情報検索行動における制約の効果

筑波大学
図書館情報メディア研究科
2013 年 3 月

藤川和也

目次

表目次

図目次

1　はじめに

インターネットの爆発的な発展と普及により，人々は日常の中でインターネットを利用した問題解決を自然に行なっている．中でもウェブ検索システムを利用した調べ事を問題解決の手法として利用することが少なくはなく，ウェブ検索システムを如何に上手く使えるかが問題解決の成否やコストに強く関わっていると言える．インターネットの黎明期にはじまり，ウェブウェブ検索システムが普及した今日にかけて，人間の情報検索行動に関する研究は広くなされてきた．それはウェブ検索システムの改良，可視化技術やデータマイニング技術の適用などによりユーザの利便性を向上させるためのものであったり，また，認知的アプローチによりユーザの視点に立った支援を目指して研究されたものであった．前者では情報システム上でユーザの行動を定量的に捉え分析し，その結果をシステムにフィードバックさせる手法をとり，後者では人間の情報検索行動のモデルや理論を適用することによりウェブ検索時の行動のパターンを記述するという手法が主にとられていた．以上のように，これまでの研究ではユーザの自然な情報検索行動の振る舞いや成果を理論的に記述しようとしたり，モデル化しようとしたりする手法が多くとられ，そこから得られた知見からウェブ検索システムを改良しようというアプローチが多かった．しかし，積極的に何らかの外的要因をもってウェブ検索システムを利用するユーザの振る舞いを制御することにより，ユーザの情報検索行動の質自体を高めようというアプローチはあまりとられてこなかった．

ここ数年話題になっている考え方として，行動経済学というものがある．行動経済学は，従来の伝統的な経済学とは違い，実際に存在する「完全ではない」人間を対象とした経済学である．伝統的な経済学で考えられていた人間とは違い，実際の人間は常に合理的な判断をするとは限らず，そして常に自分の利益・効用を最大化する行動を常に行えるわけではない．このような人間の性質を，行動経済学では「不合理」と呼んでいる．行動経済学の研究事例として Ariely と Wertenbroch の「先延ばし問題」[6] があり，ここでは，人間は問題解決にあたる際，外的要因として期限を設けられた場合のほうがを設けられていない場合と比べその成果が良いものになるという結果が報告されている．つまり，人間は不合理であるから，問題解決にあたる際，期限を自らで設けた場合は自己制御が上手く機能せず，思った通りの成果をあげられないということである．本研究ではこの「期限を設ける」という部分に着目した．すなわち，何らかの「期限」，もっと広く考えて「制限」を設けることにより，問題解決にあたる人間の振る舞いや成果はどのように変化するのかという命題に取り組んだ．

本研究では，行動経済学の考えを基に，「期限」や「制限」を設けたウェブ検索システムを開発し，そのシステム上でのユーザの情報検索行動を観察することにより，制限がかけられた際の人間の情報探索時の振る舞いと成果を分析し，そこから得られた知見を人間の情報検索行動の理解と情報システムの改良に役立てられないかと考え，被験者実験を行う．「どのような」制限を「どの程度」設ければ，どのようにユーザの振る舞いや成果は変化するのか．もしユーザの情報検索行動をより良くする「制限」を理解することができれば，ユーザに対してウェブ検索システムを利用した問題解決をより良くするための手法を提示できるかもしれないし，逆に考えてユーザがより良い成果を出せるような形にシステムを改良されることが期待される．

2 章では関連研究を中心に背景と目的を述べ，3 章では 1 つ目の実験について，実験デザイン，結果，考察を示す．4 章では 2 つ目の実験について，実験デザイン，結果，考察を示す．5 章では先行研究との比較を行う．6 章では研究の限界や研究の適用範囲について述べ，7 章では結論を述べる．

2　背景と目的

2.1　情報検索行動

情報検索行動に関する研究の初期では，コンピュータを利用した情報ウェブ検索システムを取り扱うものではなく，図書館やテレビ，新聞といった種々の情報源，メディアを主に取り扱っていた．ここでされていた研究としては，利用者のシステムや資源に対する要求，利用者の自覚・意識，利用者データなどに関するものがある．それらの研究結果より得られたものとしては，種々のサービスやメディアを利用する際に利用者の振る舞いがどのようなものであるのか，利用者にとって情報サービスは情報提供者の観点から期待されるよりも価値が低いように思えるとの結果 (最小努力の原理)，タスク分析の示唆，非学術的情報捜索への着目，などが挙げられる．時代が進むとともに，情報検索行動に対するアプローチも変化していった．黎明期を終えると，研究手法としては実態調査に依拠する研究から，次第にユーザ中心でその過程に着目する研究がされるようになった．Dervin や Franette　は，人が持つ広範囲に渡る関心や情報捜索における状況性 (situationality) が，「意味付けアプローチ」のおもな実証的研究成果であることを示唆している [7][8]．Ellis&Haugan[10] や Ellis, Cox および Hall らの研究 [9] では，データの収集時に被験者が遂行していた特定のタスクについての情報，すなわちある類型が特定種類のタスクでより多く使用されたかどうか，残念ながら報告されていない．それにもかかわらず，全く異なった情報捜索アプローチが利用できたるのである．Kuhlthau の提唱した情報捜索過程モデル [14] は，その後行われた複数の実証的研究において採用されている．これらのほとんどは，適合性判定あるいは情報検索行動を扱っている．人間の意思決定の過程やその働きについては，情報検索行動においても他の分野と同様にして過去に盛んに議論されてきた．「合理的意思決定論」は人間の意志や思考や判断が合理的であることを前提としている考えである．ここで言う「意思決定」とは，問題に対応してとるべき行為の選択肢を評価して選択する一連の行動のことを指す．そして，これまでの研究で，人間の意思決定が必ずしも合理的でないことは再三述べられてきた．Morehead らは，探索者の適合性判定を意思決定と位置づけて満足化モデルを適用した研究の結果から，満足化を含む複数の意思決定モデルを使って，時間の制約，新情報の減少，フラストレーションの 3 変数が検索を中断する要因であると仮定し，その妥当性を検証しようと試みた [18].

2.2　行動経済学

人間は限られた資源 (情報，時間，能力) を用いて，最善の行動を選び，実行しながらも，それでも後悔することが少なくない．人間の限定された合理性を中心に，最適な (合理的な) 行動からの乖離を経済分析の核に据える学問を「行動経済学」と呼ぶ．この行動経済学の考えでは，伝統的な経済学の考えのように，人間は経済計算だけで物事を判断するのではなく，人間の心理状況や知識からくる合理的ではない判断をする存在として考える．行動経済学の考えが一般に広く知られるようになったのは，その中心的な提唱者であったダニエル・カーネマンが，実験経済学者のバーノン・スミスと共同で，2002 年にノーベル経済学賞を受賞した前後からであるとされる．伝統的な経済学では，意思決定を行うにさいして完全な情報を有し，完全な計算能力を持ち，常に自分の利益・効用を最大化する行動を行えるホモエコノミクス (経済人) を想定している．しかし，実際の人間はホモエコノミクスから大きく乖離してい

るということは言うまでもない．このように，人間が完全情報や完全計算能力を持つと仮定して思考実験を行うことによってさまざまな予想や説明ができることは確かではあるが，しかしこの思考実験と実際に観察される事象とは違ったものであることは忘れてはいけない．伝統的経済学と行動経済学の最大の違いは，人間の「合理性」についての考え方である．人間の合理性には限界がある．先に挙げたように，ホモエコノミクスは完全な情報を有し完全な計算能力を持つとされているが，実際の人間には記憶力の限界もあり，計算能力に関しても 1 つの間違いも犯さないということはなく，そこから常に利益・効用を最大化する行動を常にとれるということも言えず，高々「十分」なそれを行うに留まる．この考えを，ハーバート・サイモンは「限定合理性」と名付けた [20]．サイモンは，人間が問題解決の可能な選択肢を選択しを発見する過程を研究すべきであると主張し，その主張の中で，人間は簡便な問題解決法を用いて，最適ではなくても満足のできる選択肢の発見に努めるとした．この簡便な問題解決法をヒューリスティクスと呼ぶ．ヒューリスティクスは，理性的で合理的であるというよりも，現実の人間が限られた資源を用いて直感的に判断を行う際に適用されるルールである．このように，行動経済学では，人間の認知の仕方や心理的バイアスがどのように意思決定に影響を与えるかを考えている．以上のように，伝統的な経済学が想定してきた理想の人間・ホモエコノミクスと現実の人間との間にはその判断の基準やそれにより生み出されるであろう結果や成果に小さくない隔たりがある．その原因としては人間が有する能力・資源であったり外的要因であったりそれによりもたらされる心理状況の変化であったり，そして心理状況の変化による合理的でない判断の結果であったりする．これまでに，理想と現実の乖離についてはいくつかの先行研究で論じられてきた．自己制御 (目標と実際の行動との間にある溝) については，Ainslie や Loewenstein らが論じ [4][15]，また，Akalov は時間と共に (締切が近づくにつれて) 変化する費用と利益との関係について検証してきた [5]．Ariely と Wertenbroch は，タスクの「期限」を様々な形で設けることによりパフォーマンスにどのような変化がみられるかを検証した [6]．ここでは「先延ばし」という行動 (心理) に焦点をあて，如何にして先延ばしを克服し，効果的な成果を出すかを考えた．また，タスクの「期限」を様々な形で設けることによりパフォーマンスにどのような変化がみられるかを検証した．この研究では，人があるタスクにおいて最大のパフォーマンスを発揮するためにはどういった期限の設定が最適であるかを考えることにより，日常の実際のタスクに活かすことを目標としていた．ここでいうタスクというのは，己に課されたタスクの場合もあれば，他人に課すタスクの場合もある (例えば，学校の先生が学生に課題を出すときに，どのような期限の設定をすれば学生はちゃんと課題をやってくるか，など)．しかし，これまでに，情報検索行動における「不合理性」について積極的に述べたものは，Nils らの研究 [19] を除けば多くはなかった．Nils らは，人間は意思決定において選択肢の内から最適なものではなく，条件を満たす最初に提示された選択肢を受け入れる，所謂「満足化」モデルを実証した．

2.3 制約

今日，ウェブ検索システムは技術の発展により誰でも，無料で，様々な情報を得られるようになっている．しかし過去には，利用に際して料金がかかったり，システムのレスポンスも遅く，検索できる情報にも限りがあり，利用可能な言語も限られ，その検索対象もテキストに限定されていた（表 1)．以上のように，過去のウェブ検索システムは今日のものと違い利用に様々な制約があり，利用者の情報検索行動に制限があった．

一般的には「制約」や「制限」はネガティブな要因としてとらえられているが，人間のタスクパフォーマンスにおいて有意に働くという研究結果も出ている．Junco と Cotten は，マルチタスキングにおいて

表 1　情報検索の過去と現在

過去		現在
有料	料金	無料
遅い	システムのレスポンス	早い
限定的	検索できる範囲	大きい
単一言語	利用可能な言語	複数言語
テキストのみ	検索対象	マルチメディア
古い	情報の鮮度	リアルタイム

人間の注意が散漫になることによるタスクパフォーマンスの低下について考察し，人間の行動を制限することによりタスクパフォーマンスの低下を避けることを述べた [13]．また，スクに制約の要素を取り入れることは，ゲーム性を取り入れる最も一般的な手法であり，それにより従事者の集中力や注意力を向上させる試みもなされてきた．実際，制約を取り入れることによりタスクのパフォーマンスが向上する事例も報告されている [16]．このように，集中力や注意力を向上させるために，「制約」を用いることは過去にも行われてきたが，情報検索の分野ではあまり研究されてこなかった．

以上より，本研究では，情報検索行動において不合理性や制約の考えに着想を得て，外的要因を用いて人間のタスクへの集中力や注意力を向上させることにより，成果を向上させる手法を提案し，それによる人間の意識，行動，成果への影響を分析する．その手法として，ウェブ検索システムに制約を設け，行動に制限をかけることにより，注意力を向上させる．不合理性や制約といった考えはこれまでにもあったが，これらを情報検索の分野で積極的に外的要因として用いる試みはあまりなされてなかった．また，Nils らの研究 [19] とは異なり，情報検索行動や情報システムを改善するための提案の 1 つとして，制約による「不合理性」を積極的に利用する．

3　実験 1: 情報検索行動への影響

3.1　仮説と目的

本研究では，先に挙げた行動経済学の分野での「先延ばし問題」と「自己制御」の考えを情報検索行動の分野に適用し，情報検索行動の分野でこれまでになされていない手法による研究ができないかと考えた．情報検索行動における「先延ばし問題」や「自己制御」の例としては，ある特定の「情報」を探すという行為 (=タスク) 内で利用できる資源の制限を加えることが考えられる．ここで，利用できる資源の制限がない場合と比較して，タスク中の振る舞いや得られる成果どのような変化が見られるだろうか．ここで加えられる制限は，ウェブ検索システムを情報検索行動に用いることを想定すると，例えば「ウェブページを閲覧できる回数」,「ウェブ検索システムに発行できるクエリーの回数」,「利用できるウェブ検索システム」，そして「ウェブ検索システムを利用できる時間」といったものが考えられる．これらの情報検索行動に使える資源を制限することにより，ユーザの意識や振る舞い，成果にどのような影響が出るだろうか．以上より，本研究で立てた仮説は，

- 制約は人の注意力に影響を与える．
- 注意力に影響が出ることにより，情報検索行動に影響を与える．
- 情報検索行動に影響が出ることにより，タスクの効率に影響を与える．

の 3 つである．この 3 つの仮説を検証するために本研究の目的として設定したことは，

- 制約が人に与える意識への影響を調べる
- 意識への影響によりあらわれる行動への影響を調べる
- 行動への影響によりあらわれる成果への影響を調べる

の 3 つである．次節では，この 3 つの目的の達成のために用いた実験デザインについて述べる．

3.2 実験デザイン

3.2.1 実験参加者

筑波大学の学部生 (男性 11 名，女性 8 名) 女性と大学院生 (男性 3 名，女性 1 名)，社会人 (男性 1 名) をあわせて 24 名が実験に参加した．年齢層は，10 代が 1 名，20 代が 23 名であった．参加した学部生，大学院生の専攻する分野とその分布は，情報学系が 7 名，情報工学系が 7 名，人文学・社会学系が 5 名，心理学系が 3 名，理工学系が 2 名であった．実験参加希望者の募集は，同大学内のメーリングリストを利用し，実験の実施は参加希望者とスケジュールを調整し，順に行なっていった．実験は 2011 年 12 月から 2012 年 1 月にかけて行い，実験参加者には謝金として，1,500 円分のアマゾンギフト券を支払った．

3.2.2 タスクとテストコレクション

本研究では，行動の制限が人間の情報検索行動の振る舞いと成果に与える影響をまず調べたいと考えた．そこで，こちらが設定した行動に対する制限の下で，あるトピックに対する調べごとの際の振る舞いと成果を観察できるようなタスクを設定した．今回の実験では，NTCIR の提供するテストコレクションの内，NTCIR-5CLIR[1] と NTCIR-6CLIR[2] を用いた．このテストコレクションには，2000 年から 2001 年の毎日新聞と読売新聞の記事が約 170 万件格納されており，情報検索行動のタスクのために様々なパラメータが設定されている．テストコレクションにはいくつかの「トピック」が設定されており，そのトピックについて書かれている記事を「適合文書」と呼んでいる．この「適合文書」の条件はトピック毎に定められており，同時にそのトピックの適合文書としてふさわしい，またはふさわしくない文書のリストも予め定められている．タスクの対象となるトピックは，こちらの用意したリストから実験参加者に選んでもらった（2)．トピックは表のように 6 つ用意してあり，その中から「興味のあるもの」を 3 つ選んでもらった．ここで，トピックによる振る舞い，成果の偏りを考慮して，6 種のトピックは取り扱っている分野や領域の偏りがないように用意した．また，トピックの難易度 (正答文書の数) が近い値のものを選別した．各トピックや適合文書については表 3 のようにテストコレクションで定義されている．

テストコレクションには記事とその適合性が格納されているが，NTCIR CLIR のテストコレクションの中に，適合性判断がされていない文書が含まれていたので，こちらで独自に適合性を判断した．対象となるトピックとその中に含まれていた適合性判断がされていなかった文書の数は，表 2 中の「5-008「ILOVEYOU」，コンピュータ・ウィルス」が 5 件，「5-016 歴史教科書論争，第二次世界大戦」が 31 件，

「5-018 たばこ，告訴，賠償金」が 76 件，「5-044 異常気象，災害，原因」が 67 件，「6-017 研究，取り組み，後天性免疫不全症候群（AIDS）」が 6 件であった．これらの適合性判断の詳細については付録を参照されたい．

実験参加者にはこのテストコレクションを全文検索できるウェブ検索システムを利用して調べごとをしてもらった．まずトピックを選択してもらい，そのトピックの適合文書のガイドラインを示し，ウェブ検索システムを使い適合文書を探してもらった．実際に実験参加者に伝えたタスクの内容は以下のようなものである．

「あなたは今，大学の講義で出た課題に取り組もうとしています．その課題とは，あるトピックについてウェブ検索システムを使って詳しく調べるというものです．先生からは，トピックのキーワード (タイトル) と，その簡単な概要だけ伝えられました．あなたはその情報を元に，ウェブ検索システムを使ってインターネット上からそのトピックについて述べられている文書を探すことにしました．課題の評価は，トピックに該当する文書をどれだけ多く見つけられたかによってされます．」

実験に移る前に，トレーニングとしてこちらの指定したトピック (表 2,「デリバティブ，損失」) についてのタスクをこなしてもらった．このトレーニングは，タスクの把握，ユーザインターフェイスの把握，テストコレクションについての理解を目的としており，実験参加者が概ね把握できたところでトレーニングを終了してもらった．また，トレーニング中に不明な点があれば質問してもらい，それに答えた．

3.2.3 制約

情報検索行動にかける制約として，本研究では 3 つ用意した．まず 1 つ目は，時間制限である．この制約では，ウェブ検索システムを使える時間を 15 分に制限する．制限時間を超過した場合，その時点でタスクを強制的に終了する．2 つ目は，クエリ発行回数制限である．この制限では，ウェブ検索システムでのクエリの発行回数を 10 回に制限する．発行回数が上限に達すると，それ以降クエリの発行はできなくなる．しかし，最後に発行したクエリによって表示された検索結果の閲覧は可能である．なお，制限するのはあくまでクエリの発行回数のみであり，クエリに使う単語の数，単語の長さ，そして内容に関しては自由とする．そして 3 つ目は，全文閲覧可能回数制限である．この制限では，検索結果一覧から記事全文を参照できる回数を 20 回に制限する．閲覧回数が上限に達すると，その時点でタスクを強制的に終了する．以上，3 つである．以上のように，制限を加えるとは言うものの，一般的なタスクをこなすのには問題のない設定となっており，著しくそれを困難にするものではなく，あくまで実験参加者の意識に与える影響を調べるのが目的として設けられているものである．クエリ発行回数の上限を 10 回，全文閲覧可能回数の上限を 20 回に設定した根拠は，類似したタスクを実施した先行研究 [3][12] の結果による．

実際には，上記 3 つの制限を組み合わせた制限でタスクに取り組んでもらう．その組み合わせは，1) 時間制限，2) 時間制限 + クエリ発行回数制限，3) 時間制限 + 全文閲覧可能回数制限，の 3 つである．これら 3 つの制限について，それぞれ実験参加者が選択したトピックに関してタスクをこなしてもらった．また，それぞれの制限の上限に達していない場合でも，実験参加者がもう十分にタスクをこなした，成果をあげたと感じたときには任意のタイミングでタスクを終了できるようにした．

表 2　実験 1: 使用したトピックのリスト

トピックの ID	トピックのタイトル	トピックの概要
5-018	たばこ，告訴，賠償金	たばこ会社に対する告訴および法定が決定した賠償に関する記事を探したい.
5-008	「ILOVEYOU」，コンピュータ・ウィルス	ラブ・バグ・ウイルスによって引き起こされたパソコンへの被害に関する記事を探したい.
5-044	異常気象，災害，原因	異常気象が原因とみられる災害について検索する.
5-006	クルスク，潜水艦事故，国際援助	ロシアの原子力潜水艦クルスクの沈没事故および救助を待つ状態に関する記事を探したい.
5-016	歴史教科書論争，第二次世界大戦	日本の文部科学省が認可した，異論の多い第二次世界大戦に関する歴史教科書についての記事を探したい.
6-017	研究，取り組み，後天性免疫不全症候群（AIDS）	後天性免疫不全症候群（AIDS）を克服するための取り組みと研究に関する記事を探したい.
6-043	デリバティブ，損失	デリバティブとは何か，またそれによる損失の事例があったら知りたい.

すべてのトピックは NTCIR CLIR より.
また，「デリバディブ，損失」はトレーニングのトピックとして使用.

3.2.4　手続き

表 4 は実験手続きの流れを示している．実験参加者には，はじめにコンピュータの使用歴，使用頻度，ウェブ検索システムの使用頻度についてのアンケートに回答してもらった．その後，実験に移る前に実験システムの使い方とトピックについて簡単に説明をし，トレーニングとしてこちらの提示したトピック，制限についてタスクをこなしてもらった．そしてトレーニング終了後，本実験に移った．本実験の流れは，まず，実験参加者に 6 つあるトピックの中から興味のあるものを 3 つ選択してもらい，タスクをこなす順番を指定してもらった．その後，こちらから制限の提示を行い，タスクにとりかかってもらった．実験は 15 分を 1 セットとし，1 つの制限と 1 つのトピックを対として，タスクをこなしてもらった．15 分間のタスクの後，個別アンケートに回答してもらった．個別アンケートではトピックに対しての事前知識や，クエリ生成の際の意思決定について，検索結果から適切な文書を選択できたか，文書を読む際に注意を払ったか，(主観での) タスクの難易度はどうだったか，また，タスクに取り組むにあたってどのような方針で行動したかを回答してもらった．この，「実験を行う」から「個別アンケートに回答する」という一連の流れを 3 種類の制限についてそれぞれ行なってもらった．そして，3 回のタスクが終わった後には最終アンケートに回答してもらった．最終アンケートでは，3 種類の制限のうち一番思考に負担を掛けた制限はどれか，一番いい成果をあげられたと思う制限はどれかという設問に回

表 3　実験 1: トピックの詳細

タイトル	デリバティブ，損失
トピック ID	6-043
概要	デリバティブとは何か， またそれによる損失の事例が あったら知りたい．
バックグラウンド	デリバティブとは先物やオプションなどの金融派生商品 のことである．1998 年にはヤクルトが 648 億円の損失 を， クレスベール証券が扱っていたプリンストン債は 全額債務不履行になるなど投機的性格が強いということ を如実に表す結果となった．
適合性の判断基準	デリバティブの説明や， それによる損失の事例を含む 記事であれば， 要求を満たす．デリバティブが原因で 訴訟となった事件についての記事であれば， 部分的に 要求を満たす．金融についての一般的な内容の中で， デリバティブに触れているだけの記事は適合しない．

表 4　実験 1: 実験手順

	内容	所要時間
1	エントリーシートの記入	5 分
2	トレーニング	15 分
3	トピックと制約の設定	1 分
4	タスク	15 分
5	個別アンケート	5 分
6	休憩	5 分
7	最終アンケート	5 分
8	謝金の受け渡し	5 分

3〜6 を 3 回繰り返す

答してもらった．実験にはこちらで準備したラップトップ PC を使用した．ウェブブラウザは Firefox を用い，今回の実験のために開発したウェブ検索システムでタスクをこなしてもらった．なお，ブラウザの操作に関して，タブブラウジングとブラウザバックの使用を禁止し，ページの遷移はマウス左ボタンのシングルクリックによるものだけに限定した．また，ここで，制限の加わる順番による学習効果の影響を考慮して，制限を加える順番を実験協力者毎に変え，ローテーションさせる手法をとった．制限 3 種で 6 通りの組み合わせを成し，24 人の実験参加者で計 4 巡させた（表 5）．

3.2.5　実験システム

実験には，本研究のために開発したウェブ検索システムを利用した．図 1(a)，図 1(b)，図 2 はその画面である．

図 1(a) はシステムへのログイン画面であり，ユーザはユーザ ID を入力し，タスクの対象となるトピックを選択し，制約の種類を選択する．各種情報を入力・選択した後，*login* ボタンをクリックし，タスク開始となる．

表 5　コンディションのローテーション

被験者番号	1 回目のセッション	2 回目のセッション	3 回目のセッション
n	C_1	C_2	C_3
n+1	C_1	C_3	C_2
n+2	C_2	C_1	C_3
n+3	C_2	C_3	C_1
n+4	C_3	C_1	C_2
n+5	C_3	C_2	C_1

図 1(b) はシステムのメイン画面である．画面右側には一般的なウェブ検索システムと同様に，クエリ入力エリアと結果表示エリアが設けられている．結果表示エリアには記事の見出し，記事のハイライト，掲載紙面，記事の文字数，記事の掲載日時が表示される．記事の見出しをクリックすると全文閲覧画面（図 2）に遷移する．画面左側にはタスクの情報が表示される．表示される項目は，ユーザ ID，トピック ID，タスクの残り時間であり，これに加え，C_2 ではその時点でのクエリ発行回数と上限が表示され，C_3 ではその時点での全文閲覧回数と上限が表示される．また，画面左下の「終了する」ボタンを押すとその時点でタスクが終了できる．

図 2 は全文閲覧画面である．メイン画面で記事の見出しをクリックするとこのページに遷移する．画面には記事の ID，記事の見出し，掲載紙面，掲載日時，記事の全文，適合性判断ボタンが表示される．ユーザは記事全文を読み，その記事がトピックで適合とされている文書に当てはまるかどうかを判断し，該当する適合性判断ボタンをクリックする．

(a) 実験 1: ユーザインターフェイス（ログイン画面）　(b) 実験 1: ユーザインターフェイス（検索画面）

図 1　実験 1: ユーザインターフェイス（ログイン画面と検索画面）

3.2.6　分析

各制限の振る舞いや成果への影響を分析するために，様々なデータを収集した．分析は，意識分析，行動分析，成果分析の 3 つ行った．意識分析では，各タスク後の個別アンケート中でのタスクに関する 12 個の設問の回答と，最終アンケートの回答の結果を用いた．アンケートではリカートスケールにより回答を収集・分析した．行動分析では，クエリ発行回数 (Query Count)，全文閲覧回数 (View Count)，記事選択時間 (Click Time)，適合性判断時間 (Judge Time)，発行されたクエリ内に含まれる語彙 (Vocabulary)

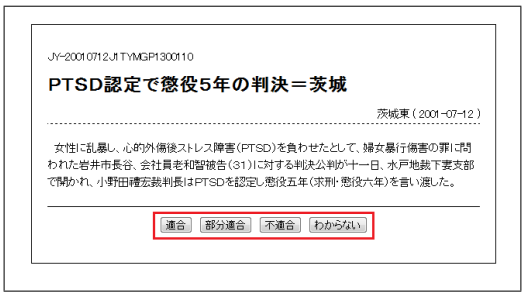


図 2　実験 1: ユーザインターフェイス（全文閲覧画面）

の 5 つのデータを用いた．成果分析には，実験参加者が適合とした文書の数 (User Relevant)，実験参加者が適合とした文書の中で実際に NTCIR のテストコレクションで適合と定められていた文書だった数 (Query Relevant)，実験参加者が適合性判断をした全文書に対する適合とした文書の数の割合 (SCTR URel, Successful Click-Through Rate by User Relevant)，実験参加者が適合性判断をした全文書に対するテストコレクションのそのトピック内で適合とされている文書の数の割合 (SCTR QRel, Successful Click-Through Rate by Query Relevant)，実験参加者が閲覧した文書全体の中で適合だった文書の数との割合 (Precision)，トピック毎に定められている適合文書全体に対するユーザが見つけられた適合文書の割合 (Recall) の，6 つのデータを用いた．また，行動分析，成果分析に関しては，実験参加者がタスクに要した時間を均等に 3 分割 (Bin に分割) し，それぞれの区間内で同様に分析，比較も行った．ここで，行動分析，成果分析に用いるデータは，開発したウェブ検索システムにより収集された．
データの分析の際，t テスト，one-way anova テスト，two-way anova テスト，TukeyHSD を用いた．補正にはボンフェローニ法を利用した．効果量の大小の判断の基準には Atsushi MIZUMOTO と Osamu TAKEUCHI の基準 [17] を用いた．各分析手法とそれによって得られた効果量の組み合わせ，そして解釈の仕方については表 3 の通りである．例えば，one-way anova を用いた多重比較を行った場合，その効果量 r が 0.30 ≦ r<0.50 のとき，効果量は中程度である，という読み方である．

表 6　効果量の指標

Test	Metrics	Small (S)	Midium (M)	Large (L)
ANOVA	η^2	.01	.06	.14
Student's t	r	.10	.30	.50

3.3　結果

本章では，実験結果の行動データおよびアンケートに対して結果を述べる．なお，これ以降示される表内の値の意味は，特に注記のない場合は—AVG は平均値，SD は標準偏差，p は有意水準，es は効果量，n はサンプル数である．

3.3.1 意識への影響

表 7　実験 1: 個別アンケートの結果（N=24）

		C_1:Time	C_2:Query	C_3:View	p	es	size
Q1	私はこのトピックについて精通していた.	1.8 (0.8)	2.2 (0.9)	2.1 (1.1)	.239	.06	*M*
Q2	このトピックは簡単だった.	2.8 (1.3)	3.1 (1.1)	3.1 (1.3)	.626	.02	*S*
Q3	最初のクエリーはすぐに思いついた.	3.9 (1.3)	4.2 (0.9)	4.0 (0.9)	.743	.01	*S*
Q4	その後も新しいクエリーをすぐに思いついた.	3.7 (1.2)	3.3 (1.1)	3.3 (1.0)	.493	.03	*S*
Q5	検索結果から適当な文章を効率的に選択できた.	3.2 (1.0)	3.2 (1.0)	3.5 (1.3)	.450	.03	*S*
Q6	検索結果から選んだ文章は期待していた内容出会った.	3.3 (0.9)	3.6 (1.1)	3.3 (1.4)	.608	.02	*S*
Q7	検索結果から選んだ文章の適合性は容易に判断できた.	3.1 (1.3)	3.4 (1.2)	3.5 (1.4)	.448	.03	*S*
Q8	新しいクエリーを試すか，別の文章を閲覧するかの判断に，迷うことがあった.	2.5 (1.4)	2.7 (1.2)	2.7 (1.4)	.800	.01	*S*
Q9	クエリーの生成に注意を払った.	2.5 (1.4)	3.4 (1.4)	2.0 (1.0)	.001	.24	*L*
Q10	検索結果の閲覧に注意を払った.	3.0 (1.4)	2.8 (1.2)	4.4 (0.6)	.000	.39	*L*
Q11	全文の閲覧に注意を払った.	3.6 (1.3)	3.5 (1.1)	3.6 (1.2)	.934	.00	
Q12	作業中，インターフェイス左側のカラムが気になった.	1.7 (1.1)	2.4 (1.4)	1.8 (1.2)	.007	.19	*L*

1: 全くそうは思わない 3: どちらとも言えない 5: とてもそう思う

表 7 は，各タスク終了後の個別アンケートの回答を分析したものである．アンケートの設問は Q1 から Q12 の 12 個である．評価にはリカートスケールを用い，その評価基準は 1:全くそうは思わない，2:そうは思わない，3:どちらともいえない，4:そう思う，5:とてもそう思う，である．

Q1「私はこのトピックについて精通していた.」という設問に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 1.75(0.85)，クエリ発行回数制限では 2.21(0.93)，全文閲覧可能回数制限では 2.08(1.14) であった．統計分析の結果，p=0.239，es=0.06 であった．p=0.239! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．

Q2「このトピックは簡単だった.」という設問に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 2.79(0.85)，クエリ発行回数制限では 2.21(0.93)，全文閲覧可能回数制限では 3.12(1.30) であった．統計分析の結果，p=0.626，es=0.02 であった．p=0.626! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．

Q3「最初のクエリーはすぐに思いついた.」という設問に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 3.92(1.28)，クエリ発行回数制限では 4.17(0.87)，全文閲覧可能回数制限では 4.00(0.93) であった．統計分析の結果，p=0.743，es=0.013 であった．p=0.743! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．

Q4「その後も新しいクエリーをすぐに思いついた.」という設問に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 3.67(1.17)，クエリ発行回数制限では 3.33(1.09)，全文閲覧可能回数制限では 3.33(1.01) であった．統計分析の結果，p=0.493，es=0.03 であった．p=0.493! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．

Q5「検索結果から適当な文章を効率的に選択できた.」という設問に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 3.17(1.05)，クエリ発行回数制限では 3.17(1.01)，全文閲覧可能回数制限では 3.54(1.25) であった．統計分析の結果，p=0.45，es=0.034 であった．p=0.45! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．

Q6「検索結果から選んだ文章は期待していた内容であった.」という設問に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 3.25(0.94)，クエリ発行回数制限では 3.58(1.06)，全文閲覧可能回数制限では 3.29(1.43) であった．統計分析の結果，p=0.608，es=0.021 であった．p=0.608! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．

Q7「検索結果から選んだ文章の適合性は容易に判断できた.」という設問に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 3.08(1.32)，クエリ発行回数制限では 3.38(1.17)，全文閲覧可能回数制限では 3.54(1.41) であった．統計分析の結果，p=0.448，es=0.034 であった．p=0.448! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．

Q8「新しいクエリーを試すか，別の文章を閲覧するかの判断に，迷うことがあった.」という設問に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 2.46(1.32)，クエリ発行回数制限では 2.67(1.17)，全文閲覧可能回数制限では 2.67(1.43) であった．統計分析の結果，p=0.800，es=0.01 であった．p=0.800! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．

Q9「クエリの生成に注意を払った.」という設問に関しては，特にクエリ発行回数制限においての数値が他の制限と比較した際に差が出ると期待されていた．その平均値 (偏差) は，時間制限では 2.46(1.32)，クエリ発行回数制限では 3.42(1.44)，全文閲覧可能回数制限では 2.04(1.00) であった．統計分析の結果，p=0.172e-2，es=0.242 であった．p=0.172e-2¡0.05 で有意水準 p<0.05 を満たしており，有意な差を示すデータを得られ，効果量も高いことがわかった．多重比較の結果，時間制限とクエリ発行回数制限との間では p=0.14，es=0.743，時間制限と全文閲覧可能回数制限との間では p=0.53，es=0.323，クエリ

発行回数制限と全文閲覧可能回数制限との間では p=0.11e-2，es=1.066 であった．クエリ発行回数制限と全文閲覧可能回数制限との間で p=0.11e-2 で有意水準 p<0.05 を満たしており，効果量も es=1.066 と高いことがわかった．また，時間制限とクエリ発行回数制限との間でも，有意水準には達していない (p=0.14! <0.05) が，比較的高い効果量 (es=0.743) が見られる．これは期待通りの結果であった．

Q10「検索結果の閲覧に注意を払った.」という設問に関しては，特に全文閲覧可能回数制限においての数値が他の制限と比較した際に差が出ると期待されていた．その平均値 (偏差) は，時間制限では 2.83(1.24)，クエリ発行回数制限では 3.00(1.44)，全文閲覧可能回数制限では 4.38(0.65) であった．統計分析の結果，p=0.101e-4，es=0.393 であった．p=0.101e-4 で有意水準 p<0.05 を満たしており，有意な差を示すデータを得られ，効果量も高いことがわかった．多重比較の結果，時間制限とクエリ発行回数制限との間では p=1.00，es=0.743，時間制限と全文閲覧可能回数制限との間では p=0.53，es=0.323，クエリ発行回数制限と全文閲覧可能回数制限との間では p=0.11e-2，es=1.066 であった．クエリ発行回数制限と全文閲覧可能回数制限との間で p=0.11e-2 で有意水準 p<0.05 を満たしており，効果量も es=1.066 と高いことがわかった．また，時間制限とクエリ発行回数制限との間でも，有意水準には達していない (p=0.14! <0.05) が，比較的高い効果量 (es=0.743) が見られる．これは期待通りの結果であった．

Q11「全文の内容に注意を払った.」という設問に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 3.50(1.10)，クエリ発行回数制限では 3.58(1.28)，全文閲覧可能回数制限では 3.58(1.21) であった．統計分析の結果，p=0.934，es=0.003 であった．p=0.934! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．

Q12「作業中，インターフェイス左側のカラムが気になった.」という設問に関しては，特にクエリ発行回数制限においての数値が他の制限と比較した際に差が出ると期待されていた．その平均値 (偏差) は，時間制限では 2.42(1.44)，クエリ発行回数制限では 1.67(1.13)，全文閲覧可能回数制限では 1.83(1.17) であった．統計分析の結果，p=0.871e-2，es=0.193 であった．p=0.871e-2 で有意水準 p<0.05 を満たしており，有意な差を示すデータを得られ，効果量も高いことがわかった．多重比較の結果，時間制限とクエリ発行回数制限との間では p=0.038，es=0.598，時間制限と全文閲覧可能回数制限との間では p=1.00，es=-0.133，クエリ発行回数制限と全文閲覧可能回数制限との間では p=0.082，es=0.465 であった．時間制限とクエリ発行回数制限との間で p=0.038 で有意水準 p<0.05 を満たしており，効果量も es=0.598 と高いことがわかった．また，クエリ発行回数制限と全文閲覧回数制限との間でも，有意水準には達していない (p=0.082! <0.05) が，比較的高い効果量 (es=0.465) が見られる．これは期待通りの結果であった．

表 8 は，3 回のタスクをすべて終えた後に回答してもらった最終アンケートの結果を分析したものである．
Q1「3 種類の制限のなかで，自分の思考に一番負担を与えたと思うものはどれですか.」という設問では，C_4（時間制限 + クエリ発行回数制限）が一番多い回答を集めた．
Q2「3 種類の制限の中で，自分が一番いい成果をあげられたと思うものはどれですか.」という設問では，「時間制限」が一番多い回答を集めた．

表 8　実験 1: 最終アンケートの結果（N=24）

		C_1:Time	C_2:Query	C_3:View
Q1	3 種類の制限のなかで，自分の思考に一番負担を与えたと思うものはどれですか.	1	13	10
Q2	3 種類の制限の中で，自分が一番いい成果をあげられたと思うものはどれですか.	11	7	6

3.3.2 行動への影響

表 9　実験 1: 行動への影響（N=24）

	C_1:Time	C_2:Query	C_3:View	p	es	
Query	11.6 (8.0)	6.9 (3.3)	11.3 (6.8)	.007	.19	*L*
Vocabulary	10.3 (7.2)	7.4 (2.8)	9.6 (4.7)	.090	.050	*S*
View	16.9 (7.7)	20.8 (11.4)	15.0 (4.0)	.023	.15	*L*
Click Interval (sec)	20.8 (13.8)	19.8 (11.3)	21.3 (11.4)	.873	.00	
Judge Time (sec)	29.1 (21.9)	21.9 (19.3)	20.3 (10.2)	.054	.11	*M*

表 10　実験 1: 行動への影響（多重比較）

	C_1 - C_2		C_1 - C_3		C_2 - C_3	
	p	es	p	es	p	es
Query	0.01	-0.73	1.00	0.04	0.02	-0.70
Vocabulary	0.349	0.48	0.602	0.23	0.049	0.71
View	1.00	-0.08	1.00	-0.04	1.00	-0.12
Click Time (sec)	0.25	-0.41	0.11	0.49	1.00	0.09
Judge Time (sec)	0.15	-0.55	0.91	0.12	0.30	-0.43

表 9 は行動への影響の分析の結果である.

クエリ発行回数 (Query Count) に関しては，特にクエリ発行回数制限においての数値が他の制限と比較した際に差が出ると期待されていた．その平均値 (偏差) は，時間制限では 11.58(7.99)，クエリ発行回数制限では 6.92(3.33)，全文閲覧可能回数制限では 11.33(6.84) であった．統計分析の結果，p=0.01，es=0.192 であった．p=0.01 で有意水準 p<0.05 を満たしており，有意な差を示すデータを得られ，効果量も高いことがわかった．多重比較の結果，時間制限とクエリ発行回数制限との間では p=0.01，es=-0.735，時間制限と全文閲覧可能回数制限との間では p=1.00，es=0.039，クエリ発行回数制限と全文閲覧可能回数制限との間では p=0.02，es=-0.70 であった．時間制限とクエリ発行回数制限との間で p=0.01 で有意水準 p<0.05 を満たしており，効果量も es=-0.735 と高いことがわかった．また，クエリ

表 11　実験 1: 行動への影響（Bin による比較）

	Condition	Bin1	Bin2	Bin3
Query	C_1	4.17 (2.84)	3.50 (4.14)	3.88 (4.57)
	C_2	2.87 (1.58)	1.73 (1.79)	1.92 (1.67)
	C_3	4.63 (3.06)	3.38 (2.75)	3.21 (2.54)
Vocabulary	C_1	4.71 (2.69)	5.46 (4.32)	5.78 (3.85)
	C_2	3.96 (1.61)	4.13 (1.71)	4.04 (1.58)
	C_3	5.13 (2.47)	5.25 (2.56)	5.21 (2.43)
Click	C_1	5.46 (3.55)	6.04 (2.84)	5.91 (3.56)
	C_2	5.78 (3.78)	7.83 (.55)	6.50 (4.14)
	C_3	4.29 (2.05)	5.13 (2.40)	5.42 (2.70)
Click Time(sec.)	C_1	16.01 (8.61)	20.66 (16.28)	26.62 (30.90)
	C_2	16.65 (12.17)	21.87 (16.65)	23.77 (18.76)
	C_3	23.06 (13.40)	32.39 (51.66)	24.2 (18.30)
Judge Time(sec.)	C_1	24.61 (15.32)	28.78 (29.36)	27.23 (25.24)
	C_2	21.82 (16.72)	23.61 (24.01)	19.48 (19.72)
	C_3	22.03 (10.64)	20.32 (16.62)	22.84 (15.73)

発行回数制限と全文閲覧可能回数制限との間で p=0.02 で有意水準 p<0.05 を満たしており，効果量も es=-0.70 と高いことがわかった．これは期待通りの結果であった．Bin による分割後の各区間での値の遷移を各コンディション間で比較（図 3(a)）すると，時間制限（time）とクエリ発行回数制限（query）においては，Bin1 がもっとも多く，Bin2 で一旦減少した後に，再び増加するといった傾向が見られる．全文閲覧可能回数制限（view）では，Bin1 が最も多く，Bin2，Bin3 と減少していく傾向が見られる．

全文閲覧回数 (View Count) に関しては，特に全文閲覧可能回数制限においての数値が他の制限と比較した際に差が出ると期待されていた．その平均値 (偏差) は，時間制限では 16.88(7.69)，クエリ発行回数制限では 20.79(11.36)，全文閲覧可能回数制限では 14.96(4.02) であった．統計分析の結果，p=0.02，es=0.151 であった．p=0.02 で有意水準 p<0.05 を満たしており，有意な差を示すデータを得られ，効果量も高いことがわかった．多重比較の結果，時間制限とクエリ発行回数制限との間では p=0.349，es=0.48，時間制限と全文閲覧可能回数制限との間では p=0.602，es=0.23，クエリ発行回数制限と全文閲覧可能回数制限との間では p=0.049，es=0.71 であった．クエリ発行回数制限と全文閲覧可能回数制限との間で p=0.049 で有意水準 p<0.05 を満たしており，効果量も es=0.71 と高いことがわかった．これは期待通りの結果であった．Bin による分割後の各区間での値の遷移を各コンディション間で比較（図 4(a)）すると，時間制限（time）においては，Bin1 が最も少なく，Bin2 で増加し，Bin3 で減少していく傾向が見られる．クエリ発行回数制限（query）においては，Bin1 が最も少なく，Bin2 で極めて多くなり，Bin3 で大きく減少する傾向が見られる．全文閲覧回数制限（view）においては，Bin1 が最も少なく，Bin2，Bin3 と増加していく傾向が見られる．

記事選択時間 (Click Time) に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均

値 (偏差) は，時間制限では 20.82(13.83)，クエリ発行回数制限では 19.82(11.29)，全文閲覧可能回数制限では 21.29(11.41) であった．統計分析の結果，p=0.873，es=0.006 であった．p=0.873! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．Bin による分割後の各区間での値の遷移を各コンディション間で比較（図 4(b)）すると，時間制限（time）においては，Bin1 が最も短く，Bin2，Bin3 と長くなっていく傾向が見られる．クエリ発行回数制限（query）においては，Bin1 が最も短く，Bin2，Bin3 と長くなっていく傾向が見られる．全文閲覧回数制限（view）においては，Bin1 が最も短く，Bin2 で極めて長くなり，Bin3 で大きく減少する傾向が見られる．

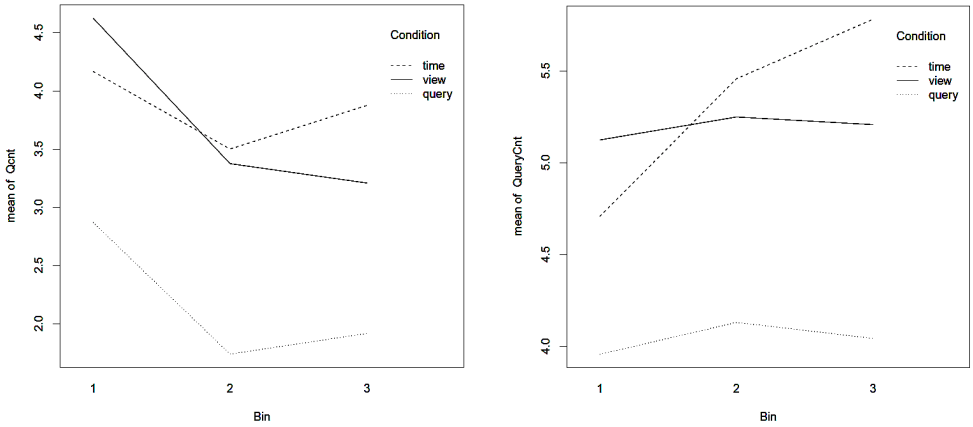
適合性判断時間に (Judge Time) に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 29.12(21.86)，クエリ発行回数制限では 21.87(19.31)，全文閲覧可能回数制限では 20.33(10.16) であった．統計分析の結果，p=0.05，es=0.119 であった．p=0.05! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．しかし，有意水準には満たなかったが，効果量は高かった (es=0.119)．多重比較の結果，時間制限とクエリ発行回数制限との間では p=0.25，es=-0.41，時間制限と全文閲覧可能回数制限との間では p=0.11，es=0.49，クエリ発行回数制限と全文閲覧可能回数制限との間では p=1.00，es=0.09 であった．各制限間での有意性は数値上見いだせないが，時間制限とクエリ発行回数制限との間で高い効果量 (es=-0.41) が見られ，また，時間制限と全文閲覧可能回数制限との間でも，高い効果量 (es=0.49) が見られた．Bin による分割後の各区間での値の遷移を各コンディション間で比較（図 5）すると，それぞれの制限間で違う傾向が見られることがわかる．時間制限（time）においては，Bin1 が最も短く，Bin2 で大きく長くなり，Bin3 で短くなる傾向が見られる．クエリ発行回数制限（query）においては，Bin1 から Bin2 にかけて長くなり，Bin3 で極めて短くなる傾向が見られる．全文閲覧回数制限（view）においては，Bin1 から Bin2 にかけて短くなり，Bin3 で長くなる傾向が見られる．

発行されたクエリ内に含まれる語彙 (Vocabulary) に関しては，特にクエリ発行回数制限においての数値が他の制限と比較した際に差が出ると期待されていた．その平均値 (偏差) は，時間制限では 10.25(7.22)，クエリ発行回数制限では 7.35(2.81)，全文閲覧可能回数制限では 9.63(4.66) であった．統計分析の結果，p=0.09，es=0.055 であった．p=0.055! <0.05 で有意な差を示すデータを得られなかった．しかし，有意水準には満たなかったが，効果量は高かった (es=0.055)．多重比較の結果，時間制限とクエリ発行回数制限との間では p=0.15，es=-0.55，時間制限と全文閲覧可能回数制限との間では p=0.91，es=0.12，クエリ発行回数制限と全文閲覧可能回数制限との間では p=0.30，es=-0.43 であった．各制限間での有意性は数値上見いだせないが，時間制限とクエリ発行回数制限との間で高い効果量 (es=-0.55) が見られ，また，クエリ発行回数制限と全文閲覧可能回数制限との間でも，高い効果量 (es=-0.43) が見られた．これは期待通りの結果であった．Bin による分割後の各区間での値の遷移を各コンディション間で比較（図**??**）すると，時間制限（time）においては，Bin1 が最も少なく，Bin2，Bin3 と増加していく傾向が見られる．クエリ発行回数制限（query）と全文閲覧可能回制限（view）においては，Bin1 が最も少なく，Bin2 で増加し，その後，Bin3 で減少する傾向が見られる．

また，各 Bin ごとにコンディション間での二元配置多重比較による分析も行った (表 12)．どのパラメータでも有意な差は見られなかった．しかし，ClickTime では，p=0.641 で有意な差は出なかったが，es=0.011 と，比較的大きな効果量が見られた．

表 12　実験 1: 行動への影響（二元配置多重比較）

	p	es		
		by Condition	By Bin	By Condition * Bin
Query	0.852	0.024	0.024	0.004
Click	0.600	0.019	0.019	0.009
Click Time	0.641	0.016	0.016	0.011
Judge Time	0.714	0.001	0.001	0.005



(a) クエリ発行回数（Bin による分析）　(b) クエリの語彙（Bin による分析）

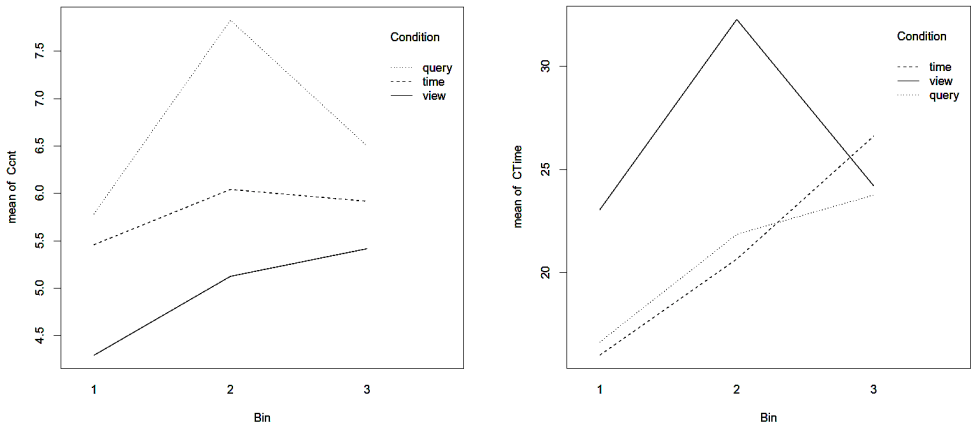
図 3　クエリ発行回数とクエリの語彙の Bin による分析

3.3.3　成果への影響

表 13　実験 1: 成果への影響（N=24）

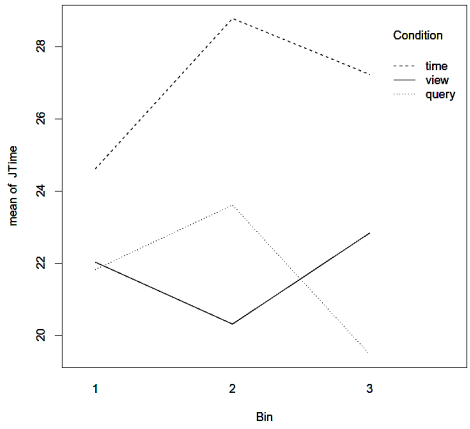
	C_1:Time	C_2:Query	C_3:View	p	es	
URel	13.0 (6.1)	16.4 (11.1)	13.0 (5.0)	.073	.10	*M*
QRel	7.8 (4.8)	12.0 (10.6)	9.5 (5.7)	.138	.08	*M*
Precision	0.61 (0.24)	0.70 (0.24)	0.72 (0.24)	.229	.06	*M*
Recall	0.13 (0.06)	0.15 (0.12)	0.13 (0.09)	.637	.01	*S*
SCTR URel	0.71 (0.15)	0.76 (0.20)	0.78 (0.21)	.461	.03	*S*
SCTR QRel	0.44 (0.20)	0.54 (0.24)	0.58 (0.26)	.126	.08	*M*

表 13 は成果への影響の分析の結果である．



(a) 全文閲覧回数（Bin による分析）　　(b) 記事選択時間（Bin による分析）

図 4　クリック回数と記事選択時間の Bin による分析



(a) 適合性判断時間（Bin による分析）

図 5　適合性判断時間の Bin による分析

実験参加者が適合とした文書の数 (U Rel) に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 12.96(6.12)，クエリ発行回数制限では 16.42(11.14)，全文閲覧可能回数制限では 12.95(5.03) であった．統計分析の結果，p=0.073，es=0.108 であった．p=0.873! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．Bin による分割後の各区間での値の遷移を各コンディション間で比較すると，それぞれの制限間で違う傾向が見られることがわかる．時間制限（time）においては，Bin1 が最も低く，Bin2，Bin3 と増加していく傾向が見られる．クエリ発行回数制限（query）においては，Bin1 から Bin2 にかけて大きく増加し，Bin3 で大きく減少している傾向が見られる．全文閲覧回数制限（view）においては，Bin1 から Bin2 にかけてゆるやかに増加し，Bin2 から Bin3 にかけてゆるやかに減少する傾向が見られる．クエリ発行回数制限と全文閲覧可

表 14　実験 1: 成果への影響（多重比較）

	C_1 - C_2		C_1 - C_3		C_2 - C_3	
	p	es	p	es	p	es
URel	0.33	0.438	1.00	0.106	0.22	0.544
QRel	0.21	0.562	0.85	-0.222	0.90	0.339
Precision	0.50	0.393	0.39	-0.476	1.00	-0.082
Recall	1.00	0.238	1.00	-0.043	1.00	0.195
SCTR URel	0.41	0.429	10.19	-0.585	1.00	-0.156
SCTR QRel	1.00	0.232	0.72	-0.320	1.00	-0.088

表 15　実験 1: 成果への影響（Bin による比較）

	Condition	Bin1	Bin2	Bin3
URel	C_1	3.67 (2.82)	4.25 (2.64)	4.71 (3.52)
	C_2	3.75 (3.69)	6.87 (5.28)	4.79 (4.18)
	C_3	4.22 (2.29)	4.04 (2.26)	3.92 (2.39)
QRel	C_1	2.38 (2.55)	2.50 (1.96)	2.75 (2.01)
	C_2	2.91 (3.65)	4.78 (5.03)	3.75 (3.29)
	C_3	3.08 (2.26)	2.96 (2.33)	3.08 (2.43)
Precision	C_1	.57 (.39)	.55 (.39)	.55 (.28)
	C_2	.63 (.42)	.62 (.36)	.76 (.33)
	C_3	.73 (.34)	.63 (.43)	.69(.32)
Recall	C_1	.037(.031)	.041 (.030)	.045 (.037)
	C_2	.035 (.038)	.061 (.066)	.045 (.035)
	C_3	.044 (.028)	.039 (.035)	.045 (.051)
SCTR URel	C_1	.64 (.32)	.68 (.30)	.73 (.28)
	C_2	.69 (.29)	.87 (.22)	.70 (.34)
	C_3	.83 (.28)	.78 (.25)	.74 (.31)
SCTR QRel	C_1	.44 (.35)	.42 (.32)	.43 (.24)
	C_2	.46 (.38)	.56 (.34)	.57 (.30)
	C_3	.66 (.32)	.51 (.38)	.54 (.28)

能回数制限とを比較すると，Bin 間の大小の関係は同じであるが，その値は大きく違うことがわかる．

実験参加者が適合とした文書の中で実際に NTCIR のテストコレクションで適合と定められていた文書だった数 (Q Rel) に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 7.79(4.79)，クエリ発行回数制限では 12.00(10.61)，全文閲覧可能回数制限では 9.46(5.73) であった．統計分析の結果，p=0.138，es=0.083 であった．p=0.083! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．Bin による分割後の各区間での値の遷移

を各コンディション間で比較すると，それぞれの制限間で違う傾向が見られることがわかる．時間制限（time）においては，Bin1 が最も低く，Bin2，Bin3 と増加していく傾向が見られる．クエリ発行回数制限（query）においては，Bin1 が最も低く，Bin1 から Bin2 にかけて大きく増加し，Bin2 から Bin3 にかけて大きく減少する傾向が見られる．全文閲覧回数制限（view）においては，Bin1 から Bin2 にかけてゆるやかに減少し，Bin2 から Bin3 にかけてゆるやかに増加する傾向が見られ，終始変化が少ないことがわかる．

実験参加者が閲覧した文書全体の中で適合だった文書の数との割合 (Precision) に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 0.61(0.24)，クエリ発行回数制限では 0.70(0.24)，全文閲覧可能回数制限では 0.72(0.24) であった．統計分析の結果，p=0.229，es=0.062 であった．p=0.229! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．Bin による分割後の各区間での値の遷移を各コンディション間で比較すると，それぞれの制限間で違う傾向が見られることがわかる．時間制限（time）においては，Bin1 から Bin2，Bin2 から Bin3 にかけて徐々に減少する傾向が見られる．クエリ発行回数制限（query）においては，Bin1 から Bin2 にかけてゆるやかに減少し，Bin2 から Bin3 にかけて急激に増加する傾向が見られる．全文閲覧回数制限（view）においては，Bin1 が最も高く，Bin1 から Bin2 にかけて大きく減少し，Bin2 から Bin3 にかけて大きく増加するが Bin1 よりかは低くなる傾向が見られる．

トピック毎に定められている適合文書全体に対するユーザが見つけられた適合文書の割合 (Recall) に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 0.13(0.06)，クエリ発行回数制限では 0.15(0.12)，全文閲覧可能回数制限では 0.13(0.09) であった．統計分析の結果，p=0.637，es=0.019 であった．p=0.637! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．Bin による分割後の各区間での値の遷移を各コンディション間で比較すると，それぞれの制限間で違う傾向が見られることがわかる．時間制限（time）においては，Bin1 から Bin2，Bin2 から Bin3 にかけて徐々に増加していく傾向が見られる．クエリ発行回数制限（query）においては，Bin1 ではとても低く，Bin2 で極めて高くなり，Bin2 から Bin3 にかけて大きく減少する傾向が見られる．全文閲覧回数制限（view）においては，Bin1 から Bin2 にかけて減少し，Bin2 から Bin3 にかけて増加し Bin3 が Bin1 よりも高くなる傾向が見られる．

実験参加者が適合性判断をした全文書に対する適合とした文書の数の割合 (SCTR URel, Successful Click-Through Rate on User Relevant) に関しては，3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 0.44(0.20)，クエリ発行回数制限では 0.54(0.24)，全文閲覧可能回数制限では 0.58(0.26) であった．統計分析の結果，p=0.126，es=0.086 であった．p=0.086! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．Bin による分割後の各区間での値の遷移を各コンディション間で比較すると，それぞれの制限間で違う傾向が見られることがわかる．時間制限（time）においては，Bin1 が最も低く，Bin2，Bin3 と徐々に増加していく傾向が見られる．クエリ発行回数制限（query）においては，Bin1 から Bin 2 にかけて大きく増加し，Bin2 から Bin3 にかけて大きく減少する傾向が見られる．全文閲覧回数制限（view）においては，Bin 1 から Bin2，Bin3 にかけて減少していく傾向が見られる．

実験参加者が適合性判断をした全文書に対するテストコレクションのそのトピック内で適合とされている文書の数の割合 (SCTR QRel, Successful Click-Through Rate on QRel) に関しては，有 3 つの制限間で特に差が出ないと期待されていた．その平均値 (偏差) は，時間制限では 0.71(0.15)，クエリ発行回数制限では 0.76(0.20)，全文閲覧可能回数制限では 0.78(0.21) であった．統計分析の結果，p=0.461，es=0.033 であった．p=0.461! <0.05 で有意な差を示すデータを得られなかった．これは期待通りの結果であった．Bin による分割後の各区間での値の遷移を各コンディション間で比較すると，それぞれの制限間で違う傾向が見られることがわかる．時間制限（time）においては，Bin1 から Bin2 にかけて減少し，Bin2 から Bin3 にかけて小さく増加する傾向が見られ，終始変化は少ない．クエリ発行回数制限時間制限（query）においては，Bin1 が最も低く，Bin1 から Bin2 にかけて大きく増加し，Bin2 から Bin3 にかけて小さく増加する傾向が見られる．全文閲覧回数制限においては（view），Bin1 でとても高い値を示し，Bin1 から Bin2 にかけて大きく減少し，Bin2 から Bin3 にかけて増加する傾向が見られる．時間制限と全文閲覧可能回数制限とを比較すると，Bin 間の大小関係は同じであるが，値やその変動の幅は大きく異なっている．

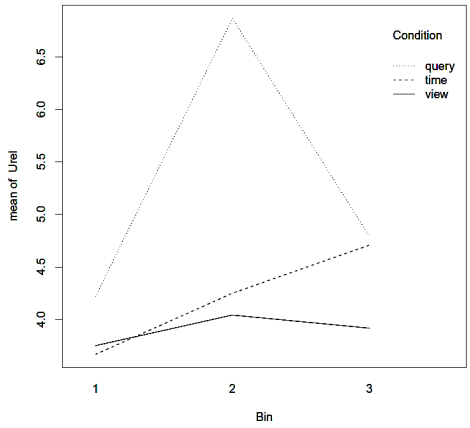
また，各 Bin ごとにコンディション間での二元配置多重比較による分析も行った (表 16)．どのパラメータでも有意な差は見られなかった．しかし，UserRelevant では，p=0.0765 で有意な差は出なかったが，es=0.024 と，比較的大きな効果量が見られた．同様に，SCTRUrel では，p=0.143 で優位な差は出なかったが，es=0.029 と，比較的大きな効果量が見られた．その他でも，比較的高い効果量が見られるものが幾つかあった．

表 16　実験 1: 成果への影響（二元配置多重比較）

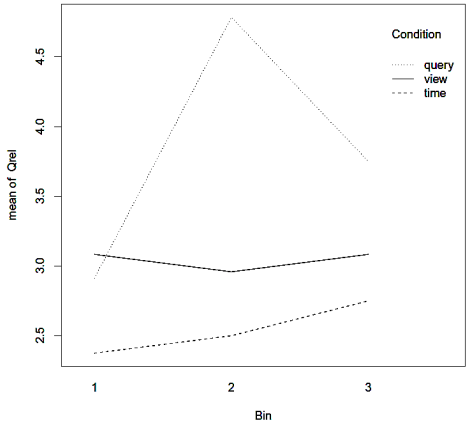
	p	es		
		by Condition	By Bin	By Condition * Bin
URel	.077	.032	.021	.024
QRel	.218	.031	.008	.016
Precision	.529	.024	.006	.009
Recall	.190	.004	.008	.002
SCTR URel	.143	.020	.008	.029
SCTR QRel	.225	.033	.001	.020

3.3.4　その他

個別アンケートでは，タスクに取り組むにあたって立てた方針についても記述してもらった．最終アンケートでは，実験全体を通しての感想も記述してもらった．ここではその中で特に多かったものを取り上げる．クエリ発行回数制限時に立てた方針としては，「ひとつのクエリから十分な情報を得るように心がけた」「質の高いクエリで検索しようと努力した」「クエリの生成に時間を掛け，少ないクエリで効率的に情報を得られるよう努力しました」など，クエリの「質」に注意してタスクに取り組んだという意見が多かった．また，「始めはなるべく検索結果の範囲が広いクエリを選んで，あとからより限定的な結果が得られるクエリを用いた．」「最初に簡単なクエリで検索し全文を見て情報を調べ，多く出て

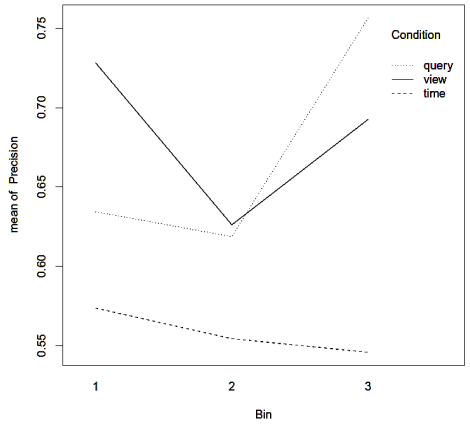


(a) URel（Bin による分析）

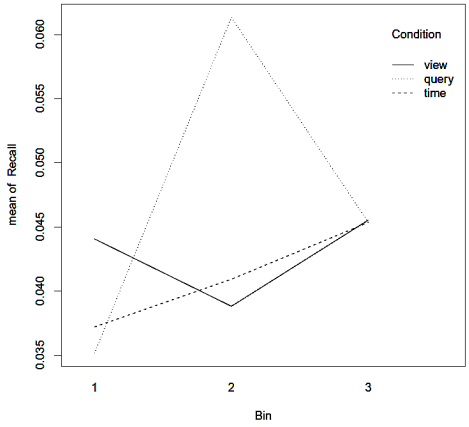


(b) QRel（Bin による分析）

図 6　URel と QRel の Bin による分析



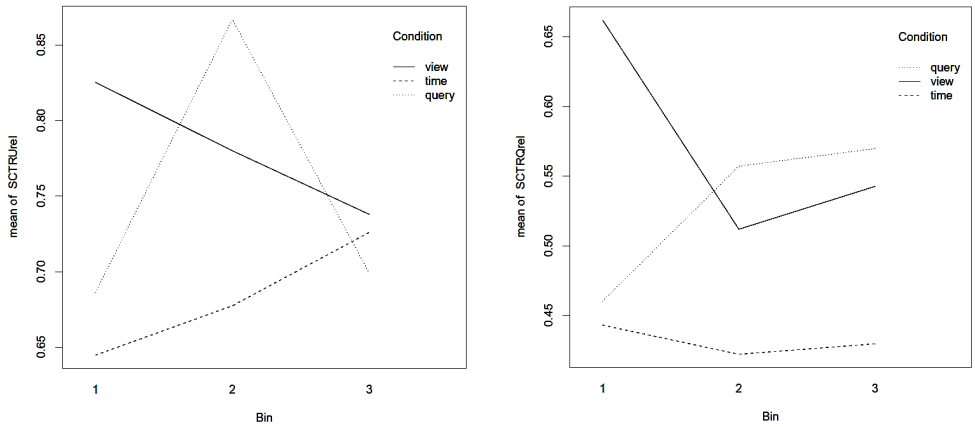
(a) Precision（Bin による分析）



(b) Recall（Bin による分析）

図 7　Precision と Recall の Bin による分析

きた単語などからよりよい結果が出そうなクエリを考えて実行した」など，クエリ生成のプロセスについて注意したという意見も多く見られた．全文閲覧回数制限時に立てた方針としては，「閲覧回数に注意したため，要約文からできる限り内容を推測してから，閲覧するように心がけた.」「文字数を気にするようになった.」「記事のタイトルや文字数から有意なものであるかを判断してから閲覧するようにした」「記事のタイトルや検索結果に表示された略文をメインに判断した.」など，検索結果一覧に表示される諸情報から，全文を読む前にある程度適合性の判断をしてから全文を閲覧したという意見が見られた．また，「クエリはどんどん新しいものを生成し，どんどん検索しました.」「検索の一覧からの情報と文字数で内容を推測し，キーワードとなりそうなものからクエリを作成しました.」「なるべく数多くのクエリを検証して，よりよいクエリを見つけることから検索を始めた.」など，質の高いクエリを生成



(a) SCTR URel（Bin による分析）　　(b) SCTR QRel（Bin による分析）

図 8　SCTR URel と SCTR QRel の Bin による分析

することにより適合しそうな文書を検索結果に表示させてから閲覧していったという意見も見られた．

3.4　考察

3.4.1　仮説の検証

この節では，仮説についての検証を行う．ここで改めて，この研究をするにあたって立てた仮説を挙げる．また，検証にあたって，それぞれの仮説について意識への影響，行動への影響，成果への影響として以下のように対応付ける．

- 制約は人の注意力に影響を与える:意識への影響
- 注意力に影響が出ることにより，情報検索行動に影響を与える:行動への影響
- 情報検索行動に影響が出ることにより，タスクの効率に影響を与える:成果への影響

本研究では，仮説を検証するための実験として，制限付きウェブ検索システムを用いた被験者実験を行った．この実験により，ユーザの情報検索行動におけるタスク中の振る舞い (行動)，タスクの成果を定量的に評価した．また，実験後にアンケートを実施することにより，ユーザの意識についても評価した．その結果は第 4 章で述べたとおりである．

これ以降，第 4 章を踏まえ，制限付きウェブ検索システムを利用した際のユーザの意識への影響，行動への影響，成果への影響について考察する．

3.4.2　意識への影響について

Q9「クエリの生成に注意を払った.」という設問で有意な差が得られた理由としては，クエリ発行回数制限の状況下では，発行できるクエリの回数に制限があるため，一度のクエリ発行で得られる検索結果をよりよいものにしようとした思考が働いたものと思われる．

Q10「検索結果の閲覧に注意を払った.」という設問で有意な差が得られた理由としては，全文閲覧可能回数制限の状況下では，記事の全文を参照する回数が制限されているため，検索結果一覧に表示されるハイライトをはじめとする種々の情報を元に記事の全文を予想する必要があるという思考が働いたからだと思われる．

Q12「作業中，インターフェイス左側のカラムが気になった.」という設問で有意な差が得られた理由としては，クエリ発行回数が 10 回に制限されるという制限が，自分の情報検索行動を縛る要因になると特に意識したからだと思われる．これは，全文閲覧可能回数制限の場合には現れていない傾向である．また，Q9 と Q10 を比較して，クエリ発行回数制限の状況下では，時間制限の場合よりも検索結果の閲覧に対する注意が低くなっていることから，クエリ発行回数に強い注意が向いていることがわかる．
この結果より，3 種類の制限の中では，意識への影響は，クエリ発行回数制限によるものが特に強いと解釈できる．

制約による意識への影響について，まとめると，

- Q9 では特にクエリ発行回数制限の影響が強いことがわかった．
- Q10 では特にクエリ発行回数の影響が強いことがわかった．
- Q12 では特にクエリ発行回数の影響が強いことがわかった．

ということが挙げられる．

3.4.3 行動への影響について

クエリ発行回数 (QueryCount) で有意な差が得られた理由としては，クエリ発行回数制限の状況下ではクエリの生成に注意し，極力少ない発行回数で有意な検索結果を得ようと思考するために，クエリ発行回数が他の 2 種の制限と比べて大幅に小さい値を示したからだと思われる．Bin による分割の結果を見ると，全文閲覧可能回数制限においてクエリ発行回数が右肩下がりの傾向を示す結果が出たが，これは，検索結果一覧から適合する文書を探すという行為に注意が向き，検索結果一覧からページランクの低い文書までしっかりと探索するようになり，結果としてクエリ発行回数が減少したのではないかと考えられる．

全文閲覧回数 (ViewCount) で有意な差が得られた理由としては，全文閲覧可能回数制限の状況下では全文を参照できる回数が制限されるため全文閲覧回数が少なくなっていることは自明であるが，それ以上に，クエリ発行回数制限の状況下で全文閲覧回数が他の 2 種の制限と比較して高くなっていることに注目したい．これは，クエリの発行回数が制限されることにより，1 回のクエリ発行に対して閲覧する記事の数が増えているからだと思われる．Bin による分割の結果を見ると，クエリ発行回数制限においては時間の経過と共に全文閲覧回数が増えていっていることがわかるが，これは，タスク開始から暫くはよりよいクエリの生成に努めているため閲覧回数は増えず，よいクエリが生成されてからは次々に記事を閲覧していっているからだと考えられる．

適合性判断時間 (JudgeTime) で有意な差が得られた理由としては，クエリ発行回数制限の状況下ではクエリの生成に注意を払い，有意なクエリを生成できることが他の 2 種の制限よりも多くなり，その結果，適合している文書を検索結果一覧から容易に選べているからだと思われる．Bin による分割の結果を見ると，全文閲覧可能回数制限では Bin3 が一番高い値を示しているが，これは，閲覧可能回数の上限が近づくにつれ，より慎重に記事の適合性判断を行なっているからと考えられる．
この結果より，3 種類の制限の中では，行動への影響は，クエリ発行回数制限によるものが特に強いと解釈できる．

制約による行動への影響について，まとめると，

- クエリの発行回数に関して，C_2: クエリ発行回数制限が特に強い影響を与えることがわかった．
- 全文閲覧回数に関して，C_2: クエリ発行回数制限が特に強い影響を与えることがわかった．
- 発行されたクエリの語彙に関して，C_2: クエリ発行回数制限が特に強い影響を与えることがわかった．
- 制約の種類によって行動の傾向に差が出ることがわかった（Bin による分析より）．

ということが挙げられる．

3.4.4 成果への影響について

実験参加者が適合とした文書の数 (UserRelevant) で高い効果量が見られた理由としては，クエリ発行回数制限においてはクエリの生成に注意をはらうことにより，よいクエリが生成され，検索結果一覧にユーザが適合と判断できる文書が多く表示されたからだと考えられる．この傾向は，実験参加者が適合とした文書の中で実際に NTCIR のテストコレクションで適合と定められていた文書だった数 (SCTRQRel) に関しても見られる．

実験参加者が適合性判断をした全文書に対する適合とした文書の数の割合 (SCTRURel) で高い効果量が見られた理由としては，全文閲覧可能回数制限においてはより慎重に記事の適合性判断を行うようになり，また，それはタスクの経過により身についた知識とあわせ，より厳密に適合性判断を行うようになったため，適合とする文書の割合が減少していったのではないかと考えられる．

実験参加者が適合性判断をした全文書に対する NTCIR のテストコレクションで適合とされている文書の数の割合 (SCTRQRel) で高い効果量が見られた理由としては，時間制限においては特に行動に対する制限が加えられてないためクエリの生成にはじまり，適合性判断にも注意が向かず，有意な文書の検索，適合性判断ができていないからだと考えられる．クエリ発行回数制限においその値が右肩あがりになっている理由としては，タスクの経過によりよいクエリを生成できるようになり，検索結果一覧に表示される記事が有意なものとなり，適合性判断の精度が増したからだと考えられる．

実験参加者が閲覧した文書全体の中で適合だった文書の数との割合 (Precision) で高い効果量が見られた理由としては，時間制限においては特に行動に対する制限が加えられてないためクエリの生成や全

文閲覧前の検索結果一覧から得られる情報による適合性判断に注意が向かず，全文を閲覧してもその記事が適合している文書でない場合が多くなってしまったからだと考えられる．

トピックで定められている適合文書全体から見つけられた適合文書の数との割合 (Recall) の Bin の結果を見ると，クエリ発行回数制限において Bin2 が特に高い値を示していることがわかるが，これは，Bin1 の段階ではよいクエリが生成できなかったが，Bin2 の段階ではよいクエリが生成できており，多くの適合性判断を行えたからだと思う．Bin3 で値が減少しているのは，先に生成したよいクエリによる記事の閲覧・適合性判断が一段落つき，次の別のクエリの生成に移っているからだと考えられる．

この結果より，3 種類の制限の中では，成果への影響は，クエリ発行回数制限によるものが特に強いと解釈できる．

制約による成果への影響について，まとめると，

- クエリの発行回数に関して，C_2: クエリ発行回数制限が特に強い影響を与えることがわかった．
- 全文閲覧回数に関して，C_2: クエリ発行回数制限が特に強い影響を与えることがわかった．
- 発行されたクエリの語彙に関して，C_2: クエリ発行回数制限が特に強い影響を与えることがわかった．
- 制約の種類によって行動の傾向に差が出ることがわかった（Bin による分析より）．

ということが挙げられる．

3.5 リソースの消費傾向の分析と考察

これまでに 探索型検索（曖昧な情報要求，検索対象に関する知識が乏しい状態で始まる検索行動）の研究では，ユーザの情報検索行動を支援するための手法に関する研究がなされてきた．情報検索行動を支援するためにはユーザの探索プロセスの理解が必要不可欠であるが，これまでは情報検索行動全体を通しての「成果」の部分に注目し，その成果を元に探索プロセス自体を後付けで評価するという手法が多くとられてきた．しかしそれでは，成果の伴わない探索プロセスの評価は不可能であった（良い成果があがった情報検索行動の探索プロセスが必ずしも良いものではないし，逆に成果があがらなかった情報検索行動の探索プロセスが悪いものだとも言い切れない）．つまり，従来の手法では「探索プロセス」自体を支援するための知見を得ることはかなわなかった．情報検索行動における探索プロセスを支援するためには，探索プロセスを評価するための指標が必要になる．そしてそのためには，どのような探索プロセスが「良いもの/悪いもの」であるかを理解する必要がある．そこで我々は，情報検索行動の「成果」から切り離された独立した「探索プロセス自体」を評価するための指標がわかれば，探索プロセス自体を評価し，支援することが可能になると考えた．

過去に Xie と Joo は同様にユーザの探索プロセスに関する研究を行った [21] が，これは探索プロセスの「状態遷移」に着目し分析したものであった．対して今回私は探索プロセス内の「リソースの消費傾向」に着目し，収集したデータからタスクの成績が良いセッション/悪いセッションを抽出し，その探索プロセスを分析・比較することにより，探索プロセスを評価するための指標について考察した．

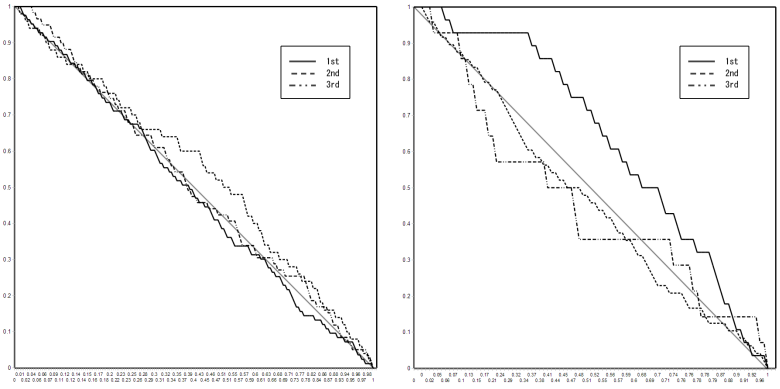
本考察では，ログから得られるクエリ発行，記事選択，適合性判断を「行動」としてとらえ，セッション中に行われた全行動数が「ユーザが所持していたリソース」と考え，それぞれの行動にかかった時間，セッション中に行われた全行動に対するそれぞれの行動の割合を算出し，「行動」を「リソースの消費」として考えて，時間の経過との関係をグラフ上にプロットした．そしてプロットされたそれぞれのグラフを比較し，成績の「良いセッション」と「悪いセッション」のリソースの消費傾向にどのような違いが見られるかを分析し，考察した．

図 9 は C_1 において成績 (= *QRel*) が良かった上位 3 セッションと，悪かった下位 3 セッションのリソースの消費傾向をプロットしたものであり，横軸が *time*[%]，縦軸が *resource*[%] である．

図 10(a) と図 10(b) はそれぞれ，C_2，C_3 において成績が悪かった下位 3 セッションのリソースの消費傾向をプロットしたもので，横軸が *time*[%]，縦軸が *resource*[%] である．

表 17 は，各制約におけるユーザのクエリ発行，記事選択，適合性判断の各行動から行動の遷移に要した時間の割合を示したものである．ここで示す値は，各制約における成績が良かった上位 3 セッションの平均値，成績の悪かった下位 3 セッションの平均値である．

図 9，図 10 中には，リソースの消費傾向を分析する際のリファレンスラインとして， *time* = *resource* のラインも示した．



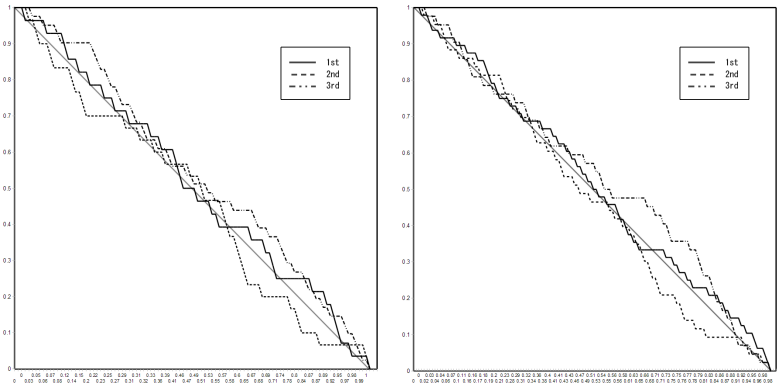
(a) best 3 ユーザーズセッション on C_1　　(b) worst 3 ユーザーズセッション on C_1

図 9　best/worst 3 ユーザーズセッション on C_1

表 17　実験 1: 行動のインターバルで見たリソースの消費傾向

interval[s]	C_1		C_2		C_3	
	best 3	worst 3	best 3	worst 3	best 3	worst 3
< 5	.263	.148	.422	.170	.309	.128
5 < 10	.241	.170	.205	.034	.242	.241
10 < 15	.165	.148	.115	.239	.134	.120
15 <	.331	.534	.258	.557	.315	.511

図 9 は，今回の分析のベースラインとなる C_1 におけるリソースの消費傾向を示している．プロッ



(a) worst 3 ユーザーズセッション on C_2　　(b) worst 3 ユーザーズセッション on C_3

図 10　worst 3 ユーザーズセッション on C_2/C_3

トされたラインから読み取れることとして，まず，成績の良いセッション（図 9(a)）では，リソースの消費が時間の経過に伴ってコンスタントに行われているということだ．したがって，消費傾向を示すラインがリファレンスラインに近いものになっている．逆に成績の悪いセッション（図 9(b)）では，リソースの消費が局所的に行われている（各行動の間の時間が長かったり短かったり，「波」がある）ということがわかる．結果的に，消費傾向を示すラインがリファレンスラインから大きく外れたものになっている．

C_2，C_3 に関しても，成績の良いセッションのリソース消費傾向は図 9(a) で示したものと似た傾向を示した．ここから，成績の良いセッションは制約の種類に関わらず，一定の傾向を示すということがわかる．対して図 10(a)，図 10(b) は C_2 と C_3 における成績の悪いセッション下位 3 つを示したものであるが，これらは図 9(b) で示されたものと傾向が違うことが見て取れる．C_1 と C_2，C_3 の成績の悪いセッションのリソース消費傾向を比較した際，C_1 におけるものよりも C_2，C_3 におけるもののほうがリソースの消費はコンスタントに行われており，また，リファレンスラインにより近いものを示していることがわかる．これらより，制約の存在は成績の悪いセッションの探索プロセスを，成績の良いセッションの探索プロセスのものにより近づけさせる作用があると推測できる．

表 17 は，各制約で成績が良かった/悪かった 3 セッションの各行動と行動の間隔がどのように分布しているかを示したものである（例えば，C_4 でのベスト 3 セッションでは，行動と行動の間にかかった時間は 5 秒以上 10 秒未満だったものが 24 %程度の割合であった，というような読み方）．

表から，成績の良いセッションは 間隔 が短くなる傾向があることがわかる．これは，各行動と行動の間が短い，行動にかける時間が少ないということである．対して成績の悪いセッションは， 間隔 が長くなる傾向があることがわかる．制約ごとにみると，成績の良いセッションで比較した際，C_2 が特にインターバルが短くなる傾向があることがわかる．対して成績の良くないセッションで比較した際，そういった傾向は特には見られないが，図 9(b) と図 10 を比較してみると，リソースの消費傾向には差異があることがわかる．これは，各行動間のインターバルの傾向は似ているが，それらの分布の仕方によってグラフの形が変わったからだと思われる．つまり，単純にインターバルの割合を比較するだけでは分からない制約の間での差があるということである．表 17 では制約の有無によって探索プロセスに影響が出たかどうかは読み取れないが，図 9，図 10 を見るとそこには確かに何らかの影響が働いていることがわかる．

図 9，図 10，表 17 を見て C_1 と C_2，C_3 とを比較すると，C_2 がユーザの情報探索プロセスに特に強い影響を与えていると解釈できる．

ここでは，探索型検索における情報検索行動の成果から独立した，探索プロセスを評価するための指標を考えるために，探索プロセスの良い/悪いについて考察した．結果，成績の良いセッションの探索プロセスは，コンスタントにリソースを消費しており，対して，成績の悪いセッションでは，リソースの消費が局所的に行われていることがわかった．また，制約の存在によって，成績の悪いセッションのリソースの消費傾向が，成績の良いセッションのリソースの消費傾向に近づくことがわかった．

3.6 実験 1 のまとめ

実験 1 では，制約が与える情報検索行動について複数の制約を用いて実験を行い比較し，各制約が与える意識，行動，成果への影響を検証した．その結果としては，

- 意識への影響は，3 種類の制約の中では，C_2（クエリ発行回数制限）によるものが特に強いことがわかった．
- 行動への影響は，3 種類の制約の中では，C_2（クエリ発行回数制限）によるものが特に強いことがわかった．
- 成果への影響は，3 種類の制約の中では，C_2（クエリ発行回数制限）によるものが特に強いことがわかった．
- 成績の良いセッションの探索プロセスは，コンスタントにリソースを消費していることがわかった．
- 成績の悪いセッションでは，リソースの消費が局所的に行われていることがわかった．
- 制約の存在によって，成績の悪いセッションのリソースの消費傾向が，成績の良いセッションのリソースの消費傾向に近づくことがわかった．

といったことが挙げられる．

4　実験 2: 戦略性の成長

実験 1 では，時間，クエリ発行回数，全文閲覧回数などの制約が与える情報検索行動について実験を行い比較し，各制約が与える意識，行動，成果への影響を検証した．その結果，制約は確かに意識，行動，成果に影響を与えることが確かめられ，特に C_2（クエリ発行回数制限）が強い影響を与えることがわかった．また，リソースの消費傾向の分析と考察も行い，良いセッションと悪いセッションとでその傾向に違いがあることも確かめた．この分析と考察はリソースの消費傾向を情報検索行動の「戦略性」を評価するための指標として用いるのが目的としてあったが，実験 2 では，この着想を元に実験を行った．

4.1　仮説と目的

実験 1 では，タスク内で利用できるリソースに着目し，利用できるリソースの制限がある場合とない場合と比較して，タスク中の意識，行動，成果にどのような変化が見られるかどうかを被験者内計画で設計した実験により検証した．制約が加わることにより情報検索行動における意識や行動，成果に影響が出ることは実験 1 で検証し，3.5 リソースの消費傾向の分析 では「戦略性」について考察したが，実験 2 では，リソースの制限がある場合とない場合とで，複数回繰り返されるタスクの中で情報検索行動の「戦略性」の変化・発展というところに着目した．「戦略性」の変化・発展は制約の有無によってどのような違いが見られるだろうか．以上より，本研究で立てた仮説は，

- 制約は情報検索行動に影響を与える
- 制約は複数回繰り返される情報検索行動における「戦略性」の変化・発展に影響を与える
- 情報検索行動を複数回繰り返す内に徐々にそのプロセスは効率の良いものになっていくが，制約がある場合はない場合と比較してそれが顕著に現れる

の 2 つである．1 つ目の仮説に関しては実験 1 で述べた．2 つ目と 3 つ目の仮説を検証するために設定した目的としては，

- 複数回繰り返される情報検索行動における「戦略性」の変化・発展の制約の有無による差を調べる

である．

実験 2 では，制約の有無 と 試行回数 を独立変数として扱い，その他の意識，行動，成果への影響として現れるデータを従属変数とする実験デザインを行った．ここで，制約の有無 とは，ユーザの情報検索行動を制限する制約が有る場合と無い場合とを指し，試行回数 とは，複数回行う情報検索行動（タスク）の回数を指す．制約の有無 と 試行回数 によって意識，行動，成果への影響がどのように変化するか，それを確かめることが実験 2 の目的である．

次節では，この目的の達成のために用いた実験デザインについて詳しく述べる．

4.2 実験デザイン

4.2.1 実験参加者

筑波大学の学部生２９名と大学院生７名あわせて 36 名が実験に参加した．年齢層は，10 代が 5 名，20 代が 31 名であった．男女の内訳は，男性 19 名，女性 17 名であった．参加した学部生，大学院生の専攻する分野とその分布は，情報学系が 18 名，情報工学系が 5 名，人文学・社会学系が 5 名，心理学系が２名，理工学系が 3 名，医学系が 1 名， 芸術系が２名であった．また，学習効果の影響を考え，実験 1 の参加者はこの実験には参加できないこととした．実験参加希望者の募集は，同大学内のメーリングリストを利用し，実験の実施は参加希望者とスケジュールを調整し，順に行なっていった．実験は 2012 年 12 月に行い，実験参加者には謝金として 1,500 円分のアマゾンギフト券を支払った．

4.2.2 タスクとテストコレクション

タスクの対象となるトピックは，こちらの用意したリストから実験参加者に選んでもらった (表 18)．トピックは 6 つ用意してあり，その中から「興味のあるもの」を 3 つ選んでもらった．ここで，トピックによる振る舞い，成果の偏りを考慮して，6 種のトピックは取り扱っている分野や領域の偏りがないように用意した．また，トピックの難易度 (正答文書の数) が近い値のものを選別した．各トピックや適合文書については表 19 のようにテストコレクションで定義されている．

実験参加者にはこのテストコレクションを全文検索できるウェブ検索システムを利用して調べごとをしてもらった．まずトピックを選択してもらい，そのトピックの適合文書のガイドラインを示し，ウェブ検索システムを使い適合文書を探してもらった．実際に実験参加者に伝えたタスクの内容は以下のようなものである．

「あなたは今，大学の講義で出た課題に取り組もうとしています．その課題とは，あるトピックについてウェブ検索システムを使って詳しく調べるというものです．先生からは，トピックのキーワード (タイトル) と，その簡単な概要だけ伝えられました．あなたはその情報を元に，ウェブ検索システムを使ってインターネット上からそのトピックについて述べられている文書を探すことにしました．課題の評価は，トピックに該当する文書をどれだけ多く見つけられたかによってされます．」

実験に移る前に，トレーニングとしてこちらの指定したトピック (表 18 中，「長寿，秘訣，アントニオ・トッディ」) についてのタスクをこなしてもらった．このトレーニングは，タスクの把握，ユーザインターフェイスの把握，テストコレクションについての理解を目的としており，実験参加者が概ね把握できたところでトレーニングを終了してもらった．また，トレーニング中に不明な点があれば質問してもらい，それに答えた．

4.2.3 制約

実験 1 では C_1: タスクに使える時間を 15 分に制限する，C_2: C_1+ 発行できるクエリを 10 回に制限する，C_3: C_1+ 記事を閲覧できる回数を 20 回に制限する，とい 3 つの制限について観察した．これらの制限は，各行動単体に対する制限であったが，実験 2 では「行動全体」に制限をかけ，制約がある場合とない場合とで比較する被験者間計画を採用した．

今回設定した制約は，C_4: タスクに使える時間を 15 分に制限する，C_5: C_4+ クエリの発行回数と記事

表 18　実験 2: 使用したトピックのリスト

トピックの ID	トピックのタイトル	トピックの概要
5-029	代替エネルギー，大気汚染，電力	環境にやさしい発電用代替エネルギーの開発に関する文書を探したい.
5-038	コソボ紛争，NATO，国連	コソボ紛争における NATO の攻撃と，それに対する国連の対応を検索する.
6-018	ティーンエージャー，社会問題	ティーンエージャーの社会問題を扱っている記事を探したい.
6-023	離婚，家族の不和，批判	離婚，別居など，家族の不和に関する批判について述べている記事を探したい.
6-033	研究，タンパク質	病気を根絶するためのタンパク質の研究に関する記事を探したい.
6-106	飲酒運転についての法規と侵害	飲酒による交通事故で引き起こされる人命の損失や物的な被害，飲酒防止に適用される法律について述べた文書にはどのようなものがあるか.
5-027	長寿，秘訣，アントニオ・トッディ	イタリアのアントニオ・トッディのような長寿の秘訣に関する文書を探している.

すべてのトピックは NTCIR CLIR より.
また，「長寿，秘訣，アントニオ・トッディ」はトレーニングのトピックとして使用.

表 19　実験 2: トピックの詳細

タイトル	長寿，秘訣，アントニオ・トッディ
トピック ID	5-027
概要	イタリアのアントニオ・トッディのような長寿の秘訣に関する文書 を探している.
バックグラウンド	イタリアの世界最長寿の男性は, 兄弟を愛し, 赤ワインを毎日飲めば, 長生きができると語っている.
適合性の判断基準	世界最長寿者やその日課を紹介している, 長寿の秘訣に関する文書 を適合とする．彼らの家庭環境に関する文書は部分的に適合とする． 長寿の村に関する環境調査計画についての文書は不適合とする.

を閲覧できる回数とページネーションできる回数あわせて 30 回に制限する，という 2 種類である．ここで，実験 2 では C_4 を「制約なし」の状態と考えベースラインとして用い，C_5 を「制約あり」の状態と考え，それぞれを比較する.

4.2.4　手続き

表 20 は実験手続きの流れを示している．実験参加者には，はじめにコンピュータの使用歴，使用頻度，ウェブ検索システムの使用頻度についてのアンケートに回答してもらった．その後，実験に移

表 20　実験 2: 実験手順

	内容	所要時間
1	エントリーシートの記入	5 分
2	トレーニング	15 分
3	事前調査書の記入	5 分
4	トピックの設定	1 分
5	タスク	15 分
6	個別アンケート	5 分
7	休憩	5 分
8	最終アンケート	5 分
9	謝金の受け渡し	5 分

4〜7 を 3 回繰り返す

る前に実験システムの使い方とトピックについて簡単に説明をし，トレーニングとしてこちらの提示したトピック，制限についてタスクをこなしてもらった．そしてトレーニング終了後，事前調査書（NASA-TLX[11] を参考に作ったワークロードに関するアンケート）の記入をしてもらい，その後本実験に移った．本実験の流れは，まず，実験参加者に 6 つあるトピックの中から興味のあるものを 3 つ選択してもらい，タスクをこなす順番を指定してもらった．その後，制約ありのグループに属する人は「制約あり」を，制約なしのグループに属する人は「制約なし」を選択し，タスクにとりかかってもらった．15 分間のタスクの後，個別アンケートに回答してもらった．個別アンケートではトピックに対しての事前知識や，クエリ生成の際の意思決定について，検索結果から適切な文書を選択できたか，文書を読む際に注意を払ったか，(主観での) タスクの難易度はどうだったか，また，タスクに取り組むにあたってどのような方針で行動したかを回答してもらった．この，「実験を行う」から「個別アンケートに回答する」という一連の流れを 3 回繰り返してもらった．そして，3 回のタスクが終わった後には最終アンケートに回答してもらった．最終アンケートでは，全 3 回のタスクを，自身が満足できた・目標を達成できたと思う順に並べ替えてください，全 3 回の試行を重ねるにつれ，あなたの検索に関する考え方・方針はどのように変化していきましたか，という質問に回答してもらった．実験にはこちらで準備したラップトップ PC を使用した．ウェブブラウザは Firefox を用い，今回の実験のために開発したウェブ検索システムでタスクをこなしてもらった．なお，ブラウザの操作に関して，タブブラウジングとブラウザバックの使用を禁止し，ページの遷移はマウス左ボタンのシングルクリックによるものと，*Ctrl – F* によるページ内検索だけに限定した．

4.2.5　事前調査書

トレーニング終了後，実験協力者には事前調査書（表 21）に記入してもらった．これは，取り組むタスクについて自身が感じた身体的・精神的負荷がどの程度であるかを測るものであり，NASA-TLX[11] を元に作成した．各タスクの後に行う個別アンケートでも同様の内容で調査を行い，試行を重ねる毎に作業従事者（実験協力者）の身体的・精神的負荷がどのように変化していくかを主観的に評価してもらい，その統計をとることが目的である．

表 21　事前調査書

	設問	指標
Q1	どの程度の知的・知覚的活動（考える，決める，計算する，記憶する，見るなど）を必要としましたか	小/大
Q2	どの程度の身体的活動を必要としましたか	小/大
Q3	仕事のペースや課題が発生する頻度のために感じる時間的切迫感はどの程度でしたか	弱/強
Q4	作業指示者（またはあなた自身）によって設定された課題の目標をどの程度達成できたと思いますか	良/悪
Q5	作業成績のレベルを達成・維持するために，精神的・身体的にどの程度いっしょうけんめいに作業しなければなりませんでしたか	少/多
Q6	作業中に，不安感，落胆，いらいら，ストレス，悩みをどの程度感じましたか	低/高
Q7	さまざまな負荷要因，負荷原因，部分部分の課題内容を総合すると，全体としてどの程度の作業負担を感じましたか	低/高

1〜10 での 10 段階評価

4.2.6　実験システム

実験には，本研究のために開発したウェブ検索システムを利用した．図 11(a)，図 11(b)，図 12 はその画面である．

図 11(a) はシステムへのログイン画面であり，ユーザはユーザ ID を入力し，タスクの対象となるトピックを選択し，制約の有無を選択し，何回目のタスクかを選択する．*Try* の「トレーニング」はトレーニングの際に選択するものである．各種情報を入力・選択した後，*login* ボタンをクリックし，タスク開始となる．

図 11(b) はシステムのメイン画面である．画面右側には一般的なウェブ検索システムと同様に，クエリ入力エリアと結果表示エリアが設けられている．結果表示エリアには記事の見出し，記事のハイライト，掲載紙面，記事の文字数，記事の掲載日時が表示される．記事の見出しをクリックすると全文閲覧画面（図 12）に遷移する．画面左側にはタスクの情報が表示される．表示される項目は，ユーザ ID，トピック ID，タスクの残り時間であり，これに加え，C_5 ではその時点でのクエリ発行回数，クリック回数，結果ページ移動回数（ページネーション回数），そしてその合計回数が表示される．画面左下の「終了する」ボタンは，ログイン画面の *Try* で「トレーニング」を選択したときのみ表示され，このボタンを押すとその時点でタスクが終了できる．

図 12 は全文閲覧画面である．メイン画面で記事の見出しをクリックするとこのページに遷移する．画面には記事の ID，記事の見出し，掲載紙面，掲載日時，記事の全文，適合性判断ボタンが表示される．ユーザは記事全文を読み，その記事がトピックで適合とされている文書に当てはまるかどうかを判断し，該当する適合性判断ボタンをクリックする．

(a) 実験 2: ユーザインターフェイス（ログイン画面）　(b) 実験 2: ユーザインターフェイス（検索画面）

図 11　実験 2: ユーザインターフェイス（ログイン画面と検索画面）

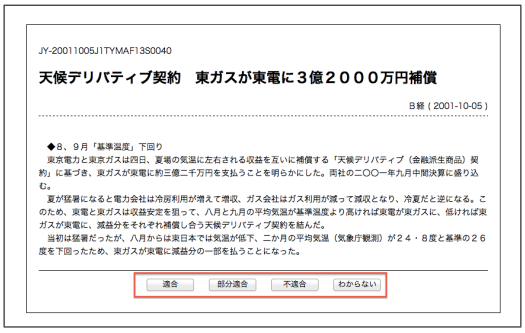


図 12　実験 2: ユーザインターフェイス（全文閲覧画面）

4.2.7　分析

各制限の振る舞いや成果への影響を分析するために，様々なデータを用いた．分析は，意識分析，行動分析，成果分析の 3 つ行った．意識分析では，各タスク後の個別アンケート中でのタスクに関する 11 個の設問の回答と，最終アンケートの回答の結果を用いた．アンケートではリカートスケールにより回答を収集・分析した．また，トレーニング後には NASA-TLX を元に作成したアンケートに回答してもらい，実験協力者のタスクに対する身体的・精神的な負荷についてのデータを集めた．行動分析では，クエリ発行回数 (Query)，発行されたクエリ内に含まれる語彙 (Vocabulary)，記事選択回数 (Click)，ページネーション回数 (Pagenation)，記事選択時間 (Click Time)，適合性判断時間 (Judge Time)，クリックした文書のページランク（Rank）の 6 つのデータを用いた．成果分析には，実験参加者が適合とした文書の数 (UserRelevant)，実験参加者が適合とした文書の中で実際に NTCIR のテストコレクションで適合と定められていた文書だった数 (Q Relevant)，実験参加者が適合性判断をした全文書に対する適合とした文書の数の割合 (SCTR URel, Successful Click-Throught Rate by User Relevant)，実験参加者が適合性判断をした全文書に対するテストコレクションのそのトピック内で適合とされている文書の数の割合 (SCTR QRel, Successful Click-Throught Rate by Query Relevant)，実験参加者が閲覧した文書全体の中で適合だった文書の数との割合 (Precision)，トピック毎に定められている適合文書全体に対するユーザが見つけられた適合文書の割合 (Recall) の，6 つのデータを用いた．ここで，行動分析，成果分析に用いるデータは，開発したウェブ検索システムにより収集された．

データの分析の際，t 検定，one-way anova，two-way anova，TukeyHSD を用いた．補正にはボンフェローニ法を利用した．効果量の大小の判断の基準には Atsushi MIZUMOTO と Osamu TAKEUCHI の基準 [17] を用いた．各分析手法とそれによって得られた効果量の組み合わせ，そして解釈の仕方については表 3 の通りである．例えば，one-way anova を用いた多重比較を行った場合，その効果量 r が 0.30 ≦ r<0.50 のとき，効果量は中程度である，という読み方である．

表 22　効果量の指標（再掲）

Test	Metrics	Small (S)	Midium (M)	Large (L)
ANOVA	η^2	.01	.06	.14
Student's t	r	.10	.30	.50

4.3　結果

本章では，実験結果の行動データおよびアンケートに対して結果を述べる．なお，これ以降示される表内の値の意味は，特に注記のない場合は AVG は平均値，SD は偏差，p は有意水準，es は効果量，n はサンプル数である．

4.3.1　意識への影響

表 23 は，各タスク終了後の個別アンケートの回答を分析したものである．アンケートの設問は Q1 から Q12 の 12 個である．評価にはリカートスケールを用い，その評価基準は 1:全くそうは思わない，2:そうは思わない，3:どちらともいえない，4:そう思う，5:とてもそう思う，である．

Q1「私はこのトピックについて精通していた.」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 2.44(1.03)，2 回目では 2.36(1.10)，3 回目では 2.33(0.83) であった．統計分析の結果，p=0.873，es=0.004 であった．p=0.873! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 2.72(1.23)，2 回目では 2.50(1.29)，3 回目では 2.22(0.81) であった．統計分析の結果，p=0.430，es=0.048 であった．p=0.430! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 2.17(0.71)，2 回目では 2.22(0.88)，3 回目では 2.44(0.86) であった．統計分析の結果，p=0.413，es=0.051 であった．p=0.! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

Q2「このトピックは簡単だった.」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 2.78(0.93)，2 回目では 3.22(3.06)，3 回目では 3.06(1.04) であった．統計分析の結果，p=0.257，es=0.038 であった．p=0.257! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 2.78(0.81)，2 回目では 3.28(1.27)，3 回目では 3.22(1.11) であった．統計分析の結果，p=0.394，es=0.053 であった．p=0.394! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 2.78(3.17)，2 回目では 3.17(1.34)，3 回目では 2.89(0.96) であった．統計分析の結果，p=0.575，es=0.032 であった．p=0.575! <0.05 で有意な差を示すデータを得られな

表 23　実験 2: 個別アンケートの結果（N(total)=36, N(C_4)=18, N(C_5)=18）

			1st	2nd	3rd	p	es	size
Q1	私はこのトピックについて精通していた.	total	2.44 (1.03)	2.36 (1.10)	2.33 (0.83)	.873	.004	
		C_4	2.72 (1.23)	2.50 (1.29)	2.22 (0.81)	.430	.048	*S*
		C_5	2.17 (0.71)	2.22 (0.88)	2.44 (0.86)	.413	.051	*S*
Q2	このトピックは簡単だった.	total	2.78 (0.93)	3.22 (1.29)	3.06 (1.04)	.257	.038	*S*
		C_4	2.78 (0.81)	3.28 (1.27)	3.22 (1.11)	.394	.053	*S*
		C_5	2.78 (1.06)	3.17 (1.34)	2.89 (0.96)	.575	.032	*S*
Q3	最初のクエリーはすぐに思いついた.	total	4.11 (0.82)	3.97 (0.94)	4.08 (0.65)	.698	.010	*S*
		C_4	4.33 (0.49)	4.28 (0.96)	4.11 (0.68)	.522	.034	*S*
		C_5	3.89 (1.02)	3.67 (0.84)	4.06 (0.64)	.368	.057	*S*
Q4	その後も新しいクエリーをすぐに思いついた.	total	3.44 (0.97)	3.39 (1.05)	3.11 (1.12)	.278	.036	*S*
		C_4	3.89 (0.68)	3.56 (1.10)	3.44 (1.20)	.430	.070	*M*
		C_5	3.00 (1.03)	3.22 (1.00)	2.78 (0.94)	.430	.048	*S*
Q5	新しいクエリーを試すか，別の文章を閲覧するかの判断に，迷うことがあった.	total	3.64 (1.10)	3.08 (1.23)	3.11 (1.14)	.008	.129	*M*
		C_4	3.33 (1.14)	2.89 (1.28)	2.94 (1.16)	.180	.096	*M*
		C_5	3.94 (1.00)	3.28 (1.18)	3.28 (1.13)	.048	.163	*L*
Q6	クエリの生成に注意を払った.	total	3.28 (1.14)	3.14 (1.02)	3.14 (0.99)	.737	.009	
		C_4	3.06 (1.21)	3.11 (1.02)	3.06 (0.87)	.473	.001	
		C_5	3.50 (1.04)	3.17 (1.04)	3.22 (1.11)	.473	.043	*S*
Q7	検索結果から適当な文章を効率的に選択できた.	total	2.86 (0.96)	3.22 (0.99)	3.25 (0.91)	.116	.060	*M*
		C_4	2.89 (0.90)	3.56 (0.92)	3.39 (0.85)	.589	.138	*M*
		C_5	2.83 (1.04)	2.89 (0.96)	3.11 (0.96)	.589	.031	*S*
Q8	検索結果から選んだ文章は期待していた内容であった.	total	3.08 (0.97)	3.14 (0.99)	3.50 (0.97)	.114	.060	*M*
		C_4	3.28 (0.75)	3.22 (0.94)	3.61 (0.92)	.291	.057	*S*
		C_5	2.89 (1.13)	3.06 (1.06)	3.39 (1.04)	.291	.070	*M*
Q9	検索結果から選んだ文章の適合性は容易に判断できた.	total	2.50 (1.03)	3.19 (1.14)	3.33 (1.04)	.003	.157	*L*
		C_4	2.44 (0.92)	3.56 (0.92)	3.56 (1.04)	.363	.342	*L*
		C_5	2.56 (1.15)	2.83 (1.25)	3.11 (1.02)	.363	.058	*S*
Q10	検索結果の閲覧に注意を払った.	total	3.86 (0.96)	3.81 (0.92)	3.78 (0.68)	.907	003	
		C_4	3.67 (1.19)	3.89 (0.90)	3.78 (0.43)	.691	.022	*S*
		C_5	4.06 (0.64)	3.72 (0.96)	3.78 (0.89)	.464	.044	*S*
Q11	作業中，インターフェイス左側のカラムが気になった.	total	2.33 (1.39)	2.11 (1.14)	2.17 (1.18)	.553	.017	*S*
		C_4	2.22 (1.48)	1.67 (0.77)	2.06 (1.16)	.081	.137	*M*
		C_5	2.44 (1.34)	2.56 (1.29)	2.28 (1.23)	.715	.020	*S*

1: 全くそうは思わない, 3: どちらともいえない, 5: とてもそう思う

かった．これらは期待していた通りの結果であった．

Q3「最初のクエリーはすぐに思いついた.」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 4.11(0.82)，2 回目では 3.97(0.94)，3 回目では 4.08(0.65) であった．統計分析の結果，p=0.698，es=0.010 であった．p=0.698! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 4.33(0.49)，2 回目では 4.28(0.96)，3 回目では 4.11(0.68) であった．統計分析の結果，p=0.522，es=0.034 であった．p=0.522! <0.05 で有意な差を示

すデータを得られなかった．C_5 の 1 回目では 3.89(1.02)，2 回目では 3.67(0.84)，3 回目では 4.06(0.64) であった．統計分析の結果，p=0.368，es=0.057 であった．p=0.368! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

Q4「その後も新しいクエリーをすぐに思いついた．」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 3.44(0.97)，2 回目では 3.39(1.05)，3 回目では 3.11(1.12) であった．統計分析の結果，p=0.278，es=0.036 であった．p=0.278! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 3.89(0.68)，2 回目では 3.56(1.10)，3 回目では 3.44(1.20) であった．統計分析の結果，p=0.430，es=0.070 であった．p=0.430! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 3.00(1.03)，2 回目では 3.22(1.00)，3 回目では 2.78(0.94) であった．統計分析の結果，p=0.430，es=0.048 であった．p=0.430! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

Q5「新しいクエリーを試すか，別の文章を閲覧するかの判断に，迷うことがあった．」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 3.64(1.10)，2 回目では 3.08(1.23)，3 回目では 3.11(1.14) であった．統計分析の結果，p=0.008，es=0.129 であった．p=0.08 で有意水準 p<0.05 を満たしており，効果量も中程度であった．C_4 の 1 回目では 3.33(1.14)，2 回目では 2.89(1.28)，3 回目では 2.94(1.16) であった．統計分析の結果，p=0.180，es=0.096 であった．p=0.180! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 3.94(1.00)，2 回目では 3.28(1.18)，3 回目では 3.28(1.13) であった．統計分析の結果，p=0.048，es=0.163 であった．p=0.048 で有意水準 p<0.05 を満たしており，効果量も高かった．これらは期待していた通りの結果であった．

Q6「クエリの生成に注意を払った．」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 3.28(1.14)，2 回目では 3.14(1.02)，3 回目では 3.14(0.99) であった．統計分析の結果，p=0.737，es=0.009 であった．p=0.737! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 3.06(1.21)，2 回目では 3.11(1.02)，3 回目では 3.06(0.87) であった．統計分析の結果，p=0.473，es=0.001 であった．p=0.473! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 3.50(1.04)，2 回目では 3.17(1.04)，3 回目では 3.22(1.11) であった．統計分析の結果，p=0.473，es=0.043 であった．p=0.! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

Q7「検索結果から適当な文章を効率的に選択できた．」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 2.86(0.96)，2 回目では 3.22(0.99)，3 回目では 3.25(0.91) であった．統計分析の結果，p=0.116，es=0.060 であった．p=0.116! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 2.89(0.90)，2 回目では 3.56(0.92)，3 回目では 3.39(0.85) であった．統計分析の結果，p=0.589，es=0.138 であった．p=0.589! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 2.83(1.04)，2 回目では 2.89(0.96)，3 回目では 3.11(0.96) であった．統計分析の結果，p=0.589，es=0.31 であった．p=0.589! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であっ

た．

Q8「検索結果から選んだ文章は期待していた内容であった．」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 3.08(0.97)，2 回目では 3.14(0.99)，3 回目では 3.50(0.97) であった．統計分析の結果，p=0.114，es=0.060 であった．p=0.114! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 3.28(0.75)，2 回目では 3.22(0.94)，3 回目では 3.61(0.92) であった．統計分析の結果，p=0.291，es=0.057 であった．p=0.291! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 2.89(1.13)，2 回目では 3.06(1.06), 3 回目では 3.39(1.04) であった．統計分析の結果, p=0.291, es=0.070 であった．p=0.! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

Q9「検索結果から選んだ文章の適合性は容易に判断できた．」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 2.50(1.03)，2 回目では 3.19(1.14)，3 回目では 3.33(1.04) であった．統計分析の結果，p=0.003，es=0.157 であった．p=0.003 で有意水準 p<0.05 を満たしており，効果量も高かった．C_4 の 1 回目では 2.44(0.92)，2 回目では 3.56(0.92)，3 回目では 3.56(1.04) であった．統計分析の結果，p=0.363，es=0.342 であった．p=0.363! <0.05 で有意な差を示すデータを得られなかったが，効果量は高かった．C_5 の 1 回目では 2.56(1.15), 2 回目では 2.83(1.25), 3 回目では 3.11(1.02) であった．統計分析の結果, p=0.363, es=0.058 であった．p=0.363! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

Q10「検索結果の閲覧に注意を払った．」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 3.86(0.96)，2 回目で 3.81(0.92) は，3 回目では 3.78(0.68) であった．統計分析の結果，p=0.907，es=0.003 であった．p=0.! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 3.67(1.19)，2 回目では 3.89(0.90)，3 回目では 3.78(0.43) であった．統計分析の結果，p=0.691，es=0.022 であった．p=0.691! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 4.06(0.64)，2 回目では 3.72(0.96)，3 回目では 3.78(0.89) であった．統計分析の結果，p=0.464，es=0.044 であった．p=0.464! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

Q11「作業中，インターフェイス左側のカラムが気になった．」という設問に関しては，C_4 と C_5 の間で特に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 2.33(1.39)，2 回目では 2.11(1.14)，3 回目では 2.17(1.18) であった．統計分析の結果，p=0.553，es=0.017 であった．p=0.553! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 2.22(1.48)，2 回目では 1.67(0.77)，3 回目では 2.06(1.16) であった．統計分析の結果，p=0.081，es=0.137 であった．p=0.081! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 2.44(1.34)，2 回目では 2.56(1.29)，3 回目では 2.28(1.23) であった．統計分析の結果，p=0.715，es=0.020 であった．p=0.715! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

表 24　実験 2: 最終アンケートの結果（N(total)=36, N(C_4)=18, N(C_5)=18）

			1*st*	2*nd*	3*rd*
Q	全 3 回のタスクを，自身が満足できた・目標を達成できたと思う順に並べ替えてください．	total	9	12	15
		C_4	3	8	7
		C_5	6	4	8

表 24 は，3 回のタスクを全て終了したあとに回答してもらった最終アンケートの結果を分析したものである．アンケートの設問は，Q) 全 3 回のタスクを，自身が満足できた・目標を達成
できたと思う順に並べ替えてください．であった．C_4 では，2 回目のタスクが一番いい成果をあげられたという人が多く，1 回目のタスクが一番よくない成果だったという人が多かった．C_5 では，3 回目のタスクが一番いい成果を上げられたという人が多く，2 回目のタスクが一番よくない成果だったという人が多かった．

4.3.2　行動への影響

クエリ発行回数 (Query) に関しては，C_5 が C_4 と比較した時に値が小さくなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 8.83(4.72)，2 回目では 7.50(3.89)，3 回目では 9.58(6.11) であった．統計分析の結果，p=0.43，es=0.30 であった．p=0.43 で有意水準 p<0.05 を満たしており，有意な差を示すデータを得られ，効果量も高いことがわかった．C_4 の 1 回目では 10.67(7.50)，2 回目では 7.50(4.02)，3 回目では 10.28(7.65) であった．統計分析の結果，p=0.68，es=0.56 であった．p=0.63! <0.05 で有意な差を示すデータを得られなかったが，効果量は中程度であった．C_5 の 1 回目では 7.00(2.50)，2 回目では 7.50(3.87)，3 回目では 8.89(4.16) であった．統計分析の結果，p=0.037，es=0.50 であった．p=0.037 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られたが，効果量はそれほど高くはなかった．これらは期待していた通りの結果であった．

発行されたクエリの語彙 (Vocabulary) に関しては，C_4 と C_5 とを比較した時に差が出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 7.94(3.45)，2 回目では 7.53(3.66)，3 回目では 8.47(4.55) であった．統計分析の結果，p=0.896，es=0.010 であった．p=0.896! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 8.61(4.27)，2 回目では 7.11(2.91)，3 回目では 8.44(5.36) であった．統計分析の結果，p=0.168，es=0.025 であった．p=0.168! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 7.28(2.30)，2 回目では 7.94(4.33)，3 回目では 8.50(3.73) であった．統計分析の結果，p=0.332，es=0.020 であった．p=0.332! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

クリック回数 (Click) に関しては，C_5 が C_4 と比較した時に値が小さくなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも回数が多くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 11.42(5.47)，2 回目では 13.11(5.73)，3 回目では 13.81(5.08) であった．統計分析の結果，p=0.028，es=0.007 であった．p=0.028 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られ，効果量も高いことがわかった．C_4 の 1 回目では 12.78(6.86)，2 回目では 14.67(6.89)，

表 25　実験 2: 行動への影響（N(total)=36, N(C_4)=18, N(C_5)=18）

		1st	2nd	3rd	p	es	
Query	total	8.83 (4.72)	7.50 (3.89)	9.58 (6.11)	.043	.030	S
	C_4	10.67 (5.71)	7.50 (4.02)	10.28 (7.65)	.068	.056	S
	C_5	7.00 (2.50)	7.50 (3.87)	8.89 (4.16)	.037	.050	S
Vocabulary	total	7.94 (3.45)	7.53 (3.66)	8.47 (4.55)	.896	.010	S
	C_4	8.61 (4.27)	7.11 (2.91)	8.44 (5.36)	.168	.025	S
	C_5	7.28 (2.30)	7.94 (4.33)	8.50 (3.73)	.332	.020	S
Click	total	11.42 (5.47)	13.11 (5.73)	13.81 (5.08)	.028	.034	S
	C_4	12.78 (6.86)	14.67 (6.89)	15.56 (5.88)	.204	.032	S
	C_5	10.06 (3.24)	11.56 (3.87)	12.06 (3.47)	.219	.058	S
Pagenation	total	11.03 (6.76)	11.75 (9.03)	12.61 (8.34)	.001	.007	
	C_4	13.61 (7.39)	15.89 (10.44)	17.17 (9.00)	.228	.027	S
	C_5	8.44 (5.03)	7.61 (4.75)	8.06 (4.28)	.727	.006	
Click Time (sec)	total	28.75 (16.17)	22.85 (10.64)	21.89 (9.45)	.707	.058	S
	C_4	30.00 (20.63)	22.90 (12.91)	22.58 (10.07)	.034	.051	S
	C_5	27.52 (10.46)	22.81 (8.14)	21.20 (9.03)	.068	.082	M
Judge Time (sec)	total	35.56 (18.55)	28.8 (18.31)	21.15 (13.00)	.880	.112	M
	C_4	32.63 (17.46)	28.85 (19.12)	22.98 (16.35)	.044	.051	S
	C_5	38.48 (19.63)	28.91 (18.02)	19.31 (8.54)	.000	.199	L
Rank	total	15.56 (9.85)	16.46 (12.45)	19.01 (22.07)	.137	.009	
	C_4	16.78 (8.97)	19.48 (13.99)	22.90 (28.69)	.577	.018	S
	C_5	14.34 (10.78)	13.43 (10.20)	15.11 (12.17)	.866	.004	

3 回目では 15.56(5.88) であった．統計分析の結果，p=0.204，es=0.032 であった．p=0.204 ! < 0.05 で有意水準を満たしておらず，有意な差を示すデータは得られなかった．C_5 の 1 回目では 10.06(3.24)，2 回目では 11.56(3.87)，3 回目では 12.06(3.47) であった．統計分析の結果，p=0.219，es=0.058 であった．p=0.219 で有意水準 p 0.05 を満たしておらず，有意な差を示すデータは得られなかった．これらの結果は期待していた通りの結果であった．

ページネーション回数 (Pagenation) に関しては，C_5 が C_4 と比較した時に値が小さくなると期待されていた．また，C_4 では 3 回目のタスクのほうが 1 回目のタスクよりも回数が多くなり，C_5 では特に差は出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 11.03(6.76)，2 回目では 11.75(9.03)，3 回目では 12.61(8.34) であった．統計分析の結果，p=0.001，es=0.007 であった．p=0.001 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られ，効果量も高いことがわかった．C_4 の 1 回目では 13.61(7.39)，2 回目では 15.89(10.44)，3 回目では 17.17(9.00) であった．統計分析の結果，p=0.228，es=0.027 であった．p=0.228! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 8.44(5.03)，2 回目では 7.61(4.75)，3 回目では 8.06(4.28) であった．統計分析の結果，p=0.727，es=0.006 であった．p=0.727! <0.05 で有意な差を示すデータを得られなかった．これ

らは期待していた通りの結果であった．

記事選択時間 (Click Time) に関しては，C_4 と C_5 とを比較した時に差が出ないと期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも値が小さくなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 28.75(16.17)，2 回目では 22.85(10.64)，3 回目では 21.89(9.45) であった．統計分析の結果，p=0.707，es=0.058 であった．p=0.707! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 30.00(20.63)，2 回目では 22.90(12.91)，3 回目では 22.58(10.07) であった．統計分析の結果，p=0.034，es=0.051 であった．p=0.034 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られたが，効果量はそれほど高くはなかった．C_5 の 1 回目では 27.52(10.46)，2 回目では 22.81(8.14)，3 回目では 21.20(9.03) であった．統計分析の結果，p=0.068，es=0.082 であった．p=0.068 で有意水準 p<0.05 を満たしておらず，有意な差を示すデータは得られなかったが，効果量は中程度であった．これらは期待していた通りの結果であった．

適合性判断時間 (Judge Time) に関しては，C_4 と C_5 とを比較した時に差が出ないと期待されていた．また，学習効果により，3 回目のタスクのほうが 1 回目のタスクよりも値が小さくなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 35.56(18.55)，2 回目では 28.80(18.31)，3 回目では 21.15(13.00) であった．統計分析の結果，p=0.88，es=0.112 であった．p=0.88! <0.05 で有意な差を示すデータを得られなかったが，効果量は中程度であった．C_4 の 1 回目では 32.63(17.46)，2 回目では 28.85(19.12)，3 回目では 22.98(16.35) であった．統計分析の結果，p=0.044，es=0.051 であった．p=0.044 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られた．C_5 の 1 回目では 38.48(19.63)，2 回目では 28.91(18.02)，3 回目では 19.31(8.54) であった．統計分析の結果，p=0.000，es=0.199 であった．p=0.000 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られ，効果量も高かった．これらは期待していた通りの結果であった．

選択した記事のランク (Rank) に関しては，C_5 が C_4 と比較した時に値が小さくなると期待されていた．また，C_4 では 3 回目のタスクのほうが 1 回目のタスクよりも数値が小さくなり，C_5 では特に差は出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 15.56(9.85)，2 回目では 16.46(12.45)，3 回目では 19.01(22.07) であった．統計分析の結果，p=0.137，es=0.009 であった．p=0.137! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 16.78(8.97)，2 回目では 19.48(13.99)，3 回目では 22.90(28.69) であった．統計分析の結果，p=0.577，es=0.018 であった．p=0.577! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 14.34(10.78)，2 回目では 13.43(10.20)，3 回目では 15.11(12.17) であった．統計分析の結果，p=0.866，es=0.004 であった．p=0.866! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

C_5 では 30 回の行動回数制限の下でタスクを行ったが，C_4 では 15 分間すべてをタスクに使えたため，行動数が C_5 のセッションに比べて多くなったものも少なくなかった．そのため，この 2 つを比較する際，リソース消費傾向の観察という観点から，C_4 のセッションを行動数 30 回の時点で区切ったものを用いた比較も行った（表 26）．

クエリ発行回数 (Query) に関しては，C_5 が C_4 と比較した時に値が小さくなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 7.81(3.38)，2 回目では 6.72(3.36)，3 回目では 8.16(4.31) であった．統計分析の結果，p=0.84，es=0.27 であった．p=0.084! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 8.61(3.99)，2 回目では 5.94(2.65)，3 回目では 7.44(4.45) であった．統計分析の結果，p=0.51，es=0.81 であった．p=0.51! <0.05 で有意な差を示すデータを得られなかったが，効果量は中程度であった．これらは期待していた通りの結果であった．

発行されたクエリの語彙 (Vocabulary) に関しては，C_5 が C_4 と比較した時に多くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 7.28(2.70)，2 回目では 7.17(3.59)，3 回目では 7.75(3.73) であった．統計分析の結果，p=0.540，es=0.006 であった．p=0.540! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 7.28(3.12)，2 回目では 6.39(2.55)，3 回目では 7.00(3.68) であった．統計分析の結果，p=0.489，es=0.015 であった．p=0.489! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

クリック回数 (Click) に関しては，C_5 が C_4 と比較した時に大きくなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも回数が多くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 10.39(4.62)，2 回目では 11.75(4.03)，3 回目では 11.50(3.30) であった．統計分析の結果，p=0.130，es=0.033 であった．p=0.130! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 10.39(4.62)，2 回目では 11.94(4.29)，3 回目では 10.94(3.11) であった．統計分析の結果，p=0.409，es=0.026 であった．p=0.40.9 ! < 0.05 で有意水準を満たしておらず，有意な差を示すデータは得られなかった．これらの結果は期待していた通りの結果であった．

ページネーション回数 (Pagenation) に関しては，C_5 が C_4 と比較した時に値が小さくなると期待されていた．また，C_4 では 3 回目のタスクのほうが 1 回目のタスクよりも回数が多くなり，C_5 では特に差は出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 8.93(5.15)，2 回目では 8.83(4.96)，3 回目では 9.31(4.12) であった．統計分析の結果，p=0.855，es=0.002 であった．p=0.885! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 9.50(5.35)，2 回目では 10.06(4.99)，3 回目では 10.56(3.65) であった．統計分析の結果，p=0.756，es=0.009 であった．p=0.756! <0.05 で有意な差を示すデータを得られなかった．これらは期待していたものとは違った結果となった．

記事選択時間 (Click Time) に関しては，C_4 と C_5 とを比較した時に差が出ないと期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも短くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 28.47(16.78)，2 回目では 22.55(10.34)，3 回目では 21.55(9.15) であった．統計分析の結果，p=0.003，es=0.058 であった．p=0.003 で有意水準 p<0.05 で有意な差を示すデータが得られたが，効果量は低かった．C_4 の 1 回目では 29.49(21.62)，2 回目では 22.29(12.40)，3 回目では 21.90(9.53) であった．統計分析の結果，p=0.046，es=0.051 であった．p=0.046 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られたが，効果量は低かった．これらは期待していた通りの結果であった．

適合性判断時間 (Judge Time) に関しては，C_5 と C_4 とを比較した時に短くなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも値が小さくなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 36.01(18.71)，2 回目では 28.85(18.32)，3 回目では 21.60(13.10) であった．統計分析の結果，p=0.000，es=0.111 であった．p=0.00 で有意水準 p<0.05 で有意な差を示すデータを得られ，効果量も中程度であった．C_4 の 1 回目では 33.51(17.97)，2 回目では 28.78(19.14)，3 回目では 23.89(16.40) であった．統計分析の結果，p=0.057，es=0.049 であった．p=0.059p0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

選択した記事のランク (Rank) に関しては，C_5 が C_4 と比較した時に値が小さくなると期待されていた．また，C_4 では 3 回目のタスクのほうが 1 回目のタスクよりも値が小さくなり，C_5 では特に差は出ないと期待されていた．その平均値 (偏差) は，全体として 1 回目では 14.85(10.36)，2 回目では 15.90(11.67)，3 回目では 17.73(19.64) であった．統計分析の結果，p=0.613，es=0.007 であった．p=0.613! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 15.45(10.18)，2 回目では 18.37(12.79)，3 回目では 20.36(25.13) であった．統計分析の結果，p=0.626，es=0.014 であった．p=0.626! <0.05 で有意な差を示すデータを得られなかった．これらは期待していたものとは違った結果となった．

4.3.3 成果への影響

実験参加者が適合とした文書の数 (User Relevant) に関しては，C_5 が C_4 と比較した時に値が小さくなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも回数が多くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 8.81(4.55)，2 回目では 10.28(5.42)，3 回目では 12.08(5.64) であった．統計分析の結果，p=0.049，es=0.063 であった．p=0.049 で有意水準 p0.05 を満たしており，有意な差を示すデータが得られ，効果量も中程度であった．C_4 の 1 回目では 9.72(5.06)，2 回目では 11.16(6.20)，3 回目では 13.94(6.21) であった．統計分析の結果，p=0.362，es=0.087 であった．p=0.362! <0.05 で有意な差を示すデータを得られなかったが，効果量は中程度であった．C_5 の 1 回目では 7.89(3.89)，2 回目では 9.39(4.53)，3 回目では 10.22(4.44) であった．統計分析の結果，p=0.254，es=0.051 であった．p=0.254! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

実験参加者が適合とした文書の中で実際に NTCIR のテストコレクションで適合と定められて いた文書だった数 (Q Relevant) に関しては，C_5 が C_4 と比較した時に値が小さくなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも回数が多くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 2.78(2.60)，2 回目では 2.81(2.74)，3 回目では 3.58(2.97) であった．統計分析の結果，p=0.043，es=0.018 であった．p=0.043 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られたが，効果量は低かった．C_4 の 1 回目では 3.00(3.01)，2 回目では 2.94(2.86)，3 回目では 4.72(3.30) であった．統計分析の結果，p=0.186，es=0.071 であった．p=0.186! <0.05 で有意な差を示すデータを得られなかったが，効果量は中程度であった．C_5 の 1 回目では 2.56(2.18)，2 回目では 2.67(2.70)，3 回目では 2.44(2.12) であった．統計分析の結果，p=0.963，es=0.000 であった．

表 26　実験 2: 行動への影響（N(total)=36, N(C_4)=18, N(C_5)=18, normalized）

		1st	2nd	3rd	p	es	
	total	7.81 (3.38)	6.72 (3.36)	8.16 (4.31)	.084	.027	S
Query	C_4	8.61 (3.99)	5.94 (2.65)	7.44 (4.45)	.051	.081	M
	C_5	7.00 (2.50)	7.50 (3.87)	8.89 (4.16)	.037	.050	S
	total	7.28 (2.70)	7.17 (3.59)	7.75 (3.73)	.540	.006	
Vocabulary	C_4	7.28 (3.12)	6.39 (2.55)	7.00 (3.68)	.489	.015	S
	C_5	7.28 (2.30)	7.94 (4.33)	8.50 (3.73)	.332	.020	S
	total	10.19 (3.90)	11.75 (4.03)	11.50 (3.30)	.130	.033	S
Click	C_4	10.39 (4.62)	11.94 (4.29)	10.94 (3.11)	.409	.026	S
	C_5	10.06 (3.24)	11.56 (3.87)	12.06 (3.47)	.219	.058	S
	total	8.97 (5.15)	8.83 (4.96)	9.31 (4.12)	.855	.002	
Pagenation	C_4	9.50 (5.35)	10.06 (4.99)	10.56 (3.65)	.756	.009	
	C_5	8.44 (5.03)	7.61 (4.75)	8.06 (4.28)	.727	.006	
	total	28.47 (16.78)	22.55 (10.34)	21.55 (9.15)	.003	.058	S
Click Time (sec)	C_4	29.49 (21.62)	22.29 (12.40)	21.90 (9.53)	.046	.051	S
	C_5	27.52 (10.46)	22.81 (8.14)	21.20 (9.03)	.068	.082	M
	total	36.01 (18.71)	28.85 (18.32)	21.60 (13.10)	.000	.111	M
Judge Time (sec)	C_4	33.51 (17.97)	28.78 (19.14)	23.89 (16.40)	.057	.049	S
	C_5	38.48 (19.63)	28.91 (18.02)	19.31 (8.54)	.000	.199	L
	total	14.85 (10.36)	15.90 (11.67)	17.73 (19.64)	.613	.007	
Rank	C_4	15.45 (10.18)	18.37 (12.79)	20.36 (25.13)	.626	.014	S
	C_5	14.34 (10.78)	13.43 (10.20)	15.11 (12.17)	.866	.004	

p=0.963! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

実験参加者が閲覧した文書全体の中で適合だった文書の数との割合 (Precision) に関しては，C_5 が C_4 と比較した時に高くなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも高くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 0.301(0.251)，2 回目では 0.321(0.286)，3 回目では 0.300(0.202) であった．統計分析の結果，p=0.605，es=0.002 であった．p=0.605! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 0.293(0.225)，2 回目では 0.303(0.248)，3 回目では 0.359(0.211) であった．統計分析の結果，p=0.699，es=0.017 であった．p=0.699! <0.05 で有意な差を示すデータを得られなかった．C_5 の 1 回目では 0.309，2 回目では 0.338，3 回目では 0.241 であった．統計分析の結果，p=0.558，es=0.024 であった．p=0.558! <0.05 で有意な差を示すデータを得られなかった．これらは期待していたものとは違った結果となった．

トピック毎に定められている適合文書全体に対するユーザが見つけられた適合文書の割合 (Recall) に関しては，C_5 が C_4 と比較した時に値が小さくなると期待されていた．また，3 回目のタスクのほ

表 27　実験 2: 成果への影響（N(total)=36, N(C_4)=18, N(C_5)=18）

		1st	2nd	3rd	p	es	
URel	total	8.81 (4.55)	10.28 (5.42)	12.08 (5.64)	.049	.063	M
	C_4	9.72 (5.06)	11.16 (6.20)	13.94 (6.21)	.362	.087	M
	C_5	7.89 (3.89)	9.39 (4.53)	10.22 (4.44)	.254	.051	S
QRel	total	2.78 (2.60)	2.81 (2.74)	3.58 (2.97)	.043	.018	S
	C_4	3.00 (3.01)	2.94 (2.86)	4.72 (3.30)	.186	.071	M
	C_5	2.56 (2.18)	2.67 (2.70)	2.44 (2.12)	.963	.000	
Precision	total	.301 (.251)	.321 (.286)	.300 (.202)	.605	.002	
	C_4	.293 (.225)	.303 (.248)	.359 (.211)	.699	.017	S
	C_5	.309 (.282)	.338 (.325)	.241 (.179)	.558	.024	S
Recall	total	.034 (.033)	.032 (.029)	.050 (.041)	.021	.050	S
	C_4	.037 (.038)	.034 (.030)	.066 (.044)	.045	.133	M
	C_5	.032 (.029)	.030 (.028)	.034 (.032)	.937	.003	
SCTR URel	total	.76 (.21)	.76 (.21)	.86 (.19)	.881	.046	S
	C_4	.76 (.19)	.75 (.21)	.88 (.13)	.031	.100	M
	C_5	.76 (.24)	.78 (.21)	.83 (.23)	.559	.018	S
SCTR QRel	total	.24 (.20)	.22 (.20)	.26 (.20)	.496	.007	
	C_4	.22 (.17)	.21 (.17)	.32 (.21)	.190	.069	M
	C_5	.25 (.23)	.23 (.22)	.20 (.17)	.751	.012	S

うが 1 回目のタスクよりも高くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 0.034(0.033)，2 回目では 0.032(0.029)，3 回目では 0.050(0.041) であった．統計分析の結果，p=0.021，es=0.50 であった．p=0.021 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られたが，効果量は低かった．C_4 の 1 回目では 0.037(0.038)，2 回目では 0.034(0.030)，3 回目では 0.066(0.044) であった．統計分析の結果，p=0.045，es=0.133 であった．p=0.045 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られ，効果量も中程度であった．C_5 の 1 回目では 0.032(0.029)，2 回目では 0.030(0.028)，3 回目では 0.034(0.032) であった．統計分析の結果，p=0.937，es=0.003 であった．p=0.937! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．

実験参加者が適合性判断をした全文書に対する適合とした文書の数の割合 (SCTR URel, Successful Click-Through Rate on User Relevant) に関しては，C_5 が C_4 と比較した時に高くなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも高くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 0.76(0.21)，2 回目では 0.76(0.21)，3 回目では 0.86(0.19) であった．統計分析の結果，p=0.881，es=0.46 であった．p=0.881! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 0.76(0.19)，2 回目では 0.75(0.21)，3 回目では 0.88(0.13) であった．統計分析の結果，p=0.031，es=0.100 であった．p=0.031 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られ，効果量も中程度であった．C_5 の 1 回目では 0.76(0.24)，2 回目では 0.78(0.21)，3 回目で

は 0.83(0.23) であった．統計分析の結果，p=0.559，es=0.018 であった．p=0.559! <0.05 で有意な差を示すデータを得られなかった．これらは期待していたものとは違った結果となった．

実験参加者が適合性判断をした全文書に対するテストコレクションのそのトピック内で適合とされている文書の数の割合 (SCTR QRel, Successful Click-Through Rate on QRel) に関しては，C_5 が C_4 と比較した時に高くなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも高くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 0.24(0.20)，2 回目では 0.22(0.20)，3 回目では 0.26(0.20) であった．統計分析の結果，p=0.496，es=0.007 であった．p=0.496! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 0.22(0.17)，2 回目では 0.21(0.17)，3 回目では 0.32(0.21) であった．統計分析の結果，p=0.190，es=0.069 であった．p=0.190! <0.05 で有意な差を示すデータを得られなかったが，効果量は中程度であった．C_5 の 1 回目では 0.25(0.23)，2 回目では 0.23(0.22)，3 回目では 0.20(0.17) であった．統計分析の結果，p=0.751，es=0.12 であった．p=0.751! <0.05 で有意な差を示すデータを得られなかった．これは期待していたものとは違った結果となった．

C_5 では 30 回の行動回数制限の下でタスクを行ったが，C_4 では 15 分間すべてをタスクに使えたため，行動数が C_5 のセッションに比べて多くなったものも少なくなかった．そのため，この 2 つを比較する際，リソース消費傾向の観察という観点から，C_4 のセッションを行動数 30 回の時点で区切ったものを用いた比較も行った（表 26）．ここでは，C_4 のセッションを行動数 30 回の時点で区切った場合の結果を示す．

実験参加者が適合とした文書の数 (User Relevant) に関しては，C_5 が C_4 と比較した時に値が大きくなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも回数が多くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 7.89(3.63)，2 回目では 9.28(4.40)，3 回目では 10.00(3.87) であった．統計分析の結果，p=0.068，es=0.057 であった．p=0.068! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 7.96(9.17)，2 回目では 9.17(4.40)，3 回目では 9.78(3.32) であった．統計分析の結果，p=0.318，es=0.041 であった．p=0.318! <0.05 で有意な差を示すデータを得られなかった．これらは期待していた通りの結果であった．
実験参加者が適合とした文書の中で実際に NTCIR のテストコレクションで適合と定められて いた文書だった数 (Q Relevant) に関しては，C_5 が C_4 と比較した時に値が大きくなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも回数が多くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 2.36(1.85)，2 回目では 2.58(2.41)，3 回目では 2.83(2.14) であった．統計分析の結果，p=0.673，es=0.08 であった．p=0.673! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 2.22(1.63)，2 回目では 2.50(2.15)，3 回目では 3.22(2.16) であった．統計分析の結果，p=0.356，es=0.045 であった．p=0.356! <0.05 で有意な差を示すデータを得られなかった．これらは期待していたものとは違った結果となった．

実験参加者が閲覧した文書全体の中で適合だった文書の数との割合 (Precision) に関しては，C_5 が C_4 と比較した時に高くなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも高くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 0.30(0.25)，2 回目では 0.32(0.29)，

3 回目では 0.29(0.20) であった．統計分析の結果，p=0.882，es=0.003 であった．p=0.882! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 0.29(0.22)，2 回目では 0.30(0.25)，3 回目では 0.34(0.21) であった．統計分析の結果，p=0.825，es=0.009 であった．p=0.825! <0.05 で有意な差を示すデータを得られなかった．これらは期待していたものとは違った結果となった．

トピック毎に定められている適合文書全体に対するユーザが見つけられた適合文書の割合 (Recall) に関しては，C_5 が C_4 と比較した時に高くなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも高くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 0.029(0.025)，2 回目では 0.030(0.026)，3 回目では 0.040(0.031) であった．統計分析の結果，p=0.255，es=0.29 であった．p=0.225! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 0.027(0.023)，2 回目では 0.030(0.024)，3 回目では 0.045(0.031) であった．統計分析の結果，p=0.143，es=0.0.87 であった．p=0.143! <0.05 で有意な差を示すデータを得られなかったが，効果量は中程度であった．これらは期待していた通りの結果であった．

実験参加者が適合性判断をした全文書に対する適合とした文書の数の割合 (SCTR URel, Successful Click-Through Rate on User Relevant) に関しては，C_5 が C_4 と比較した時に高くなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも高くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 0.76(0.02)，2 回目では 0.76(0.22)，3 回目では 0.86(0.18) であった．統計分析の結果，p=0.041，es=0.45 であった．p=0.045 で有意水準 p0.05 を満たしており，有意な差を示すデータを得られ，．C_4 の 1 回目では 0.76(0.19)，2 回目では 0.76(0.23)，3 回目では 0.89(0.12) であった．統計分析の結果，p=0.057，es=0.95 であった．p=0.057! <0.05 で有意な差を示すデータを得られなかったが，効果量は高かった．これらは期待していたものとは違った結果となった．

実験参加者が適合性判断をした全文書に対するテストコレクションのそのトピック内で適合とされている文書の数の割合 (SCTR QRel, Successful Click-Through Rate on QRel) に関しては，C_5 が C_4 と比較した時に高くなると期待されていた．また，3 回目のタスクのほうが 1 回目のタスクよりも高くなると期待されていた．その平均値 (偏差) は，全体として 1 回目では 0.24(0.20)，2 回目では 0.23(0.20)，3 回目では 0.25(0.19) であった．統計分析の結果，p=0.860，es=0.003 であった．p=0.860! <0.05 で有意な差を示すデータを得られなかった．C_4 の 1 回目では 0.23(0.17)，2 回目では 0.22(0.17)，3 回目では 0.30(0.20) であった．統計分析の結果，p=0.346，es=0.042 であった．p=0.346! <0.05 で有意な差を示すデータを得られなかった．これは期待していたものとは違った結果となった．

4.3.4 その他

個別アンケートでは，タスクに取り組むにあたって立てた方針についても記述してもらった．最終アンケートでは，実験全体を通しての感想も記述してもらった．ここではその中で特に多かったものを取り上げる．

C_4 においては，「単一のクエリで得られた結果から網羅的に閲覧した」「思いついたクエリをすべて試していった」「特に何も考えずに検索を繰り返した」など，情報探索行動に対して強い注意の意識はなかったように見受けられる意見が複数見られた．

表 28　実験 2: 成果への影響（N(total)=36, N(C_4)=18, N(C_5)=18, normalized）

		1st	2nd	3rd	p	es	
URel	total	7.89 (3.63)	9.28 (4.40)	10.00 (3.87)	.068	.047	*S*
	C_4	7.94 (3.57)	9.17 (4.40)	9.78 (3.32)	.318	.041	*S*
	C_5	7.89 (3.89)	9.39 (4.53)	10.22 (4.44)	.254	.051	*S*
QRel	total	2.36 (1.85)	2.58 (2.41)	2.83 (2.14)	.673	.008	
	C_4	2.22 (1.63)	2.50 (2.15)	3.22 (2.16)	.356	.045	*S*
	C_5	2.56 (2.18)	2.67 (2.70)	2.44 (2.12)	.963	.000	
Precision	total	.30 (.25)	.32 (.29)	.29 (.20)	.882	.003	
	C_4	.29 (.22)	.30 (.25)	.34 (.21)	.825	.009	
	C_5	.309 (.282)	.338 (.325)	.241 (.179)	.558	.024	*S*
Recall	total	.029 (.025)	.030 (.026)	.040 (.031)	.255	.029	*S*
	C_4	.027 (.023)	.030 (.024)	.045 (.031)	.143	.087	*M*
	C_5	.032 (.029)	.030 (.028)	.034 (.032)	.937	.003	
SCTR URel	total	.76 (.02)	.77 (.22)	.86 (.18)	.041	.045	*S*
	C_4	.76 (.19)	.76 (.23)	.89 (.12)	.057	.095	*M*
	C_5	.76 (.24)	.78 (.21)	.83 (.23)	.559	.018	*S*
SCTR QRel	total	.24 (.20)	.23 (.20)	.25 (.19)	.860	.003	
	C_4	.23 (.17)	.22 (.17)	.30 (.20)	.346	.042	*S*
	C_5	.25 (.23)	.23 (.22)	.20 (.17)	.751	.012	*S*

C_5 においては，「記事の見出しや文字数に注目して関連する記事を探した」「まず抽象的なクエリで検索し、情報を集めた後に具体的なクエリを生成した」「序盤は探索的な検索を行ったが、徐々に検索すべきものがわかってきた」など，行動回数の制限によってクエリの生成や記事の閲覧に注意が向いているように見受けられる意見が複数見られた．

4.4　考察

4.4.1　仮説の検証

この節では，仮説についての検証を行う．ここで改めて，この実験をするにあたって立てた仮説を挙げる．

- 制約は情報検索行動に影響を与える
- 制約は複数回繰り返される情報検索行動における「戦略性」の変化・発展に影響を与える
- 情報検索行動を複数回繰り返す内に徐々にそのプロセスは効率の良いものになっていくが，制約がある場合はない場合と比較してそれが顕著に現れる

これ以降，複数回繰り返されるタスクの中で情報検索行動の「戦略性」の変化・発展が，制約の有無によってどのような違いを示すかについて考察し，仮説の検証を行う．

4.4.2 意識への影響について

Q5「新しいクエリーを試すか, 別の文章を閲覧するかの判断に, 迷うことがあった.」で優位な差が得られた理由としては，C_5 において 1 回目では特に高い値になっているが，2 回目，3 回目と回を重ねる毎にリソース配分についての学習がなされたため思考にかかる負荷が低くなったためだと考えられる．C_5 においても同様の傾向が見られるが，有意な差を示すほどのものではない．

Q9「検索結果から選んだ文章の適合性は容易に判断できた.」で優位な差が得られた理由としては，タスクを重ねる内に適合性判断についての学習がなされたため，思考にかかる負荷が低なったためだと考えられる．

4.4.3 行動への影響について

クエリ発行回数 (Query) で有意な差が得られた理由としては，C_5 の 1 回目ではリソースが限られているため，クエリの生成に慎重になっているが，回を重ねる毎にリソース配分についての学習がなされ，その結果，「クエリを多く生成したほうがよい」という思考が働き，クエリの発行回数が増えたためだと考えられる．対して C_4 では 1 回目，2 回目，3 回目と規則性は見いだせない．これはクエリ生成に注意が向かず，その場で思いついたまま行動に移しているために試行を重ねる間での変化が見られないからだと考えられる．

クリック数（Click）で有意な差が得られた理由としては，試行を重ねることによる学習効果で，それぞれの行為にかける時間が短くなり，結果的に行為の数が増え，クリック数も増えたのではないかと考えられる．C_4 ではこの傾向が見られるが，C_5 だけを見た場合は顕著な変化は見られない．これは，リソースが限られていることによる行為数の頭打ちが原因だと考えられる．

ページネーション回数（Pagenation）で有意な差が得られた理由としては，試行を重ねることによる学習効果で，それぞれの行為にかける時間が短くなり，結果的に行為の数が増え，クリック数も増えたのではないかと考えられる．C_4 ではこの傾向が見られるが，C_5 だけを見た場合は顕著な変化は見られない．これは，リソースが限られていることによる行為数の頭打ちが原因だと考えられる．

記事選択時間（Click Time）で有意な差が得られた理由としては，試行を重ねることによる学習効果で，それぞれの行為にかける時間が短くなり，結果的に行為の数が増え，クリック数も増えたのではないかと考えられる．C_4 ではこの傾向が見られるが，C_5 だけを見た場合は顕著な変化は見られない．これは，リソースが限られていることにより各行為に注意を払う必要が生じ，検索結果一覧から記事を選択する際にも慎重になり，C_4 の場合よりも時間をかけているからだと考えられる．

適合性判断時間（Judge Time）で有意な差が得られた理由としては，試行を重ねることによる学習効

果で，それぞれの行為にかける時間が短くなり，結果的に行為の数が増え，クリック数も増えたのではないかと考えられる．C_4，C_5 ともにこの傾向が見られるが，C_5 は特に顕著な差が見られる．これは，検索結果一覧から記事を選択する際に慎重に選んだことにより，記事全文を閲覧する前に予めある程度の適合性の判断ができているからだと考えられる．

4.4.4 成果への影響について

クエリ発行回数 (Query) で有意な差が得られた理由としては，C_5 の 1 回目ではリソースが限られているため，クエリの生成に慎重になっているが，回を重ねる毎にリソース配分についての学習がなされ，その結果，「クエリを多く生成したほうがよい」という思考が働き，クエリの発行回数が増えたためだと考えられる．対して C_4 では 1 回目，2 回目，3 回目と規則性は見いだせない．これはクエリ生成に注意が向かず，その場で思いついたまま行動に移しているために試行を重ねる間での変化が見られないからだと考えられる．

クリック数（Click）で有意な差が得られた理由としては，試行を重ねることによる学習効果で，それぞれの行為にかける時間が短くなり，結果的に行為の数が増え，クリック数も増えたのではないかと考えられる．C_4 ではこの傾向が見られるが，C_5 だけを見た場合は顕著な変化は見られない．これは，リソースが限られていることによる行為数の頭打ちが原因だと考えられる．

ページネーション回数（Pagenation）で有意な差が得られた理由としては，試行を重ねることによる学習効果で，それぞれの行為にかける時間が短くなり，結果的に行為の数が増え，クリック数も増えたのではないかと考えられる．C_4 ではこの傾向が見られるが，C_5 だけを見た場合は顕著な変化は見られない．これは，リソースが限られていることによる行為数の頭打ちが原因だと考えられる．

記事選択時間（Click Time）で有意な差が得られた理由としては，試行を重ねることによる学習効果で，それぞれの行為にかける時間が短くなり，結果的に行為の数が増え，クリック数も増えたのではないかと考えられる．C_4 ではこの傾向が見られるが，C_5 だけを見た場合は顕著な変化は見られない．これは，リソースが限られていることにより各行為に注意を払う必要が生じ，検索結果一覧から記事を選択する際にも慎重になり，C_4 の場合よりも時間をかけているからだと考えられる．

適合性判断時間（Judge Time）で有意な差が得られた理由としては，試行を重ねることによる学習効果で，それぞれの行為にかける時間が短くなり，結果的に行為の数が増え，クリック数も増えたのではないかと考えられる．C_4，C_5 ともにこの傾向が見られるが，C_5 は特に顕著な差が見られる．これは，検索結果一覧から記事を選択する際に慎重に選んだことにより，記事全文を閲覧する前に予めある程度の適合性の判断ができているからだと考えられる．

4.4.5 リソースの消費傾向

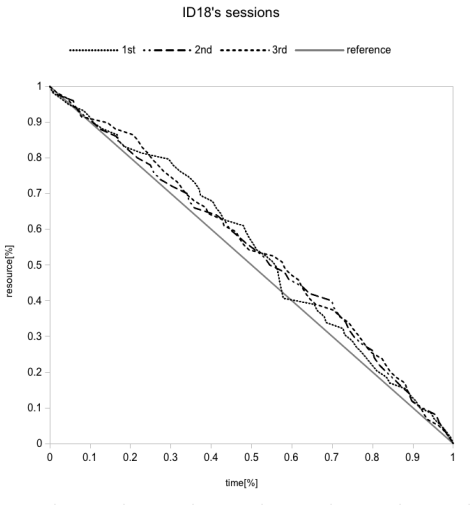
戦略性の変化を見るために，実験 1 で行ったリソース消費傾向の分析をここでも行う．3.5 リソースの消費傾向の分析と考察 では，成績の良いセッションの探索プロセスはコンスタントにリソースを消費

しており，対して，成績の悪いセッションではリソースの消費が局所的に行われていることや，制約の存在によって成績の悪いセッションのリソースの消費傾向が，成績の良いセッションのリソースの消費傾向に近づくことを知見として得られたが，ここではその傾向が複数回繰り返される情報検索行動において制約の有無によってどのような傾向や変化を示すかを考察する．
実験 1 では各制約におけるベスト/ワーストセッションを比較したが，実験 2 では戦略性の変化を見るために，1 回目のタスクでのベスト/ワーストセッションのユーザが，2 回目，3 回目のタスクでどのようなリソース消費傾向を示したかに着目した．

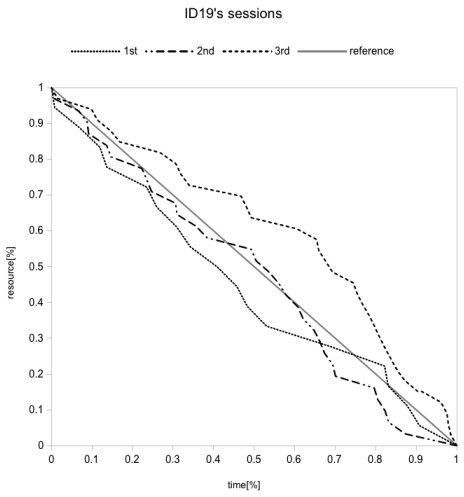
図 13(a)，図 13(b)，図 13(c) は，C_4 の 1 回目のタスクで成績が良かったユーザの，1 回目，2 回目，3 回目のタスクにおけるリソース消費傾向を示したものである．
図 14(a)，図 14(b)，図 14(c) は，C_4 の 1 回目のタスクで成績が悪かったユーザの，1 回目，2 回目，3 回目のタスクにおけるリソース消費傾向を示したものである．
図 15(a)，図 15(b)，図 15(c) は，C_5 の 1 回目のタスクで成績が良かったユーザの，1 回目，2 回目，3 回目のタスクにおけるリソース消費傾向を示したものである．
図 16(a)，図 16(b)，図 16(c) は，C_5 の 1 回目のタスクで成績が悪かったユーザの，1 回目，2 回目，3 回目のタスクにおけるリソース消費傾向を示したものである．
図中には，リソースの消費傾向を分析する際のリファレンスラインとして，*time* = *resource* のラインも示した．

成果が良かったユーザのセッションを C_4 と C_5 との間で比較すると，C_4 の 1 回目のタスクではリファレンスラインから大きく外れているセッションも存在することがわかる．C_5 の 1 回目のタスクでは，波はあるものの，3 セッションともリファレンスラインに寄り添う形で遷移していく傾向があることがわかる．C_4 の 2 回目のタスクでは，1 回目と比較してよりリファレンスラインに寄り添う形で遷移していく傾向が見て取れるが，3 回目のタスクではリファレンスラインから大きく外れるセッションも示されている．しかし，1 回目，2 回目，3 回目ともにリファレンスラインから外れる傾向を示しているセッションは同一のユーザのもののため，個人差ということも考えられる．C_5 の 2 回目のタスクでは，3 セッションともかなり波がある形で遷移していることがわかるが，3 回目のタスクでは 3 セッションとも比較的リファレンスラインと似た傾向で遷移していく傾向が見て取れる．
成果が悪かったユーザのセッションを C_4 と C_5 との間で比較すると，1 回目のタスクではともにリファレンスラインから大きく外れるセッションが示されている．しかし，2 回目，3 回目とタスクを繰り返す内に，徐々にリファレンスラインに近づいていく様子も見て取れる．C_4 と C_5 とを比較した際には，C_5 のセッションの方がよりリファレンスラインに近づいていくと仮説を立てたが，その傾向があるようにも見える．
総じて，リソースの消費傾向には制約の有無や試行の回数により傾向が違うように見られるが，C_5 のセッションの方がよりリファレンスラインに近づくという仮説を強く主張できる結果は得られなかった．

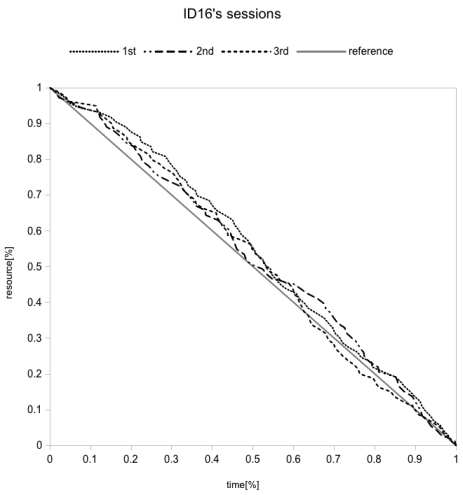
表 29 は，C_4，C_5 におけるユーザのクエリ発行，記事選択，ページネーション，適合性判断の，各行動から行動の遷移に要した時間の割合を示したものである．ここで示す値は，先ほど示した成績の良かった/悪かったユーザのセッションから算出したものである．



(a) best 1 ユーザーズセッション on C_4



(b) best 2 ユーザーズセッション on C_4

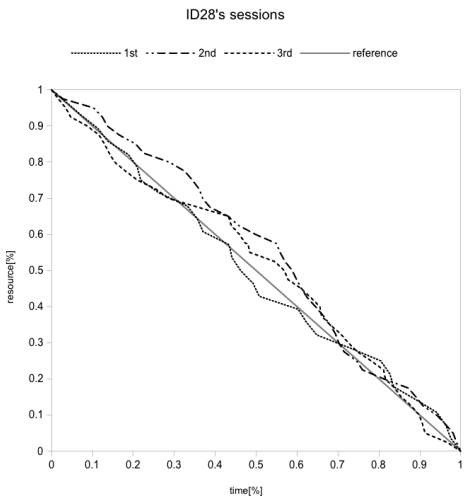


(c) best 3 ユーザーズセッション on C_4

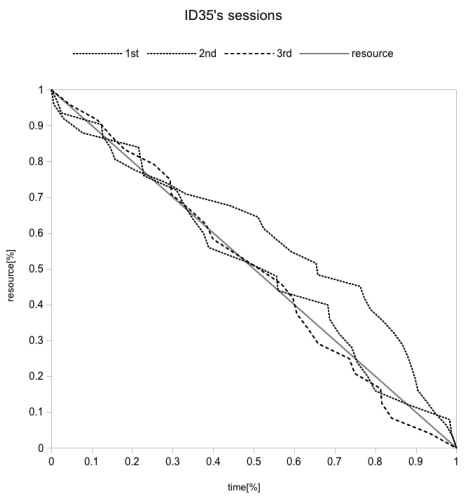
図 13　best ユーザーズセッション on C_4

3.5 リソース消費傾向の分析と考察 でも述べたように，成績の良いセッションは間隔が短くなる傾向があることがここでもわかる. 対 して，成績の悪いセッションは, 間隔が長くなる傾向があることがわかる. ここで注目したいのが，C_5 においては成績の良いセッションでも間隔が長くなる傾向があるというところである．また，C_4 では試行を重ねる毎に間隔が短くなる傾向があるのに対して，C_5 ではそういった傾向は見られない．これは，リソースの制限が設けられることにより，一つ一つの意思決定を慎重に行なっているからだと考えられる．

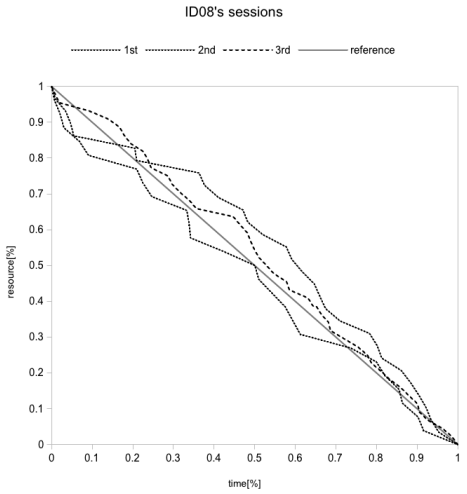
実験 1 の C_2 や C_3 と比較しても，C_5 の結果とベースラインとなる C_4(=C_1) の結果とでは大きな差が出ていることから，C_2 や C_3 よりもより強い制約が働いているとも考えられる．その結果，C_2 や C_3 では間隔を短くする作用が観察できたが，C_5 では逆に間隔を長くする作用が観察できた．これにより，制



(a) worst 1 ユーザーズセッション on C_4



(b) worst 2 ユーザーズセッション on C_4



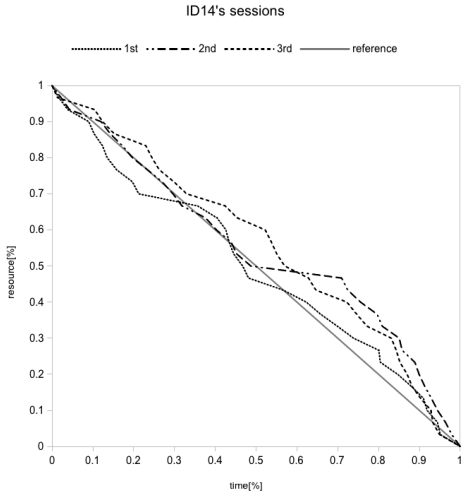
(c) worst 3 ユーザーズセッション on C_4

図 14　worst ユーザーズセッション on C_4

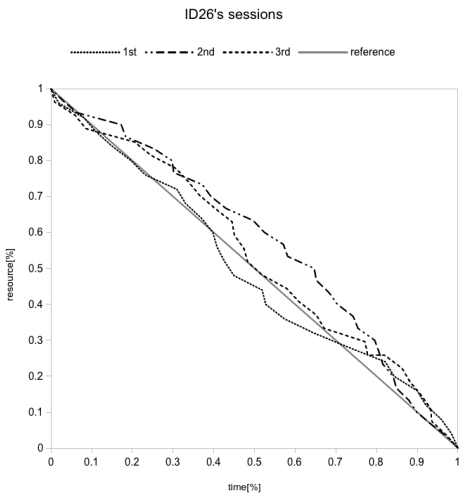
約の種類や程度によってユーザの情報検索行動のプロセスに与える影響が違うということが言えるだろう．

4.5　実験 2 のまとめ

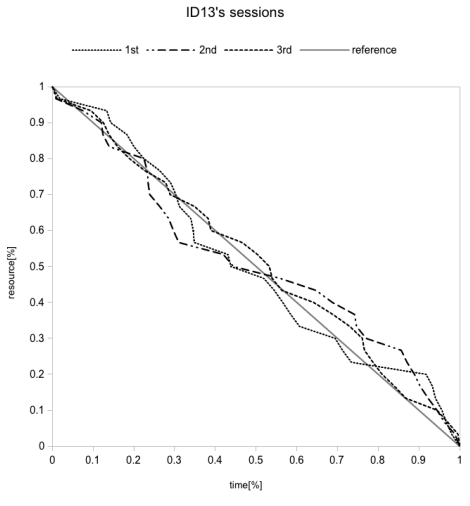
実験 2 では，リソースの制限がある場合とない場合とで，複数回繰り返されるタスクの中で情報検索行動の「戦略性」の変化・発展というところに着目し，実験により検証した．そのまとめとしては，



(a) best 1 ユーザーズセッション on C_5



(b) best 2 ユーザーズセッション on C_5

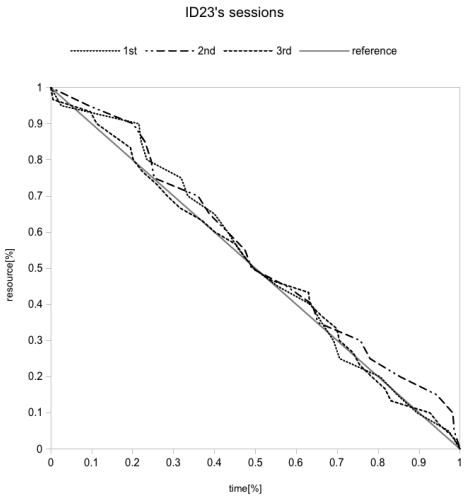


(c) best 3 ユーザーズセッション on C_5

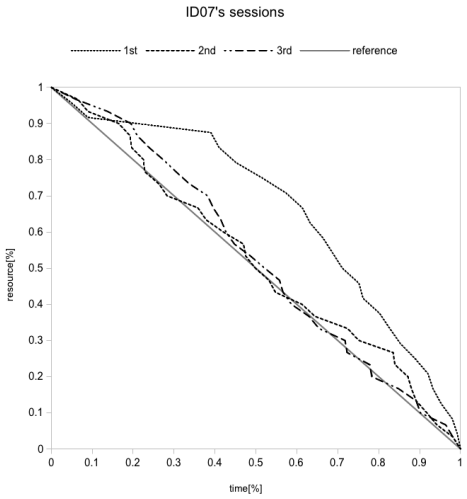
図 15　best ユーザーズセッション on C_5

- リソースの制限がある場合（=C_5），タスクをこなすうちに自身の中でのリソースの配分について学習が行われることがわかった．
- 実験 1 と同様に，成績の良いセッションは各行動の間隔が短くなる傾向があることがわかった．
- C_4 では試行を重ねる毎に行動の間隔が短くなる傾向があるのに対して，C_5 ではそういった傾向が見られないことがわかった．

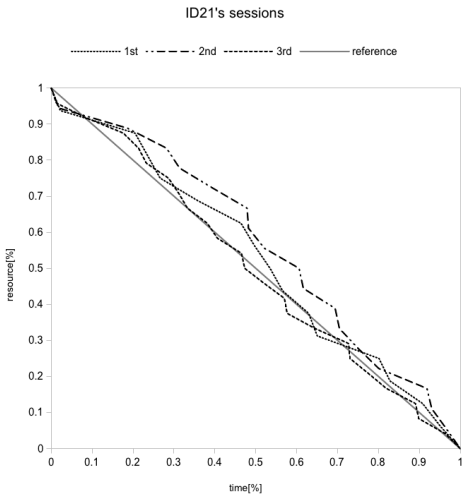
といったことが挙げられる．



(a) worst 1 ユーザーズセッション on C_5



(b) worst 2 ユーザーズセッション on C_5



(c) worst 3 ユーザーズセッション on C_5

図 16　worst ユーザーズセッション on C_5

5　全体的な議論

5.1　2 つの実験から得られた知見と考察・議論

本節では，本研究で行った 2 つの実験について比較をしながら考察を行う．

実験 1 では特定の行動に対して制限を設けることにより制約のある環境を再現した．対して，実験 2 では複数の行動に対してその全体の行動数を制限することで制約のある環境を再現した．これらは，どのような制約が，どの程度存在すれば，どの程度意識や行動，成果に影響が出るかを検証するために行ったものであった．その結果，実験 1 では 3 種類の制約の中では時間制限 + クエリ発行回数制限が特

表 29　実験 2: 行動のインターバルで見たリソースの消費傾向

interval[s]		1*st*		2*nd*		3*rd*	
		best 3	worst 3	best 3	worst 3	best 3	worst 3
< 5	total	.107	.049	.146	.088	.128	.073
	C_4	.141	.060	.177	.065	.161	.073
	C_5	.048	.034	.090	.119	.069	.072
5 < 10	total	.227	.112	.198	.088	.157	.130
	C_4	.268	.155	.238	.097	.181	.138
	C_5	.155	.051	.124	.075	.115	.120
10 < 15	total	.163	.112	.186	.188	.236	.115
	C_4	.195	.131	.207	.226	.303	.119
	C_5	.107	.085	.146	.134	.115	.108
15 <	total	.502	.727	.470	.638	.479	.682
	C_4	.396	.655	.378	.613	.355	.670
	C_5	.690	.831	.640	.672	.701	.699

に強い影響をあたえる事がわかったが，実験２では実験１で見られたような強い影響を確認することはできなかった．また，実験デザインの段階でこちらが期待していた結果とは違った結果が多く出たが，これらは，実験２の「全体の行動数を制限する」という手法や，その制限の回数の設定が原因である可能性がある．一方，実験１ではこちらの概ね期待した通りの結果が出たこともあって，仮説や実験デザインが上手くいっていたということが言えるだろう．結果として，実験１と実験２を比較したとき，実験１の手法のほうがより人間のタスクパフォーマンスに与える影響が強いということになった．

実験２ではリソースの消費傾向を観察することにより「戦略性」といった観点で考察したが，「戦略性」を確かに定義するだけの知見は得られなかった．これは，分析手法や実験デザインに問題があったと考えられる．

5.2　先行研究との比較

Kuhlthau は一般的な（適合性判断のタスクの）情報検索行動における人間の行動モデルを提唱した [14] が，本研究では，過去の一連の研究で行われてきた一般的な行動のモデルを定式化するために人間の行動プロセスを観測するのではなく，外的要因で行為者の意識へ働きかけることによりその行動プロセス自体をコントロールすることを目的としていた．その結果，確かに行為者の行動プロセスをある程度コントロールすることができた．

Ariely と Wertenbroch の「先延ばし問題」[6] では，課題の期限を，課題をこなす本人が設定する場合と，他者から設定される場合とで，成果に差が出ることが述べられていた．結果としては，後者の方がよい成果が出ることがわかった．本研究では，タスクをこなす際に利用できるリソースを制限することにより，この「先延ばし問題」のような「制約がある環境」を作り出し，被験者の行動や成果にどのよ

うな影響が出るかを実験により検証した．実験の結果は，設定する制約の種類によって行動や成果に影響が出ることを示した．その影響は，「良い影響」だと一概には言えないものであったが，しかし確実に被験者の意識，行動，成果に影響が出ていることが検証できた．情報検索行動における研究では，利用者の振る舞いを観測することにより行動理論の確立やモデル化などが行われてきたが，本研究ではその部分には触れず，利用環境に通常とは違う外的要因 (=制約) を加えることにより，利用者の思考に負荷をかけ，それにより現れる意識への影響によって行動と成果をコントロールすることを目的とした．そして，成果を制約が加わっていないときと制約が加わってるときとで比較し，どのような差が現れるかを検証した．つまり，成果を向上させることができる制約を理解することにより，利用者の情報検索行動をより良いものにしようということが本研究の目的であった．

Xie は，検索行動のプロセスを行動から行動への状態遷移（流れ）の数やその割合を観察することにより分析した [21]．本研究では，行動の状態遷移の数や割合ではなく，間隔（時間）と割合により「戦略性」という考えを提案した．また，Junco らはマルチタスキングにおけるタスクパフォーマンスと注意力による関係を議論した [13] が，本研究では注意力を外的要因によりタスクに向けさせることにより，そのタスクパフォーマンスをコントロールすることを目的としていた．その結果は先に挙げたとおりである．

5.3 研究の限界

本研究では制約が人の情報検索行動に与える影響について検証したが，様々な要素において限定的なものであったため，得られた知見の一般性は主張できない．しかし，限定的ではあるものの，その限定の仕方についても考慮した．サンプルの偏りに関しては，被験者として参加したのはほとんどが大学生であったが，その中でも大学生の所属する学部や専攻する分野は多様であったし，コンピュータに関しての習熟度も特別偏りは見られないなど，そのプロファイルは多様であった．タスクの対象となったトピックに関しては，採用したトピックの数は限られてはいるが，被験者のもつ知識量による偏りをなくすために広い分野から選択し，そのタスクの難易度も揃えた上でこちらからとりくむタスクのトピックを強制せずに被験者に選ばせた．実験に使用したシステムに関しては，バックエンドは 1 つであったが，そこで採用したアルゴリズムは一般的・代表的なものであるので，特殊なものであるとは言えないため，その点での一般性は主張できる．

タスクがトピックに対する調べごと 1 つだけであること，実験参加者の大半が学生であることから，得られた知見の一般性は主張できない．タスクに関しては，タスクに取り組む時間を 15 分に制限したため，今回の結果が情報検索行動全体のどの部分を見ているかはわからない．もしかしたら情報検索行動の極々初期の部分だけかもしれないし，または，本来タスクを終えている部分までを見ているかもしれない．しかし，過去に同様のデータコレクションを用いた被験者実験でもこの 15 分という時間制限を採用した事例はあるので，その結果との比較を行うことは出来る．

今後，これらの要素について更に考え，様々なケースを試すことは必要であると思われる．

5.4 研究の適用範囲

情報検索行動の成果を向上させるための制約の種類，程度を理解することができれば，それを種々の情報システムに応用することにより，利用者がより利用しやすいシステムの設計・開発に寄与できるかもしれない．また，本研究では「戦略性」についても議論したが，この観点はこれまでの研究にない新たなアプローチであり，後々の研究のための一つの指標になりえると考えられる．加えて，本研究から得られた知見からの発展で，合理的な情報検索行動の手法を確立することができれば，情報探索技術をはじめとした情報リテラシの指導のために，教育の場で寄与できるかもしれない．

6 結論

6.1 結論

本研究では，情報検索の分野において，人間のタスクパフォーマンスを向上させることを目的として，行動経済学の「不合理性」や，「制約」の考えを基にウェブ検索システムを利用する人間の意識を制約を用いてコントロールし，行動や成果を観察・分析する制限付きウェブ検索システムを開発し，実験を行った．これはこれまでの情報検索行動の研究とは違ったアプローチをとるものであった．行動経済学の「先延ばし問題」と同様に，ウェブ検索システムに制約を設けることにより，確かにユーザの行動や成果に変化が現れることがわかった．また，実験 1 で設けた 3 種類の制約の中では，時間制限 + クエリ発行回数制限が特にユーザの行動や成果に影響を与えることがわかった．その影響はユーザの成果を向上させるものが多かった．また，情報検索プロセスを評価するための指標を考えるために，探索プロセスの良い/悪いについて「戦略性」という観点からリソースの消費傾向の分析によって考察した．その結果，タスクの成果が良いセッションと悪いセッションとでリソースの消費傾向に差があることがわかった．加えて，制約によってリソースの消費傾向が成績の良いセッションの傾向に近づくことがわかった

実験 2 では，実験 1 の検証で得られた知見を用いて，制約の有無によって複数回の情報検索行動を繰り返す内に「戦略性」がどのように変化していくかを検証したが，制約の存在が有意に働くかどうかの明確な知見は得られなかった．

今後は，今回採用した制約以外の制約についても調査することや，また別のタスクにおける振る舞いや成果についても観察・分析することを考えている．また，制約が加わるとユーザの意識，行動に影響が出て，成果が良くなる場合もあるが，その意識への影響 (思考への影響) は利用者にとって負担になるものでもあるので，その負担を如何に軽減するか，その手法についても考慮したいと考えている．

6.2 今後の方向性

今後は，様々な制約の種類や程度を試すことにより，情報検索行動を向上させるためにはどのような制約が適当なのかを定量的に検証したい．加えて，情報検索行動における「戦略性」を評価するための指標や，「戦略性」という考え方自体の定義についても確立したい．また，5.3 制約の限界 で述べた限定性については，様々な実験環境やサンプルで実験を行うことにより，一般性を主張できるように努めていきたい．

謝辞

まずは，お忙しい中，終始熱心で丁寧なご指導を頂いた指導教員の上保秀夫准教授に感謝の意を表したいと思います．上保准教授には研究テーマの決定や研究の進め方，加えて，研究とは何か，良い研究をするにはどうすればいいかといった研究自体の方法論までご教授いただきました．この 2 年で学んだことは必ず将来自身の役に立つものと思います．

主査の中山伸一教授にもお忙しい中，様々なアドバイスをいただきました．中山教授の豊富な経験からくる客観的な考察・評価は本論文を書き上げる上で大変ありがたいものでした．その他にも様々な場面でお世話になりました中山教授に，感謝の意を表したいと思います．

上保研究室の同僚たちとのディスカッションは自身にとってもとても身になるものが多く，糧になりました．感謝の意を表します．

本研究で行った実験に協力して下さった本大学の学生方々にも，こんな場所ではありますが多大な感謝をお贈りしたいと思います．

最後に，研究で辛い時，私を支えて下さった多くの友人に，最高の感謝を．

参考文献

[1] *Overview of CLIR Task at the Fifth NTCIR Workshop*, 2005.

[2] *Overview of clir task at the sixth ntcir workshop.*, 5 2007.

[3] *Effectiveness of additional representations for the search result presentation on the web*, volume 44, 1 2008.

[4] Ainslie. A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, 82:463–496, 1975.

[5] G.A. Akalof. Procrastination and obedience. *American Economic Review*, 81:1–19, 1991.

[6] D. Ariely and K. Wertenbroch. Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological Science*, 13(3):219–224, 5 2012.

[7] B. Delvin. Empirical findings based on the sense-making methodology. *Personal communication to Kal Jarvelin*, 7 2002.

[8] B. Delvin and M. Franette. Sense-making methodology: Communicationg communicatively with campaign audiences.

[9] D. Elis, D. Cox, and K. Hall. Sense-making methodology: Communicationg communicatively with campaign audiences. *Journal of Documentation*, 49:356–359, 1993.

[10] D. Elis and M. Haugan. Modeling the infromation seeking patterns of enginerrs and researchscientists in an industrial enviroment. *Journal of Documentation*, 53(4):384–403, 1997.

[11] S.G. Hart and L.E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *P. A. Hancock and N. Meshkati (Eds.) Human Mental Workload*, 1988.

[12] H. Joho, H. David, and J.M. Joemon. Comparing collaborative and independent search in a recall-oriented task. *In Proceedings of the second international symposium on Informa- tion interaction in context*, 2008.

[13] R. Junco and S.R. Cotten. Perceived academic effects of instant messaging use. *Conputers and Education*, 56(2):370–378, 2011.

[14] C.C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Infromation Science*, 42(5):361–371, 1991.

[15] Loewenstein. Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65:272–292, 1996.

[16] J. Marguc, J. Forster, and G.A. Van Kleef. Stepping back to see the big picture: when obstacles elicit global processing. *Journal of Personality and Social Psychology*, 101(5):883–901, 2011.

[17] A. Mizumoto and O. Takeuchi. Basics and considerations for reporting effect sizes in research papers. **英語教育研究**, 31:57–66, 2008.

[18] D.R. Morehead and W.B. Rouse. Models of human behavior in information seeking tasks. *Information on Processing and Management*, 18(4):193–205, 1982.

[19] P. Nils and J. Kalervo. "irrational" searchers and ir-rational researchers. *Journal of the American Society for Infromation Science*, 57(2):222–232, 2 2006.

[20] H.A. Simon. Models of bounded rationality. *Empirically Grounded Economic Reason*, page 32, 1997.

[21] Iris Xie and Soohyung Joo. Transitions in search tactics during the web-based search process. *JASIST*, 61(11):2188–2205, 2010.