに関する特徴量から受ける影響の存在があげられる．表 4 に示したように，検索専門性のあるユーザには，閲覧ページ数と満足度との間に有意な負の相関関係が存在する（$r = -.631$，$p = .007$）．また，閲覧ページ数の増加にともない，発見される答えの数も増加することが予想される．そのため，答えのエントロピーとタスク満足度との間で確認された負の相関関係は，閲覧ページ数を交絡変数とした擬似相関であるという可能性も考えられる．入力クエリ数についても同様の議論が可能である．

そこで，検索コストに関する各特徴量の影響を取り除いたうえで，答えの数に関する特徴量と満足度との間の偏相関分析を行った．この分析によって，両者の間の相関関係のうち，検索コストでは説明できない分を計算すること

表 6　タスク間での各属性のユーザ数の分布
Table 6　Distribution of number of users among tasks for each attribute.

		Drought	Pixels	TV	Verizon
is_ir	TRUE	5 (29.4%)	5 (29.4%)	4 (23.5%)	3 (17.6%)
	FALSE	11 (22.4%)	9 (18.4%)	15 (30.6%)	14 (28.6%)
knowledge	>1	0 (0.00%)	3 (33.3%)	4 (44.4%)	2 (22.2%)
	$=1$	16 (28.1%)	11 (19.3%)	15 (26.3%)	15 (26.3%)

ができる．閲覧ページ数の影響を取り除いた偏相関分析の結果，答えの数に関する特徴量とタスク満足度との偏相関係数は依然として負値であることが確認された（総数：$r = -.354$，$p = .163$，種類数：$r = -.291$，$p = .258$，エントロピー：$r = -.432$，$p = .094$）．入力クエリ数の影響を取り除いた場合についても，同様の傾向が見られた（総数：$r = -.356$，$p = .161$，種類数：$r = -.314$，$p = .220$，エントロピー：$r = -.465$，$p = .070$）．

以上の結果から，検索コストに関する影響を取り除いた後でも，答えのエントロピーとタスク満足度との負の相関関係が有意な傾向にあるといえる．そのため，4.1.2 項で主張した，検索専門性の存在がタスク満足度に負の影響を与えるという仮説は，引き続き支持可能であると考える．

5.3　個々のタスクの影響

本研究ではユーザが発見した答えとタスクの満足度との関係性を調査するにあたって，Feild ら [9] が公開している検索ログデータの中から，分析対象を一部のタスクに限定するという方法を採用した．その結果，分析対象のセッションは合計で 66 個（表 3）に限られ，各タスクあたりの平均セッション数も 16.5 個と小さい値になった．そこで本稿では，個々のタスクにおける結果については区別せず，すべてのタスクの結果をまとめたうえでユーザ属性の影響に関する分析を行った．

しかし**表 6** から，各属性に対応するユーザの数はタスクによって異なることが分かる．たとえば，検索専門性のないユーザに関するデータは，TV タスクに 15 個存在するのに対して，Pixels タスクには 9 個しか存在しない．また，Drought タスクにいたっては，事前知識のあるユーザに関するデータがいっさい含まれていない．こうしたタスク間でのデータの偏りによって，多数派を占めるタスクの検索ログが，4 章で得られた分析結果に大きな影響を及ぼしている可能性がある．

4 章の分析結果に基づく本稿の主張は「検索専門性のあるユーザは発見した答えに対して慎重な態度を示す傾向にあり，答えの発見以後も長い時間をかけてタスクに取り組む」というものである．本節では，多くのタスクに共通してこの主張が成立するのか，あるいは特定のタスクに限定されるものなのかを，個々のタスクのデータを分析することで検証する．なお表 6 が示すように，タスクの中にはユーザ数がきわめて少ないものが存在する．そこで以降の検証では，タスクごとの結果の有意性については議論せず，全体の結果と類似した傾向が見られるかに着目する．

表 7　満足度とエントロピー間の相関係数のタスクごとの影響
Table 7　Task effect on correlation between satisfaction and entropy.

		Drought	Pixels	TV	Verizon
is_ir	TRUE	$-.478$ ($p = .415$)	$.000$ ($p = 1.00$)	$-.500$ ($p = .667$)	$-.500$ ($p = .667$)
	FALSE	$.177$ ($p = .625$)	$-.270$ ($p = .483$)	$.607$ ($p = .047$)	$-.474$ ($p = .102$)

5.3.1　答えのエントロピーとタスク満足度への影響

4.1.2 項の分析結果に基づく「答えのエントロピーが増加するとタスク満足度が低下する傾向がある」という主張について，個々のタスクによる影響を検証する．各タスクについて，答えのエントロピーとタスク満足度との間の相関係数を計算したところ，**表 7** に示す結果が得られた．

同表から，ユーザに検索専門性のない場合，両者間の相関係数の符号およびその値がタスクによって大きく異なることが分かる．そのため，検索専門性のないユーザについては，答えのエントロピーとタスク満足度との間に，全タスクで共通する関係性は存在しないものと考えられる．実際に，同ユーザのタスク全体における結果（表 4）では，両者の間の相関係数には有意性が認められなかった（$r = -.021$，$p = .896$）．

一方，検索専門性のあるユーザについては，Pixels を除くすべてのタスクにおいて，答えのエントロピーと満足度との間の相関係数が負になっている．これらの値は，同ユーザのタスク全体に対する相関係数（$r = -0.458$，$p = .074$）と同様の傾向を示している（表 4）．そのため，検索専門性のあるユーザに関するこの特徴は，多くのタスクに共通するものであると考えられる．

唯一の例外として，Pixels タスクに対してだけは，検索専門性のあるユーザの満足度とエントロピーとの間に負の相関係数が確認できなかった．Pixels タスクには，答えが記述されたオフィシャルページが存在せず，また表 2 が示すように，同タスクの答えの総種類数は Drought タスクに次いで多いという特徴が存在する．そのため我々は，両者の間にも負の相関係数が成立すると予想していたが，得られた結果は予想とは異なっていた．

この理由を明らかにするために，個々のタスクについて表 4 の各特徴量の値を計算し，その比較を行った．その結果，発見された答えの数は，タスクの答えの総種類数に比例する傾向にあり，Pixels タスクの特異性は確認できなかった（Drought：7.40，Pixels：3.60，TV：2.25，Verizon：2.33）．一方で，Pixels タスクの閲覧ページ数は，他のタスクに比べて少ないことが分かった（Drought：6.20，Pixels：3.00，TV：5.75，Verizon：4.67）．Pixels タスクの

閲覧ページ数だけが極端に低くなった理由については定かではないが，極端に低い検索コストで同タスクを終了できたことが満足度の評価を大きく左右し，その結果，答えのエントロピーによる影響が小さくなったのではないかと予想される．

5.3.2　タスク実行に費やされる時間への影響

4.2.1 項の分析結果に基づく「最初の答えの発見以降も長い時間が費やされ，タスク終了までの総時間も増加する」という主張について，個々のタスクによる影響を検証する．そのために個々のタスクについてユーザが費やした時間を，同項の分析と同様に 3 種類の区間に分割し，各区間における時間分布の平均値を計算した．異なる属性値間での経過時間の平均値の比較結果を**表 8** に示す．

同表から，検索専門性のあるユーザは専門性のないユーザに比べて総時間（Total time）が長いという傾向が，全タスクに共通するものであることが分かる．そのため，検索専門性のあるユーザはタスクの終了までに長い時間を費やすという仮説 H3 は，タスクに依存せず成立するものと考えられる．

しかし，最初の答えの発見までの時間（First answer time）とそれ以降の時間（Remaining time）に関する結果を比べることで，総時間増加の要因はタスクによって異なることが分かる．Drought タスクおよび Pixels タスクについては，総時間の増加分のうちの半分以上が最初の答え発見以降の時間の増加分に起因している．特に Drought タスクについては，全タスクの結果と同様に，総時間および最初の答え発見以降の時間の両方について有意差が確認された．一方，TV タスクおよび Verizon タスクについては，最初の答え発見までの時間の増加分が，総時間増加の主要な要因となっていることが分かる．

上述の時間の増加分に加えて，総時間のうち最初の答え発見以降の時間が占める割合についても，タスクごとの値を計算した．その結果，検索専門性のないユーザについては，タスク間でその割合に大きな差は見られなかった（Drought：.760，Pixels：.753，TV：.625，Verizon：.504）．対照的に，検索専門性のあるユーザの場合は，タスクによってその値が大きく変動していることが分かった（Drought：.915，Pixels：.723，TV：.528，Verizon：.354）．

以上の結果から，Drought タスクと Pixels タスクについては，検索専門性のあるユーザは答えの発見以降も長い時間をかけているといえる．これらのタスクの共通点として，多くの種類の答えが存在するという特徴が存在する（表 2）．そのため，検索専門性のあるユーザは，多くの答えが存在するタスクの場合に，最初の答えの発見以降も長い時間をかけてタスクに取り組むものと予想される．

表 8　答えの発見に費やされた時間のタスクごとの影響
Table 8　Task effect on elapsed time for finding answers.

タスク	時間区分	*is_ir*	
		TRUE	FALSE
Drought	Total time	900 $>$ ($p = .052$)	654
	First answer time	76 $<$ ($p = .028$)	157
	Remaining time	823 $>$ ($p = .027$)	497
Pixels	Total time	589 $>$ ($p = .624$)	519
	First answer time	163 $>$ ($p = .694$)	128
	Remaining time	426 $>$ ($p = .737$)	391
TV	Total time	754 $>$ ($p = .018$)	557
	First answer time	356 $>$ ($p = .113$)	209
	Remaining time	398 $>$ ($p = .610$)	348
Verizon	Total time	809 $>$ ($p = .126$)	432
	First answer time	523 $>$ ($p = .171$)	214
	Remaining time	286 $>$ ($p = .419$)	218

5.4　分析の限界点

最後に，本稿で行った分析の限界点について考察する．第 1 に考えられるのは，分析対象のデータ数の少なさにともない，結果を一般化することが困難であるという点である．前節で述べたように，本研究の分析で利用したデータセットの規模はきわめて小さい．そこで，4 章の分析では，$\alpha = .10$ という比較的大きな有意水準を用いて結果の有意性の検定を行った．しかし，有意水準として大きな値を採用することで，結果に有意性がないにもかかわらず帰無仮説が棄却される確率が大きくなる，という弊害が生まれ，得られた結果に対する信頼性が低くなっている．そのため今後は，大規模なデータに対して同様の分析を行うことで，本稿で有意性があると判断された結果が一般的に成立するかを検証する必要がある．

また，今回の分析で利用したデータには，前節で述べた属性内の偏りだけでなく，属性間にも偏りが存在している．表 3 より，検索専門性のないユーザ（$is\_ir = \text{FALSE}$）が行ったセッションは 49 個存在することが分かる．このうち，タスクに関する事前知識のない（$knowledge = 1$）セッションの数を調べたところ，42 個が該当した．これより，検索専門性のないユーザに関するセッション集合のうちの大部分（$\simeq 86\%$）が，事前知識のないユーザに関するセッション集合と重複しているといえる．そのため，4 章の分析で得られた両者のユーザに関する結果が，双方のユーザに対して成り立つのか，あるいはどちらか一方に対してのみ成立するのかは，今回の分析からだけでは明らかにすることが難しい．さらに，被験者間で実行タスク数が異なるという点も分析結果の解釈を難しくする．被験者 30 名それぞれに対してタスク実行数を計算したところ，5 名は 1 タスクのみ，14 名は 2 タスク，11 名は 3 タスクという結果になった．そのため，今回の分析結果ではタスクを多く実行した被験者の特性が強く現れている可能性がある．

限界点としてあげられる最後の 1 点は，本分析で得られた知見が成立するタスクの種類に関するものである．本研究では，事実発見型 [21] の検索タスクの中でも複数の矛盾する答えが存在するものに対象を定め，満足度と適合度の関係性を分析した．しかし，それ以外の検索タスクでは今回とは異なる分析結果が得られる可能性がある．検索クエリが navigational [5] な場合は，両者に強い正の相関関係が成立することが既存研究 [17] によって報告されている．Transactional な検索意図 [5] の場合は，閲覧したページの適合度だけでなく商品購入などの目的が達成されたかがユーザの満足度に大きく関与すると考えられる．また，曖昧な情報要求が複数回の検索を通じて明確化されるという性質を持つ探索型検索 [31] においては，多様な観点から検索ができたかや，検索対象への理解が深まったか，といった側面が満足度に影響を及ぼすことが予想される．

今後は上記の課題を解決するために，事実発見型だけでなくさまざまな検索タスクに対してユーザの検索行動の収集および分析を行うことで，タスクに対するユーザの満足度の形成過程を明らかにしていきたい．また，分析を通じて得られた知見をもとに，ユーザの満足度の向上につながる検索の仕組みについても検討を行う予定である．

6. おわりに

本稿では，検索専門性と事前知識という 2 種類のユーザ属性が，検索行動および満足度の評価基準に与える影響を調査した．我々は，事実発見型の中でも複数の矛盾する答えが存在する検索タスクを分析対象に選び，(1) タスク実行中に発見された答えの数，(2) タスクの達成に要する時間，および (3) 発見される答えの時間経過による変化，という 3 種類の観点から，ユーザの検索行動の分析を行った．

Web 上で公開されている事実発見型タスクの検索ログに対して，閲覧ページに含まれるタスクの答えの抽出を行い，上記観点に基づいた分析を行った結果，各属性の有無によってユーザの満足度の評価基準が異なることを示唆する以下の傾向が見られた．

- 情報検索の専門知識を持つユーザについては，発見された答えの一貫性と満足度との間に負の相関関係が存在する可能性がある．
- 情報検索の専門知識を持つユーザは，答えの発見以後も長い時間をかけてタスクに取り組む．
- 情報検索の専門知識を持たないユーザは，タスク開始から一定時間が経過した後も，特定の答えに絞り込んだ検索を行わない．

本稿ではページ中に出現する答えに着目して分析を行ったが，ユーザの検索行動，閲覧ページの適合度，およびタスクに対する満足度の関係性をより深く理解するためには，ユーザが入力したクエリやマウス操作といった，他の行動情報も含めた総合的な分析が必要になると考えられる．また，今回対象とした事実発見型の検索タスクとそれ以外のタスクによって，上記の関係性にどういった違いが生じるのかについても明らかにする必要がある．今後はこうした課題を解決するだけでなく，分析を通じて得られた知見をもとに，ユーザの満足度の向上につながる検索の仕組みについても検討を行う予定である．

謝辞　本研究の一部は，文部科学省科学研究費補助金基盤研究（A）「ウエブ検索の意図検出と多元的検索意図指標にもとづく検索方式の研究」（研究代表者：田中克己，課題番号：24240013），特別研究員奨励費「ユーザの行動モデルに基づく検索意図推定に関する研究」（研究代表者：梅本和俊，課題番号：13J06404）によるものです．ここに記して謝意を表します．

参考文献

[1] Aktolga, E. and Allan, J.: Sentiment Diversification with Different Biases, *Proc. 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.593–602 (2013).

[2] Al-Maskari, A., Sanderson, M. and Clough, P.: The Relationship Between IR Effectiveness Measures and User Satisfaction, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.773–774 (2007).

[3] Aula, A., Khan, R.M. and Guan, Z.: How does Search Behavior Change as Search Becomes More Difficult?, *Proc. 28th International Conference on Human Factors in Computing Systems*, pp.35–44 (2010).

[4] Boldi, P., Bonchi, F., Castillo, C. and Vigna, S.: From "Dango" to "Japanese Cakes": Query Reformulation Models and Patterns, *Proc. 2009 IEEE/WIC/ACM International Conference on Web Intelligence*, pp.183–190 (2009).

[5] Broder, A.: A Taxonomy of Web Search, *SIGIR Forum*, Vol.36, No.2, pp.3–10 (2002).

[6] Büttcher, S., Clarke, C.L.A. and Cormack, G.V.: *Information Retrieval: Implementing and Evaluating Search Engines*, MIT Press, Cambridge, Mass. (2010).

[7] Carterette, B.: System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation, *Proc. 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.903–912 (2011).

[8] Chapelle, O., Metlzer, D., Zhang, Y. and Grinspan, P.: Expected Reciprocal Rank for Graded Relevance, *Proc. 18th ACM Conference on Information and Knowledge Management*, pp.621–630 (2009).

[9] Feild, H.A., Allan, J. and Jones, R.: Predicting Searcher Frustration, *Proc. 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.34–41 (2010).

[10] Guo, Q., Lagun, D. and Agichtein, E.: Predicting Web Search Success with Fine-grained Interaction Data, *Proc. 21st ACM International Conference on Information and Knowledge Management*, pp.2050–2054 (2012).

[11] Hassan, A., Jones, R. and Klinkner, K.L.: Beyond DCG: User Behavior as a Predictor of a Successful Search, *Proc. 3rd ACM International Conference on Web Search and Data Mining*, pp.221–230 (2010).

[12] Hassan, A., Shi, X., Craswell, N. and Ramsey, B.: Beyond Clicks: Query Reformulation as a Predictor of Search Satisfaction, *Proc. 22nd ACM International Conference on Information & Knowledge Management*, pp.2019–2028 (2013).
[13] Hassan, A. and White, R.W.: Personalized Models of Search Satisfaction, *Proc. 22nd ACM International Conference on Information & Knowledge Management*, pp.2009–2018 (2013).
[14] 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, *JASIST*, Vol.56, No.8, pp.861–871 (2005).
[15] Hersh, W., Turpin, A., Price, S., Kraemer, D., Olson, D., Chan, B. and Sacherek, L.: Challenging Conventional Assumptions of Automated Information Retrieval with Real Users: Boolean Searching and Batch Retrieval Evaluations, *Information Processing & Management*, Vol.37, No.3, pp.383–402 (2001).
[16] Hölscher, C. and Strube, G.: Web Search Behavior of Internet Experts and Newbies, *Computer Networks*, Vol.33, No.1-6, pp.337–346 (2000).
[17] Huffman, S.B. and Hochster, M.: How Well does Result Relevance Predict Session Satisfaction?, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.567–574 (2007).
[18] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM Trans. Information Systems*, Vol.20, No.4, pp.422–446 (2002).
[19] Järvelin, K., Price, S.L., Delcambre, L.M.L. and Nielsen, M.L.: Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions, *Proc. 30th European Conference on IR Research*, pp.4–15 (2008).
[20] Kanoulas, E., Carterette, B., Clough, P.D. and Sanderson, M.: Evaluating Multi-Query Sessions, *Proc. 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.1053–1062 (2011).
[21] Kellar, M., Watters, C. and Shepherd, M.: A Field Study Characterizing Web-Based Information-Seeking Tasks, *JASIST*, Vol.58, No.7, pp.999–1018 (2007).
[22] Kelly, D. and Cool, C.: The Effects of Topic Familiarity on Information Search Behavior, *Proc. 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pp.74–75 (2002).
[23] Kim, Y., Hassan, A., White, R.W. and Zitouni, I.: Modeling Dwell Time to Predict Click-level Satisfaction, *Proc. 7th ACM International Conference on Web Search and Data Mining*, pp.193–202 (2014).
[24] Landis, J.R. and Koch, G.G.: The Measurement of Observer Agreement for Categorical Data, *Biometrics*, Vol.33, No.1, pp.159–174 (1977).
[25] Nakamura, S., Konishi, S., Jatowt, A., Ohshima, H., Kondo, H., Tezuka, T., Oyama, S. and Tanaka, K.: Trustworthiness Analysis of Web Search Results, *Proc. 11th European Conference on Research and Advanced Technology for Digital Libraries*, pp.38–49 (2007).
[26] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M. and Gatford, M.: Okapi at TREC-3, *Proc. 3rd Text Retrieval Conference*, pp.109–126 (1994).
[27] Smucker, M.D. and Clarke, C.L.: Time-Based Calibration of Effectiveness Measures, *Proc. 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.95–104 (2012).
[28] Turpin, A.H. and Hersh, W.: Why Batch and User Evaluations Do Not Give the Same Results, *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.225–231 (2001).
[29] White, R.: Beliefs and Biases in Web Search, *Proc. 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.3–12 (2013).
[30] 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*, pp.132–141 (2009).
[31] White, R.W. and Roth, R.A.: Exploratory Search: Beyond the Query-Response Paradigm, *Synthesis Lectures on Information Concepts, Retrieval, and Services*, Vol.1, No.1, pp.1–98 (2009).
[32] Xu, Y. and Mease, D.: Evaluating Web Search Using Task Completion Time, *Proc. 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.676–677 (2009).
[33] Yamamoto, Y. and Tanaka, K.: Enhancing Credibility Judgment of Web Search Results, *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pp.1235–1244 (2011).
[34] 高久雅生，江草由佳，寺井　仁，齋藤ひとみ，三輪眞木子，神門典子：タスク種別とユーザ特性の違いが Web 情報探索行動に与える影響：眼球運動データおよび閲覧行動ログを用いた分析，情報知識学会誌，Vol.20, No.3, pp.249–276 (2010).

梅本 和俊

京都大学大学院情報学研究科博士後期課程在学中．日本学術振興会特別研究員（DC1）．2013 年京都大学大学院情報学研究科修士課程修了．主に情報検索におけるユーザ行動の分析と応用に関する研究に従事．日本データベース学会学生会員．

山本 岳洋　（正会員）

京都大学大学院情報学研究科社会情報学専攻助教．2011 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主に情報検索，特に情報検索におけるユーザインタラクションに関する研究に従事．日本データベース学会会員．

田中 克己　（フェロー）

京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．博士（工学）．主にデータベース，マルチメディアコンテンツ処理，ウェブ検索の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，日本データベース学会各会員．

（担当編集委員　石田 栄美）