特集論文 「Web インテリジェンスとインタラクション 2017」

網羅性指向タスクにおける未閲覧情報量の提示

Displaying the Amount of Missed Information in Recall-Oriented Tasks

梅本 和俊 *1
Kazutoshi Umemoto
京都大学大学院 情報学研究科
Graduate School of Informatics, Kyoto University
umemoto@dl.kuis.kyoto-u.ac.jp

山本 岳洋
Takehiro Yamamoto
（同　上）
tyamamot@dl.kuis.kyoto-u.ac.jp

田中 克己
Katsumi Tanaka
（同　上）
tanaka@dl.kuis.kyoto-u.ac.jp

keywords: query suggestion interface, search stopping, information retrieval

Summary

In recall-oriented tasks, where collecting extensive information from different aspects of a topic is required, searchers often have difficulty formulating queries to explore diverse aspects and deciding when to stop searching. With the goal of helping searchers discover unexplored aspects and find the appropriate timing for search stopping in recall-oriented tasks, this paper proposes a query suggestion interface displaying the amount of *missed information* (*i.e.*, information that a user potentially misses collecting from search results) for individual queries. We define the amount of missed information for a query as the additional gain that can be obtained from unclicked search results of the query, where gain is formalized as a set-wise metric based on aspect importance, aspect novelty, and per-aspect document relevance and is estimated by using a state-of-the-art algorithm for subtopic mining and search result diversification. Results of a user study involving 24 participants showed that the proposed interface had the following advantages when the gain estimation algorithm worked reasonably: (1) users of our interface stopped examining search results after collecting a greater amount of relevant information; (2) they issued queries whose search results contained more missed information; (3) they obtained higher gain, particularly at the late stage of their sessions; and (4) they obtained higher gain per unit time. These results suggest that the simple query visualization helps make the search process of recall-oriented tasks more efficient, unless inaccurate estimates of missed information are displayed to searchers.

1. は じ め に

検索タスクの中には，1 回の検索クエリの投入や 1 つの文書の閲覧だけでは達成の困難なものが存在する．こうしたタスクは探索型検索 [White 09] と呼ばれる．探索型検索で取り組まれる問題は多角的かつ制限のないものであるため，複数の文書の収集および比較が必要になるという性質が存在する [Wildemuth 12]．医療や健康は，探索型検索が行われやすい代表的なドメインと言われている [Cartright 11]．例として，喫煙を継続すべきかどうかを判断するために,「タバコの影響」というトピックについて検索を行うユーザを考える．このトピックには「肺癌」や「値上げ」,「リラックス」などさまざまな観点が含まれており，各観点について多様な情報が存在している．ユーザがこのトピックを十分に理解した上で適切な意思決定を行うためには，一連の検索行動を通して重要な観点に関する適合文書を網羅的に収集する必要がある．本稿では，こうした複数の観点に関する情報の網羅的な収集が必要な検索タスクを**網羅性指向タスク**と呼ぶ．

網羅性指向タスクに共通する以下の問題は，ユーザが同タスクを効率的に達成することを困難なものにしている．1 つ目の問題は，多様な観点をカバーする適合文書の収集が可能な検索クエリを思いつくことが難しいということである．未探索の観点に関する適合文書が検索結果中にほとんど含まれない場合，ユーザはタスクを達成するまでに多くのクエリを投入する必要がある．別の問題として，検索をいつやめるべきかを判断することが難しいという点があげられる．必要以上に検索を続けることは時間の無駄である一方で，検索の早期終了は重要な情報の見逃しにつながりかねない．こうした問題が生じる根底には，(1) 検索対象のトピックにどういった観点が存在し，それらがどの程度重要であるかが分かりにくい，および (2) 各観点に関する重要な情報がどの程度存在しているかや，そのうち現時点までの検索で調べ切れていないものがどの程度残っているのかが分かりにくい

*1 現在は東京大学および情報通信研究機構に所属

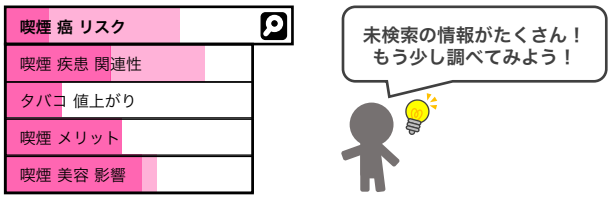


図 1 各クエリに対する未閲覧情報量の可視化

という原因が存在すると我々は考える．

本稿では，網羅性指向タスクを対象として，検索結果中の**未閲覧情報量**を提示するクエリ推薦インタフェースを提案する．我々は未閲覧情報量を検索トピックに関する重要な情報が検索結果の中に残っている度合いと定義する．提案インタフェースは，図 1 に示すように，現在の検索クエリおよび推薦クエリのそれぞれに対して未閲覧情報量を推定し，その値を棒グラフの形式でユーザに提示する．こうした情報を提示することで，ユーザは適合文書が十分に収集された観点や収集が不十分な観点を視覚的に理解し，適切なタイミングで検索を終了することが可能になると期待される．本研究では，ユーザが文書集合から獲得する**利得**を（1）観点の重要性，(2）観点の新規性，および（3）観点に関する文書の適合性に基づく尺度として定式化し，各クエリに対する未閲覧情報量をその検索結果中の未閲覧文書集合からユーザが獲得可能な追加利得として算出する．

提案インタフェースがユーザの検索戦略および検索成果に与える影響を明らかにするために，網羅性指向タスクに関する 4 種類の検索トピックを用意し，24 名の被験者を対象にユーザ実験を実施した．被験者の検索ログを用いた分析の結果，提案インタフェースの利用者は，未閲覧情報量の推定精度が高いトピックにおいて，(1）重要な情報を十分に調べた後で個々の検索を終了する，(2）重要な情報を多く含むクエリを次の検索に利用する，(3）タスクの後半にも重要な情報を継続的に収集する，および（4）単位時間あたりに獲得可能な利得が上昇することが確認された．

2. 関 連 研 究

本章では，(1）検索インタフェース，(2）ユーザの検索終了行動の理解，および（3）検索結果多様化のためのサブトピックマイニングに関する関連研究を概説する．

2·1 検索インタフェース

検索結果ページ上にリンク先の文書の手がかりとなる情報（*information scent* [Pirolli 07]）を付与することで，検索プロセスの効率化を図る研究や，ユーザの検索行動に与える影響を分析する研究がこれまでに数多く行われてきた [Kato 12, Kelly 10]．Hearst [Hearst 95] は，短時間での適合性判断支援を目的として，TileBar と呼ばれるインタフェースを提案している．TileBar は，検索結果の各文書に対して，クエリが出現する位置を矩形で可視化する．Iwata ら [Iwata 12] は，TileBar の設計思想を参考にして，多様な観点が存在する検索タスクに特化した AspecTiles と呼ばれるインタフェースを提案している．AspecTiles は，検索結果の各文書がどの観点に関する内容を含んでいるかをタイル状に表示する．Qvarfordt ら [Qvarfordt 13] は，ユーザが入力した検索クエリに対して，その検索結果に含まれる（1）既閲覧文書数，(2）過去に検索された未閲覧文書数，および（3）初めて検索された未閲覧文書数を積み上げ棒グラフの形式で可視化するインタフェースを提案している．

本研究の提案インタフェースは，Qvarfordt ら [Qvarfordt 13] と同様，クエリレベルでの手がかり情報を提示する検索インタフェースと位置付けられる．網羅性指向タスクでは，閲覧されていない文書の数よりも調べ切れていない重要な情報の量を把握することが本質的に重要であると我々は考える．そこで本研究では，文書が既に閲覧されたかどうかに加えて，検索トピックに存在する観点の重要性や新規性，ならびに文書の適合性といった情報も考慮した上で，未閲覧情報量の定式化を行う．

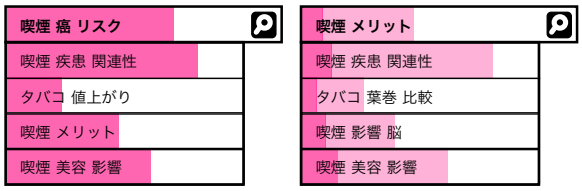
2·2 検 索 終 了

ユーザが個々の検索や検索タスク全体をいつ終了するかを理解するために，過去にさまざまな研究がなされてきた．Prabha ら [Prabha 07] は，ユーザが情報を十分に検索したと感じる基準をアンケート調査しており，定量的な基準と定性的な基準に整理している．Wu ら [Wu 14] は，ユーザ実験を完了した被験者に対して，検索終了判断に関するインタビューを行っている．Dostert ら [Dostert 09] は，ユーザ実験を通して，被験者が認識する網羅度と実際の値とのずれを指摘している．この実験に参加した被験者の多くは，タスク終了時に半数以上の適合文書を発見したと信じていたが，実際に獲得された再現率の平均値は 10%以下であった．

Dostert ら [Dostert 09] の実験結果は，情報の網羅的な収集の難しさを実証していると言える．その支援を目的として，我々は検索結果に含まれる未閲覧情報量を提示するクエリ推薦インタフェースを提案する．さらに，ユーザ実験を通して，提案インタフェースがユーザの検索終了に与える影響を定量的に分析する．

2·3 サブトピックマイニング

多くのユーザの情報要求を同時に満たすことを目的として，検索結果の多様化に関する研究が近年盛んに行われている [Dou 11, Tsukuda 13]．検索結果多様化はクエリに関連するサブトピックのモデル化に基づき行われるが，そのアプローチは明示的なものと暗黙的なものに大別される [He 12]．Carbonell ら [Carbonell 98] は，暗黙的なモデ



(a) タスク開始直後　(b) タスク終了直前

図 2 網羅性指向タスクの各状態における提案インタフェースのふるまい．最初の検索の終了後の状態は図 1 に示されている

ル化に基づく手法として，Maximal Marginal Relevance（MMR）を提案している．MMR は，文書の適合度と文書間の類似度を考慮することで新規性の高い適合文書を上位にランキングする．Agrawal ら [Agrawal 09] は，検索結果多様化を重要なサブトピックをカバーする最適な文書集合の発見と定義し，その解決策として IA-Select と呼ばれる近似アルゴリズムを提案している．彼らの多様化アルゴリズムは明示的なアプローチに分類される．

本研究が対象とする網羅性指向タスクには，検索トピックに関する観点が複数存在するという点で，多様化手法の問題領域との間に類似性が存在する．そこで我々は，多様化の評価尺度を参考に利得を定義する．また，サブトピックマイニング手法を利用して未閲覧情報量の推定を行う．

3. 提案インタフェース

本章ではまず，提案インタフェースのふるまいを説明する．次に，提案インタフェースがユーザの検索戦略および検索成果に与える影響について本研究で取り扱う研究課題を述べる．

3·1 ふ る ま い

提案インタフェースは，ユーザが入力した検索クエリおよび各推薦クエリに対して，未閲覧情報量を棒グラフの形式で可視化し，ユーザに提示する．本研究では，クエリに対する未閲覧情報量を，その検索結果中に存在する重要な情報の中で，ユーザがこれまでに閲覧していないものの量と定義する．

1 章で述べた検索トピックを再び例にとる．図 2(a) は，ユーザが “喫煙 癌 リスク” というクエリを用いて「タバコの影響」に関する網羅性指向タスクを開始した直後の提案インタフェースの状態を示している．タスク開始直後は既閲覧文書が 1 件もないため，この時点での未閲覧情報量は検索結果中に存在するトピックに関して重要な情報の総量とみなすことができる．例えば，ユーザは図 2(a) から，“喫煙 癌 リスク” というクエリは “タバコ 値上がり” というクエリに比べて検索結果中に多くの重要な情報が含まれていると推測できる．

図 1 は，初期クエリ “喫煙 癌 リスク” の検索結果からユーザが多くの適合情報を獲得した後の提案インタフェースの状態を示している．この時点での同クエリの未閲覧情報量（図中の濃桃色）は，タスク開始時の初期値（図中の薄桃色）から大幅に減少している．ユーザはこの差を見ることで，現在のクエリの検索結果中に未探索の観点に関する重要情報がほとんど残っていないことを把握できる．図 1 を詳しく見ると，現在の検索クエリに加えて，“喫煙 疾患 関連性” という推薦クエリについても未閲覧情報量が大きく減少していることが分かる．これは，両クエリの検索結果中に同一の観点に関する情報が重複して含まれているためである．一方で，“喫煙 美容 影響” など他の推薦クエリについては未閲覧情報量がほとんど変化していない．ユーザはこれを手がかりに，現時点までの検索で調べ切れていない重要情報を検索結果に多く含むクエリを知ることができる．

ユーザが本トピックに関する情報を多角的に検索し終えると，多くのクエリに対して未閲覧情報量の値が大幅に減少する．図 2(b) は，タスク終了直前における提案インタフェースの状態を示している．図中の各クエリに対する未閲覧情報量を眺めることで，ユーザは未収集の重要情報がどのクエリの検索結果にもほとんど残っていないと判断できる．

3·2 研 究 課 題

本研究では，上述の提案インタフェースがユーザの検索戦略および検索成果に与える影響を調査する．具体的には，本節で述べる 5 つの研究課題に答えることに焦点を絞る．

網羅性指向タスクでは，ユーザは複数個のクエリを投入することが予想される．最初の 2 つの研究課題は，あるクエリが投入されてから別のクエリが投入される（あるいはタスクが終了する）までの個々の検索に関するものである．

RQ1 未閲覧情報量の提示は**現在の検索の終了**に関するユーザの判断にどのような影響を与えるか

RQ2 未閲覧情報量の提示は**次の検索クエリの選択**に関するユーザの判断にどのような影響を与えるか

我々は，提案インタフェースによる未閲覧情報量の提示が個々の検索におけるユーザの検索戦略をより合理的なものにすると予想する．例えば，検索クエリの未閲覧情報量を確認することで，ユーザは重要な情報を十分に収集し終えてから現在の検索を終了することが可能になるかもしれない．また，推薦クエリ間の未閲覧情報量の比較を通して，次の検索のためのクエリが戦略的に決定されるようになる可能性も考えられる．

残りの 3 つは，検索タスク全体に関する研究課題である．

RQ3 未閲覧情報量の提示はユーザが獲得する**利得の時**

間的変化にどのような影響を与えるか

RQ4 未閲覧情報量の提示は**検索タスクの終了**に関するユーザの判断にどのような影響を与えるか

RQ5 未閲覧情報量の提示はユーザが費やす**コスト**と獲得する**利得**との**関係性**にどのような影響を与えるか

我々は，提案インタフェースがタスク全体におけるユーザの検索戦略や検索成果にも影響を与えると予想する．例えば，提案インタフェースは高い成果の継続的な獲得を可能にするかもしれない．また，個々の検索の終了に関する研究課題（**RQ1**）の類推として，未閲覧情報量の提示によって，合理的なタスクの終了判断が可能になるかもしれない．さらに，未閲覧情報量を指針とした検索を行うことで，ユーザは短い時間で高い成果を獲得できるという仮説も考えられる．

4. 未閲覧情報量の推定

本章ではまず，網羅性指向タスクにおける検索成果の評価尺度として**利得**という概念を導入する．次に，利得を用いて，本研究の根幹をなす**未閲覧情報量**を定式化する．最後に，両者の具体的な値の計算に必要となる各要素の推定手法を述べる．

4·1 利　　得

トピック t に関する網羅性指向タスクにおいて，文書集合 $D=\{d_1,d_2,\dots\}$ の閲覧を通して獲得される利得 $\text{Gain-IA}_t(D)$ について考える．1 章で述べたように，本タスクでは検索トピックに関する重要かつ多様な観点について適合文書を網羅的に収集することが望まれる．そこで我々は，利得が満たすべき要件として以下の 3 つを考える．

重要性 重要な観点に関する文書は，瑣末な観点に関する文書に比べて，高い利得をもたらす

適合性 ある観点に関する高適合文書は，その観点に関する低適合文書に比べて，高い利得をもたらす

新規性 未探索の観点に関する文書は，探索済みの観点に関する文書に比べて，高い利得をもたらす

第 1 の要件に基づき，トピック t に関する利得 $\text{Gain-IA}_t(D)$ を，t の各観点 a に関する利得 $\text{Gain}_a(D)$ へと分解する．

$$\text{Gain-IA}_t(D)=\sum_{a\in A_t}\Pr(a\,|\,t)\cdot\text{Gain}_a(D).$$

ここで，A_t はトピック t の観点の集合，$\Pr(a\,|\,t)\in[0,1]$ は t における観点 a の重要度を表し，$\sum_{a\in A_t}\Pr(a\,|\,t)=1$ を満たす．上式の分解は，意図意識型（intent-aware）の評価尺度 [Agrawal 09] と同様の考えに基づいており，重要な観点に関する利得を重視する．

次に，観点 a に関して文書集合 D から得られる利得 $\text{Gain}_a(D)$ を次式で定義する．

$$\text{Gain}_a(D)=\sum_{i=1}^{|D|}\text{Rel}_a(d_i)\cdot\text{Disc}_a(\{d_1,\dots,d_{i-1}\}).$$

ここで，$\text{Rel}_a(d)\in[0,1]$ は文書 d の観点 a に対する適合度を表す．観点ごとの利得の計算時には，第 2 の要件に基づき，各文書の適合度が足し合わされる．ただし，$\text{Gain}_a(\emptyset)=0$ とする．

上式の $\text{Disc}_a(\cdot)$ は，第 3 の要件を満たすために設計されたディスカウント項であり，閲覧済みの文書集合によって観点 a が十分にカバーされていた場合に，以降に閲覧される文書が $\text{Gain}_a(\cdot)$ に寄与する度合いを下げる役割を果たす．我々は，既閲覧文書集合 D' に対するディスカウント項を以下で定義する．

$$\text{Disc}_a(D')=\prod_{d'_j\in D'}\left(1-\text{Rel}_a(d'_j)\right).$$

上式は，高適合文書が既に多く閲覧されている観点については低い値を返す．ただし，$\text{Disc}_a(\emptyset)=1$ とする．

既存の評価尺度との関連性

前節で我々が定義した Gain-IA は，検索結果多様化の評価尺度である Intent-Aware Expected Reciprocal Rank（ERR-IA）[Chapelle 09] の定義式と類似したものになっている．ERR-IA は，クエリ q の検索結果文書リスト $R_q=\{d_1,d_2,\dots\}$ に対する評価値を次式で計算する．

$$\sum_{a\in A_q}\Pr(a\,|\,q)\sum_{i=1}^{|R_q|}\frac{\text{Rel}_a(d_i)}{i}\prod_{j=1}^{i-1}(1-\text{Rel}_a(d_j)).$$

両者の違いは，検索結果文書の逆順位（i^{-1}）に相当する項が Gain-IA には存在しないという点である．ERR-IA は，検索結果を上から順番に調べるユーザモデルを仮定しており，下位に存在する文書の価値を割り引いた評価を行う．一方で本研究の目的は，個々の文書の閲覧順序にはよらない，閲覧文書集合全体に対する尺度として利得を定量化することである．そこで我々は Gain-IA を ERR-IA の集合版，すなわち文書集合に対して定まる指標として定義した．

4·2 未閲覧情報量

本研究では，検索クエリに対する未閲覧情報量を，その検索結果中の未閲覧文書集合を閲覧することでユーザが獲得可能な追加利得と定める．具体的には，ユーザ u がトピック t に関する網羅性指向タスクに取り組んでいる際のクエリ q に対する未閲覧情報量 $\text{MI}_{u,t}(q)$ を次式で定義する．

$$\text{MI}_{u,t}(q)=\text{Gain-IA}_t(D_u\cup D_q^K)-\text{Gain-IA}_t(D_u).$$

ここで，D_u はユーザ u がこれまでに閲覧した文書集合，D_q^K はクエリ q の上位 K 件の検索結果に出現する文書集合を表す．

表 1 未閲覧情報量の計算例

A_t		$\{a_1, a_2, a_3\}$
$\Pr(a \mid t)$	a_1	0.2
	a_2	0.6
	a_3	0.2
D_u		$\{d_1\}$
D_q^K	q_1	$\{d_1, d_2\}$
	q_2	$\{d_1, d_3\}$
	q_3	$\{d_1, d_4\}$

		a_1	a_2	a_3
$\mathrm{Rel}_a(d)$	d_1	0.5	0	0
	d_2	0	0.5	0
	d_3	0	0	0.5
	d_4	0.5	0	0
$\mathrm{MI}_{u,t}(q)$	q_1			0.3
	q_2			0.1
	q_3			0.05

上式が理想的な挙動を示す状況において，クエリ q に対する未閲覧情報量 $\mathrm{MI}_{u,t}(q)$ が高いことは，q の検索結果中の未閲覧文書集合 $D_q^K \setminus D_u$ の中に，ユーザ u が見逃している重要な情報を含んだ文書が多く存在していることを示している．

例として，表 1 に示す状況を考える．この時，

$$\begin{aligned}
\text{Gain-IA}_t(D_u) &= 0.2\cdot 0.5 + 0.6\cdot 0 + 0.2\cdot 0 = 0.1,\\
\text{Gain-IA}_t(D_u \cup D_{q_1}^K) &= 0.2\cdot 0.5 + 0.6\cdot 0.5 + 0.2\cdot 0 = 0.4,\\
\text{Gain-IA}_t(D_u \cup D_{q_2}^K) &= 0.2\cdot 0.5 + 0.6\cdot 0 + 0.2\cdot 0.5 = 0.2,\\
\text{Gain-IA}_t(D_u \cup D_{q_3}^K) &= 0.2\cdot (0.5+0.25) + 0.6\cdot 0 + 0.2\cdot 0\\
&= 0.15,
\end{aligned}$$

であるため，

$$\begin{aligned}
\mathrm{MI}_{u,t}(q_1) &= 0.4 - 0.1 = 0.3,\\
\mathrm{MI}_{u,t}(q_2) &= 0.2 - 0.1 = 0.1,\\
\mathrm{MI}_{u,t}(q_3) &= 0.15 - 0.1 = 0.05,
\end{aligned}$$

となり，クエリ q_1 に対する未閲覧情報量の値が最も高くなる．これは，他のクエリ q_2 や q_3 に比べて，クエリ q_1 の検索結果には未探索かつ重要な観点 a_2 に関する適合文書（d_2）が含まれるためである．

4·3 構成要素の値の推定

利得および未閲覧情報量の計算には，トピック t の観点集合 A_t，各観点 $a \in A_t$ の重要度 $\Pr(a \mid t)$，および a に対する各文書 d の適合度 $\mathrm{Rel}_a(d)$ という 3 種類の要素の値が必要となる．本節では，その推定手法について述べる．なお，検索結果として取得する文書数 K については，実験設定に関するパラメータであるため，5 章にて実際に利用した値を報告する．

構成要素の値の推定は，明示的なサブトピックのモデル化に基づく検索結果多様化アプローチ（2 章）の対象タスクと同様のものである．そこで本研究では，同分野における最新の手法の 1 つである Tsukuda ら [Tsukuda 13] の手法（以降，佃手法と表記する）を利用することで，構成要素の値を推定する．佃手法は，情報アクセス評価型会議 NTCIR-10 の INTENT-2 タスクにおいて，サブトピックマイニングに関する全 14 件の投稿結果の中で 2 位の成績を獲得したものである [Sakai 13]．以下では，推定手法の概要を示しつつ，佃手法と異なる部分についてはその違いを明記する [*2]．なお，本手法の適用には検索システムが必要となる．本研究では，提案インタフェースが利用する検索システムと同一のものを利用する．その詳細は 5 章にて述べる．

観点集合

トピック t の観点集合 A_t を推定するために，t のサブトピック集合 S_t を 3 種類のリソースから収集する．まず，t をクエリとした際に得られるクエリ推薦をサブトピックとする．次に，クエリログから，先頭に t を含むクエリをサブトピックとして抽出する．最後に，クエリ t の検索結果に対して Zeng ら [Zeng 04] の手法を適用し，その結果得られるキーフレーズについてもサブトピックとする．こうして得られた S_t に対してウォード法を適用することでクラスタ集合 $\mathcal{C}_t$ を取得し，各サブトピッククラスタ $C \in \mathcal{C}_t$ がトピック t に存在する観点の 1 つに対応しているとみなす．5 章で述べるように，我々がユーザ実験に使用した検索トピックは過去の NTCIR で利用されたものの一部である．そこで，それらの観点集合を推定する際には，NTCIR のタスク参加者に提供されている上記リソースを利用した．

観点の重要度

トピック t における観点 a の重要度 $\Pr(a \mid t)$ を推定するために，まずは t におけるサブトピック s の重要度 $\mathrm{Imp}_t(s)$ を $\mathrm{Imp}_t(s) = \sum_{d \in D_s^N \cap D_t^N} \frac{1}{\mathrm{Rank}_t(d)}$ により推定する．ここで，$\mathrm{Rank}_t(d)$ は t をクエリとする検索結果中での文書 d の順位を表す．次に，各サブトピッククラスタ $C \in \mathcal{C}_t$ の中から，代表的なサブトピック a を $a = \arg\max_{s \in C} \mathrm{Imp}_t(s)$ により決定し，トピック t の観点の 1 つとする．こうして得られた各観点 $a \in A_t$ に対して，その重要度 $\Pr(a \mid t)$ を次式で推定する．

$$\Pr(a \mid t) = \frac{\mathrm{Imp}_t(a)}{\sum_{a' \in A_t} \mathrm{Imp}_t(a')}.$$

文書の適合度

佃手法では，各観点 $a \in A_t$ をクエリとした際に得られる上位 N 件の文書集合 D_a^N のみを対象に適合度推定を行う．これは，元論文 [Tsukuda 13] の目的が検索結果多様化であり，最終的に出力すべき文書数が 10 件程度であるという理由によるものと考えられる．一方，本研究が対象とする網羅性指向タスクでは，ユーザは検索中に複数のクエリを投入し，多くの文書を閲覧することが予想される．そのため，未閲覧情報量の値の正確な推定のためには，可能な限り多くの関連文書に対して適合度推定を行う必要がある．そこで我々は，観点 a に関する適合度の推定対象となる文書集合を $\bigcup_{s \in C_a} D_s^N$ へと拡張する．ここで，$C_a \in \mathcal{C}_t$ は観点 a に対応するサブトピッククラスタである．対象集合中の各文書 d に対して，観点

[*2] 本研究では，元論文 [Tsukuda 13] と同じパラメータ値を用いる．

図 3 提案インタフェースのスクリーンショット

a に関する適合度 $\mathrm{Rel}_a(d)$ を次式で推定する．

$$\mathrm{Rel}_a(d) = \frac{\sum_{s \in C_a} \mathrm{Imp}_t(s) \cdot \mathrm{Rel}_s(d)}{\sum_{s \in C_a} \mathrm{Imp}_t(s)}.$$

ここで，$\mathrm{Rel}_s(d)$ はサブトピック s に対する d の適合度を表す．その値は，佃手法に従い，$\mathrm{Rel}_s(d) = \frac{1}{\sqrt{\mathrm{Rank}_s(d)}}$ と推定する．

5. 実 験 設 定

本章では，我々が実施したユーザ実験の設定について述べる．

5·1 インタフェース

実験では 2 種類のクエリ推薦インタフェースを用意した．1 つ目は，3 章で説明した未閲覧情報量を提示する提案インタフェース（**w/ scent**）である（図 3）．2 つ目は，それに対する比較を目的としたベースラインインタフェース（**w/o scent**）である．ベースラインインタフェースは，未閲覧情報量がユーザに提示されないという点を除けば，提案インタフェースと同様にふるまう．ユーザが検索ボックスにクエリを入力すると，両インタフェースは Google Suggest API[*3] を通じて関連する推薦クエリ集合を取得し，入力クエリの下部に提示する．検索クエリが投入されると，次節で述べる検索システムによってランキングされた検索結果がインタフェースの右側に提示される．検索結果がクリックされると，文書ビューアがインタフェースの前面に表示され，文書の内容がビューア上で閲覧可能となる．

5·2 検索システム

我々は，Apache Solr[*4] を用いて，両インタフェースが内部的に利用する全文検索システムを構築した．検索対象の文書コーパスには ClueWeb09-JA[*5] を，文書のランキングアルゴリズムについては BM25[*6] [Robertson 09] を採用した．

この検索システムを用いた予備実験の際に，内容の重複する文書が検索結果の上位に表示されるという事例が何度か確認された．このようなランキングは，網羅的な情報収集の妨げとなるばかりでなく，計算される利得の値の正確性にも悪影響を及ぼしかねない．我々はこの問題に対する措置として，検索クエリ q が与えられた際に，BM25 によってランキングされた文書リスト R_q に MMR [Carbonell 98] を適用することで検索結果の多様化を行った．具体的には，クエリ q の検索結果リストにおいて k 位に配置される文書 $d_k \in D_q^K$ を次式によって選択した．

$$d_k = \underset{d \in R_q \setminus S^{k-1}}{\arg\max} \left[\lambda \cdot \mathrm{Rel}_q(d) - (1-\lambda) \max_{d' \in S^{k-1}} \mathrm{Sim}(d, d') \right].$$

ここで，S^{k-1} は MMR が既に選択した $k-1$ 件の文書集合，$\mathrm{Rel}_q(\cdot)$ はクエリ q に対する文書の適合度として BM25 が計算したスコア [*7]，$\mathrm{Sim}(\cdot,\cdot)$ は各文書の検索結果タイトルおよびスニペット中に出現する語集合から構成された tf-idf ベクトル間のコサイン類似度を表す．適合度と類似度を制御するパラメータ λ については，Tsukuda ら [Tsukuda 13] および Dou ら [Dou 11] に従い，$\lambda = 0.3$ と設定した．検索結果として提示する文書数 K については，Qvarfordt ら [Qvarfordt 13] に従い，$K = 100$ と設定した．

表 2 ユーザ実験に用いた検索トピックとその観点

検索トピック	トピックの観点
糖尿病の症状	原因，末期症状，自覚症状，予防，合併症，…
うつ病	症状，仕事，社会，接し方，治療，…
結婚式の服装マナー	男性，女性，親族，髪型，二次会，…
恐竜	化石，絶滅，図鑑・種類，博物館，時代，…

5·3 検索トピック

本実験では，網羅性指向タスクに関する 4 つの検索トピックを用意した．これらは，NTCIR INTENT-1 [Song 11]，INTENT-2 [Sakai 13]，および IMine-1 [Liu 14] タスクにおいて，サブトピックマイニングに関するサブタスクで提供されたものの一部である．実験で利用したトピックを表 2 に示す．

1 章でも述べたように，医療や健康は探索型検索が行われやすいドメインである [Cartright 11]．そこで，上述の候補の中からこうしたドメインに属するトピックとして「糖尿病の症状」および「うつ病」を選択した．さら

*3 http://www.google.com/complete/search
*4 http://lucene.apache.org/solr/
*5 http://www.lemurproject.org/clueweb09.php
*6 パラメータには，Solr のデフォルト値（$k_1 = 1.2$, $b = 0.75$）を用いた．
*7 1 位の文書のスコアで除算することで，値域を $[0,1]$ に正規化した．

に，それ以外のドメインから「結婚式の服装マナー」および「恐竜」の 2 トピックを選択した．表 2 には，これらのトピックに関する観点の一部も示されている．これらの観点は，NTCIR のタスク主催者が評価用に提供しているものである．同表より，各トピックには複数の観点が存在しており（1 トピックあたりの平均観点数は 10.50），タスクの達成には複数の文書の収集が必要であることが分かる．

5·4 手　　順

本実験では，上述の 2 種類のインタフェースを被験者内要因として設計した．すなわち各被験者は，全 4 トピックのうち 2 つに対して提案インタフェースを，残りの 2 つに対してベースラインインタフェースを用いて網羅的な情報の収集に取り組んだ．また，利用するインタフェースや取り組むトピックの順序による影響を可能な限り取り除くために，グレコラテン方格を用いて両者の順序の回転および無作為化を行った．

我々は，本実験の被験者として 24 名（男性 20 名，女性 4 名）の大学関係者を募集した．その内訳は学部生 16 名，大学院生 6 名，研究者 2 名であり，被験者の平均年齢は 23.29 歳であった．各被験者に実験内容を十分に理解してもらうこと，およびインタフェースの使い方に慣れてもらうことを目的として，前節で述べたトピックに関する本番タスクに先立ち，練習タスクとして「地球温暖化」に関する検索を 15 分程度行ってもらった．練習タスクの中ではインタフェースに関する説明も行った．その際に，未閲覧情報量については論文中での定義（トピックに関する重要な情報が検索結果の未閲覧文書集合中に残っている度合い）をそのまま説明した．また，被験者が未閲覧情報量を盲信してクエリ投入や検索終了に関する判断を行うことを避けるために，提案インタフェースが可視化する未閲覧情報量は推定値であり正確とは限らないことを伝えた．実験の開始前後および各タスクの実行前後にはアンケートへの回答を求めた．

提案インタフェースを適切に評価するためには，被験者に可能な限り現実的な状況の下で網羅性指向タスクに取り組んでもらう必要がある．そこで我々は，実験の開始前に「トピックに関する詳細な記事を提出するという課題が出たので，そのために検索システムを使ってトピックに関する重要な情報をさまざまな角度から調べて下さい」と被験者に伝えた．また，タスクの遂行中に記事の執筆に有用な文書を発見した際には，その文書をブックマークしてもらうよう依頼した．我々は，重要な観点に関する情報を網羅的に収集できたと感じた時点でタスクを終了するように被験者に伝え，その直感的な基準として，「現時点までに詳細な記事の執筆に必要な情報を十分に収集し，これ以上検索を続けてもより良質な記事の執筆に有用な文書を発見できないと感じた時点でタスクを終了して下さい」と説明した．

実験時の検索プロセスをより現実的なものにするために，被験者に自由なクエリの入力および任意の検索結果文書の閲覧を許可した．また，タスクの達成に有用な文書を発見した際には，その文書をブックマークすることを依頼した．ただし，閲覧中の文書からリンクを辿るという操作については許可しなかった．これは，リンク先の文書が本実験の文書コーパスである ClueWeb09-JA の中に含まれているとは限らないためである．

被験者に全トピック数を知らせると，タスクの終了時間に影響を与えてしまうおそれがある．そこで我々は，十分な数のトピックが用意されており，それらが全て完了された時点あるいは開始から 2 時間が経過した時点で実験が終了すると被験者に伝えた．各タスクは予備実験によって 20 分以内に完了されることが確認されており，実際に本実験においても，全ての被験者が一連の実験プロセスを 2 時間以内に完了した．

6. 実　験　結　果

本章では，我々が実施したユーザ実験の結果を報告し，3·2 節で列挙した研究課題に回答する．なお，実験データは分布の正規性が保証されないため，統計的有意性を検証する際にはノンパラメトリックな検定法を利用した（有意水準 $\alpha = 0.05$）．

6·1 正 解 デ ー タ

被験者が獲得した成果を評価するために，実験で用いた各トピックについて正解データを用意した．4·3 節で述べたとおり，利得や未閲覧情報量の計算には，トピックの観点集合，各観点の重要度，および各観点に対する文書の適合度という 3 種類の構成要素の値が必要である．このうち，観点集合およびそれらの重要度については，NTCIR のタスク主催者によって提供されているものを正解データとした．

適合性判定

残りの構成要素である適合度の正解データを作成するために，被験者が実験中に閲覧した各文書に対して，本稿著者らが適合性判定を排他的に行った．具体的には，判定者 3 名のうち 1 名（第 1 著者）が 2 つの実験トピック（うつ病および結婚式の服装マナー）を，残りの 2 名がそれ以外の 2 つの実験トピック（糖尿病の症状および恐竜）を 1 つずつ担当した．正解データの各観点に対する文書の適合性グレードには，irrelevant（$= 0$），partially-relevant（$= 1$），highly-relevant（$= 2$），および perfect（$= 3$）の 4 段階を利用した．判定者間での適合性のずれを可能な限り避けるために，「文書が観点に関する情報を 90%以上網羅していると思うなら perfect，60%以上なら highly-relevant，30%以上なら partially-relevant，それ未満あるいは観点と関係ない場合は irrelevant」という適合性判定基準を各判定者に事前に伝えた．こうして得られた観点 a に対する文書

表 3 推定利得と真の利得の相関係数．太字は各指標の上位 3 件を表す

検索トピック	Pearson's r	Spearman's ρ	Kendall's τ
糖尿病の症状	**0.834**	**0.851**	**0.683**
うつ病	**0.845**	**0.860**	**0.678**
結婚式の服装マナー	**0.824**	**0.862**	**0.713**
恐竜	0.710	0.702	0.529

d の適合性グレード $g_{ad} \in \{0,1,2,3\}$ に対して，Chapelle ら [Chapelle 09] に従い，変換式 $\mathrm{Rel}_a^*(d) = (2^{g_{ad}} - 1)/2^3$ を適用することで適合度の正解値を用意した．この変換により，適合度の正解値は $\{0, 0.125, 0.375, 0.875\}$ のいずれかとなる．

利得の推定精度

提案インタフェースがユーザに提示する未閲覧情報量は，4 章で述べた手法で推定された利得の値を用いて計算される．利得の推定精度の低さは，提示される未閲覧情報量の信頼性の低下につながり，ユーザの検索に悪影響を与えるおそれがある．そこで，実験で用いた各トピックについて，利得の推定精度の検証を行った．以降では，4·3 節で述べた手法により推定された利得を推定利得（estimated gain），本節で述べた正解データによる計算された利得を真の利得（oracle gain）と呼ぶ．

我々は，推定精度の検証のために，被験者が文書閲覧を行った各々の時点での推定利得と真の利得を計算した．こうして得られた利得値のペアの系列に対して，ピアソンの積率相関係数（Peason's r），スピアマンの順位相関係数（Spearman's ρ），およびケンドールの順位相関係数（Kendall's τ）を計算した．ここで，相関係数が高い場合は，両者の利得の変動に一貫性があることを意味するため，推定精度が高いとみなすことができる．トピックごとに上記の計算を行った結果を表 3 に示す．

同表より,「恐竜」というトピックは，他のトピックと比べて，各指標の相関係数が低いことが分かる．同トピックでは,「模型 販売」や「グッズ ショップ」などの正解データには含まれていない観点が 4·3 節で述べた手法によって重要と推定されていたことが判明した．この観点のずれが，推定利得が真の利得と一貫しない変動をもたらしたと言える．

前述のとおり，利得の推定精度の低さは，望ましくない結果をもたらす可能性がある．そこで以後の分析では，全トピックに関する結果に加えて，推定利得が真の利得と高く相関する 3 トピック（**高相関トピック**）に関する結果，および低く相関する 1 トピック（**低相関トピック**）に関する結果についても示し，利得の推定精度が与える影響を議論する．

6·2 現在の検索の終了

研究課題 **RQ1** に答えるために，未閲覧情報量の提示が現在の検索の終了に与える影響を分析した．ユーザが検索結果中の重要な情報を十分に収集した場合，クエリの未閲覧情報量は検索開始時点から比べて大きく減少するはずである．そこで，被験者が投入した各クエリに対して，検索の開始時点と終了時点での未閲覧情報量の変化を計算した．その結果を表 4 に示す．

表 4 検索開始時と終了時での未閲覧情報量の変化の平均（および標準偏差）．太字は有意差（$p < 0.05$）が確認された結果を表す

	全トピック		高相関トピック		低相関トピック	
	w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
ΔMI	0.109 (0.110)	**0.138** (0.134)	0.128 (0.120)	**0.186** (0.147)	0.069 (0.073)	0.071 (0.072)

表 5 クエリ投入時の未閲覧情報量の平均（および標準偏差）．太字は有意差（$p < 0.05$）が確認された結果を表す

	全トピック		高相関トピック		低相関トピック	
	w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
MI	0.211 (0.194)	**0.241** (0.191)	0.238 (0.214)	**0.298** (0.210)	0.155 (0.126)	0.162 (0.123)

全トピックに関する結果から，提案インタフェースはベースラインインタフェースに比べて，検索の前後での未閲覧情報量の変化が大きいことが分かる．両者の差は，ウィルコクソンの順位和検定によって有意であることが確認された（$p = 0.001$）．高相関トピックに関する結果についても両者の間に有意な差が確認された一方で（$p < 0.001$），低相関トピックでは未閲覧情報量の変化に関してインタフェース間に有意な差は存在しなかった（$p = 0.683$）．

以上の結果から，推定精度の高いトピックでは，提案インタフェースの利用者は重要な情報をより多く収集した後で現在の検索を終えていたことが分かる．一方，未閲覧情報量の推定が不正確なトピックでは，提案インタフェースは個々の検索で収集された重要な情報の量に大きな影響を与えなかったと言える．

6·3 次の検索クエリの選択

研究課題 **RQ2** に答えるために，未閲覧情報量の提示が次の検索クエリの選択に与える影響を分析した．我々は，未閲覧情報量の提示によって重要な情報を多く含むクエリが投入されるようになるという仮説を立てた．この仮説を検証するために，被験者が投入した各クエリに対して，投入時点でのそのクエリの未閲覧情報量を計算した．その結果を表 5 に示す．

全トピックに関する結果は，提案インタフェースではベースラインインタフェースに比べて，未閲覧情報量の多いクエリが投入されたことを示している．両者の差は，ウィルコクソンの順位和検定によって有意であることが確認された（$p = 0.002$）．前節の結果と同様，高相関トピックではその差が大きくなった一方で（$p < 0.001$），低相関トピックではインタフェース間に有意な差は存在しなかった（$p = 0.521$）．

表 6 上段：タスク開始時と終了時での未閲覧情報量の変化の平均（および標準偏差）．下段：獲得された利得の平均（および標準偏差）．太字は有意差（$p < 0.05$）が確認された結果を表す

	全トピック		高相関トピック		低相関トピック	
	w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
ΔMI	0.162 (0.132)	**0.195** (0.151)	0.188 (0.143)	**0.256** (0.158)	0.106 (0.080)	0.110 (0.086)
Gain	0.358 (0.132)	0.404 (0.147)	0.370 (0.133)	0.415 (0.150)	0.322 (0.129)	0.371 (0.138)

この結果から，提案インタフェースの利用者は，推定精度の高いトピックに関する検索時に，重要な情報をより多く含んだクエリを優先的に投入していたことが分かる．一方で，未閲覧情報量の推定が不正確なトピックについては，そのような傾向は見られず，どちらのインタフェースにおいても同程度の重要情報を含むクエリが投入されていたと言える．

6·4 利得の時間的変化

研究課題 **RQ3** に答えるために，未閲覧情報量の提示が被験者によって獲得される利得の時間的変化に与える影響を分析した．具体的には，タスク開始から 1 分が経過するごとに，その時点（elapsed time）までに各被験者が獲得した利得を計算し，その平均値をプロットした [*8]．その結果を図 4 に示す．

全トピックに対する結果（図 4(a)）からは，タスク開始から約 7 分が経過するまでは，インタフェース間で獲得される利得に大きな差は見られない．しかし，それ以降は，提案インタフェースの利用者がより高い利得を獲得していたことが分かる．高相関トピックに対する結果では，その傾向がより強くなっている（図 4(b)）．その一方で図 4(c) は，未閲覧情報量の推定精度が低い場合は，ほぼ全ての時点でベースラインインタフェースの利用者の方が高い利得を獲得していたことを示している．

6·5 検索タスクの終了

研究課題 **RQ4** に答えるために，未閲覧情報量の提示が検索タスクの終了に与える影響を分析した．6·2 節と同様の類推として，重要な観点に関する情報が十分に収集された場合，トピックに関連する多くのクエリに対して，未閲覧情報量の値が大きく減少するはずである．そこで我々は，被験者がタスク中に投入した各クエリに対して，タスクの開始時点と終了時点での未閲覧情報量の変化を計算した．さらに，タスクを通して被験者が獲得した利得についても計算した．その結果を表 6 に示す．

ウィルコクソンの順位和検定を行ったところ，全トピックに関する結果では，未閲覧情報量の変化にインタフェース間で有意な差が確認された（$p = 0.002$）．高相関トピックに関する結果についてもインタフェース間の差が有意であったのに対して（$p < 0.001$），低相関トピックでは両者の間に有意な差は認められなかった（$p = 0.845$）．タスク全体で獲得された利得についても，提案インタフェースの方がベースラインインタフェースよりも平均値が高いという結果が得られた．しかし，全トピック（$p = 0.097$），高相関トピック（$p = 0.191$），低相関トピック（$p = 0.410$）のいずれに対しても，両者の差の有意性は確認されなかった．

[*8] 各時点で検索を継続している被験者を対象として利得の平均値を計算した．

6·6 コストと利得の関係性

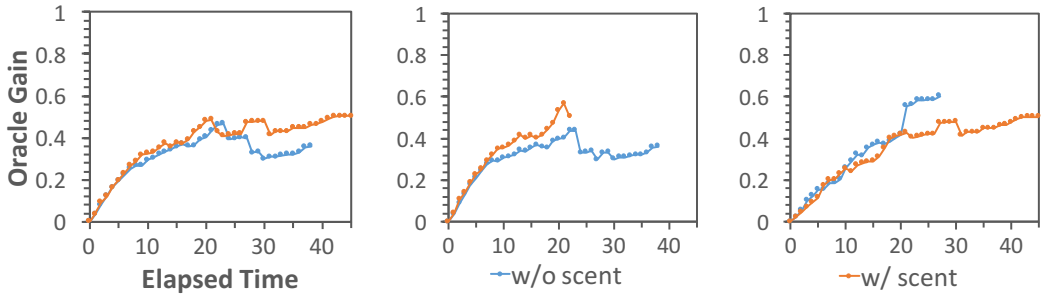
研究課題 **RQ5** に答えるために，被験者がタスク中に費やしたコストと，その結果として獲得した利得の関係性を分析した．本分析では，被験者がタスクを完了するまでに費やした時間（task completion time）をコストとみなした．コストを x 軸（分単位），利得を y 軸としてプロットした散布図を図 5 に示す．同図には，利得を目的変数，コストを説明変数として，単回帰分析を行った結果も含まれている．

全トピックに関する結果（図 5(a)）を見る限り，インタフェース間で，データ点の分布に大きな違いは認められない．単回帰分析によって得られた直線の傾き β についても，提案インタフェースが 0.008（決定係数 $R^2 = 0.187$），ベースラインインタフェースが 0.007（$R^2 = 0.158$）とほぼ同一の値となった．しかし，高相関トピックについては（図 5(b)），提案インタフェースは $\beta = 0.040$（$R^2 = 0.281$），ベースラインインタフェースは $\beta = 0.006$（$R^2 = 0.120$）と大きな差が確認された．また，低相関トピックについては（図 5(c)），提案インタフェースが $\beta = 0.007$（$R^2 = 0.310$），ベースラインインタフェースが $\beta = 0.012$（$R^2 = 0.416$）となり，大小関係が逆転した．この結果から，推定精度の高いトピックでは，未閲覧情報量の提示によって単位時間あたりに獲得可能な利得が上昇したと言える．一方で，推定精度の低いトピックでは，未閲覧情報量の提示が単位時間あたりに獲得可能な利得の低下をまねいていたことが分かる．

6·7 アンケート

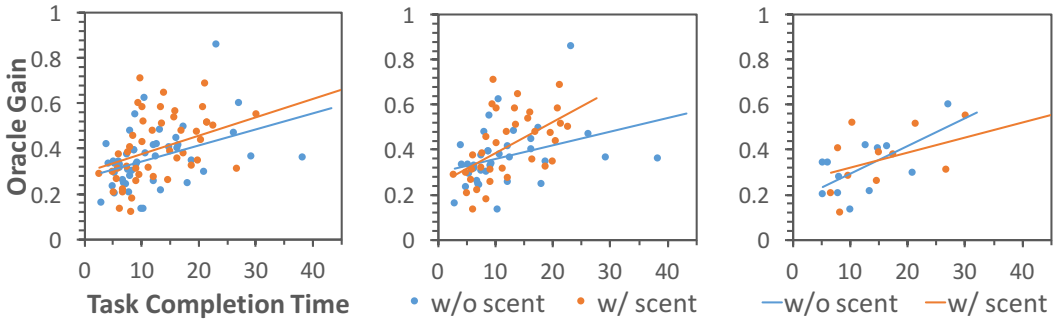
ユーザ実験において各タスクの終了時および実験全体の終了時に実施した，インタフェースのユーザ評価に関するアンケートの結果を分析した．なお，アンケートの各質問に対する回答には 1 から 5 までの 5 段階の尺度を用いた．評価値の平均（および標準偏差）を表 7 に示す．

Q1 および Q2 の結果から，多くの被験者が低相関トピックに関する情報の網羅的な収集に困難を覚えていたことが分かる．Q3 は，その一因が検索結果のランキングにあることを示唆している．このトピックでは，被験者の直感に反する未閲覧情報量が提示されており（Q10–Q11），提案インタフェースはタスクの達成に有効に機能しなかった（Q7–Q9）．Q2 の結果に対してウィルコクソンの順位和検定を行ったところ，低相関トピックに



(a) 全トピック (b) 高相関トピック (c) 低相関トピック

図 4 利得の時間的変化の平均



(a) 全トピック (b) 高相関トピック (c) 低相関トピック

図 5 コストと利得の関係性

おいては，ベースラインインタフェースを用いた方が検索の効率性が有意に高いと被験者が感じていたことが確認された（$p = 0.039$）．この結果から，推定精度の悪化は提案インタフェースに対するユーザの主観的な評価にも悪影響を及ぼすと言える．高相関トピックでは，未閲覧情報量に対する評価に改善が見られた（Q7–Q9 および Q10–Q11）．高相関トピックにおける Q12（$p = 0.008$），Q13（$p = 0.049$），および Q14（$p = 0.020$）の結果に対してウィルコクソンの符号順位検定を行ったところ，提案インタフェースに対する評価値がベースラインインタフェースに比べて有意に高いことが確認された．以上の結果から，提案インタフェースはその有用性や使いやすさという点において被験者から高い評価を受けていたと言える．

7. 考　　察

本章では，実験結果が示唆する内容について考察する．その後，本研究の限界点を整理するとともに，今後の課題を述べる．

7·1 示　　唆

前章の結果から，推定精度の高い未閲覧情報量の提示は網羅性指向タスクの実行に一定の効果をもたらすことが判明した．具体的には，提案インタフェースの利用者は，重要な情報を調べ終えた後で個々の検索を終了していた（6·2 節）．これは，現在のクエリに対して可視化された未閲覧情報量が，検索の終了判断の指針として効果的に機能したためと考えられる．また，提案インタフェースの利用者が，未収集の情報が多く含まれるクエリを次の検索に利用していたこと（6·3 節）や，タスク後半においても重要情報を継続的に収集していたこと（6·4 節）が分かった．これは，推薦クエリに対する未閲覧情報量の提示によって，情報の網羅的な収集に効果的な検索クエリの選択が可能になったためと考えられる．しかし，タスク全体で獲得される利得については，インタフェース間に有意な差が存在しなかった（6·5 節）．この理由としては，十分に検索したと感じる基準には個人差が存在し [Prabha 07]，その基準に基づきタスクが終了された可能性が考えられる．その裏付けとして，6·6 節の分析結果は，インタフェースの種類によらず，高い利得を獲得するまで検索を続ける被験者もいれば，短時間で少量の利得を獲得した時点でタスクを終了する被験者もいることを示している．

その一方で，推定精度の低い未閲覧情報量が提示されると，単位時間あたりに獲得可能な利得が低下してしまうことが分かった（6·6 節）．これは，提示された未閲覧情報量の値やその変化が被験者の直感と合わず，クエリの選択や検索の終了を適切に判断できなかったためと考えられる．

表 7 タスク終了時および実験後のアンケートの結果．太字は有意差（$p < 0.05$）が確認された結果を表す

		質問	全トピック		高相関トピック		低相関トピック	
			w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
検索タスク終了時	Q1	重要な情報の網羅的な収集は難しかったですか？	3.10 (1.32)	3.13 (1.31)	2.72 (1.23)	2.69 (1.19)	4.25 (0.87)	4.42 (0.67)
	Q2	重要な情報の網羅的な収集を効率的に行えましたか？	3.25 (1.21)	3.02 (1.34)	3.56 (1.16)	3.50 (1.18)	**2.33** (0.89)	1.58 (0.51)
	Q3	提示された検索結果は満足のいくものでしたか？	2.92 (1.18)	2.88 (1.35)	3.17 (1.08)	3.31 (1.19)	2.17 (1.19)	1.58 (0.90)
	Q4	推薦されたクエリは満足のいくものでしたか？	3.04 (1.15)	3.33 (1.06)	3.33 (1.04)	3.47 (1.00)	2.17 (1.03)	2.92 (1.16)
	Q5	一連の検索を通じて収集した情報に満足しましたか？	3.43 (1.15)	3.29 (1.30)	3.78 (0.99)	3.75 (1.11)	2.42 (1.00)	1.92 (0.79)
	Q6	調べ切れていない情報がまだ残っていると思いますか？	3.38 (1.38)	3.25 (1.18)	2.97 (1.28)	2.89 (1.09)	4.58 (0.90)	4.33 (0.65)
	Q7	未閲覧情報量は個々の検索の終了判断に役立ちましたか？	—	3.04 (1.35)	—	3.28 (1.32)	—	2.33 (1.23)
	Q8	未閲覧情報量は次の検索に用いるクエリの決定に役立ちましたか？	—	3.33 (1.24)	—	3.39 (1.27)	—	3.17 (1.19)
	Q9	未閲覧情報量は検索タスク全体の終了判断に役立ちましたか？	—	3.15 (1.32)	—	3.44 (1.32)	—	2.25 (0.87)
実験後	Q10	提示された未閲覧情報量の値は納得のいくものでしたか？	—	2.96 (0.91)	—	3.08 (1.00)	—	2.83 (0.83)
	Q11	提示された未閲覧情報量は期待通りに変化しましたか？	—	2.50 (0.83)	—	2.58 (0.90)	—	2.42 (0.79)
	Q12	重要な情報の網羅的な収集に役に立ちましたか？	2.96 (0.86)	**3.79** (1.06)	2.75 (0.87)	**3.92** (0.90)	3.17 (0.83)	3.67 (1.23)
	Q13	使いやすかったですか？	2.96 (0.95)	**3.88** (0.95)	3.00 (0.85)	**3.92** (0.90)	2.92 (1.08)	**3.83** (1.03)
	Q14	また使いたいと思いますか？	2.50 (1.14)	**3.50** (1.06)	2.67 (1.30)	**3.42** (1.24)	2.33 (0.98)	3.58 (1.44)

7·2 限　　界

本研究では未閲覧情報量の計算に必要な利得の値を，サブトピックを明示的にモデル化する手法 [Tsukuda 13] を利用することで推定した．このアプローチの限界点として，文書間で内容が重複する事例に適切に対応できないという点があげられる．我々はユーザ実験の際に，多様化された検索結果を被験者に提示することで，この問題の回避に努めた．しかし，この方法は完全な解決策とは言えない．未閲覧情報量の理想的な表現のためには，サブトピックではなく，より粒度の細かいナゲット [Clarke 08] を単位とした情報を考慮する必要があると我々は考える．この問題への対処として，階層性を考慮したサブトピックのモデリング手法 [Hu 15] を 4 章で述べた利得の計算式および推定手法に組み込むことが今後の重要な課題の 1 つと言える．

第 2 の限界は，提案インタフェースの適用範囲に関するものである．本研究では，提案インタフェースの有効性を検証するために，4 種類の検索トピックのみを対象としたユーザ実験を実施した．これは，本実験の目的に未閲覧情報量の提示がユーザのタスク終了判断に与える影響の調査（**RQ4**）が含まれており，各タスクの所要時間を明示的に設定することが困難であったという理由による（5·4 節）．本論文では，推定利得と真の利得の相関係数の値に基づき，これらの 4 トピックのうち 3 つを高相関トピック，1 つを低相関トピックに分類して分析を行った．そのため，低相関トピックに関する実験結果が，利得の推定精度に起因するものなのか，「恐竜」トピックの内容そのものに起因するものなのかは，今回の実験からは明らかになっていない．今後の別の課題として，本稿で述べた実験結果に関する主張の一般性を検証するために，より多くのトピックを用いた追加実験の実施を行う必要がある．

8. お　わ　り　に

本稿では，検索結果に含まれる未閲覧情報量を提示するクエリ推薦インタフェースを提案した．我々は，クエリに対する未閲覧情報量を，その検索結果中の未閲覧文書集合からユーザが獲得可能な追加利得として定式化し，サブトピックマイニング手法を利用してその値を推定した．24 名の被験者を対象に 4 種類の検索トピックを用いたユーザ実験を行った結果，提案インタフェースの利用者は，未閲覧情報量の推定精度の高いトピックにおいて，(1) 重要な情報を十分に調べた後で個々の検索を終了する，(2) 重要な情報を多く含むクエリを次の検索に利用する，(3) タスクの後半にも重要な情報を継続的に収集する，および (4) 単位時間あたりに獲得可能な利得が上昇することが確認された．一方で，推定精度の低い未閲覧情報量の提示は，検索終了やクエリ選択に大きな影響は与えず，単位時間あたりに獲得可能な利得はかえって低下することが判明した．

今後は，未閲覧情報量のより正確な推定のために，サブトピックの階層や文書間の重複をモデル化し，利得の計算式および推定手法に組み込む必要がある．また，より多くのトピックを用いた追加実験を行うことで，提案手法の適用範囲や主張の一般化可能性を明確にする予定である．

謝　　辞

本研究の一部は，日本学術振興会科学研究費（課題番号：16K16156，16H02906，16H01756，15H01718，13J06404）によるものです．ここに記して謝意を表します．

◇ 参　考　文　献 ◇

[Agrawal 09] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S.: Diversifying Search Results, in *WSDM*, pp. 5–14, (2009)

[Carbonell 98] Carbonell, J. and Goldstein, J.: The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries, in *SIGIR*, pp. 335–336 (1998)

[Cartright 11] Cartright, M.-A., White, R. W., and Horvitz, E.: Intentions and Attention in Exploratory Health Search, in *SIGIR*, pp. 65–74 (2011)
[Chapelle 09] Chapelle, O., Metlzer, D., Zhang, Y., and Grinspan, P.: Expected Reciprocal Rank for Graded Relevance, in *CIKM*, pp. 621–630 (2009)
[Clarke 08] Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I.: Novelty and Diversity in Information Retrieval Evaluation, in *SIGIR*, pp. 659–666 (2008)
[Dostert 09] Dostert, M. and Kelly, D.: Users' Stopping Behaviors and Estimates of Recall, in *SIGIR*, pp. 820–821 (2009)
[Dou 11] Dou, Z., Hu, S., Chen, K., Song, R., and Wen, J.-R.: Multi-dimensional Search Result Diversification, in *WSDM*, pp. 475–484 (2011)
[He 12] He, J., Hollink, V., and Vries, de A.: Combining Implicit and Explicit Topic Representations for Result Diversification, in *SIGIR*, pp. 851–860 (2012)
[Hearst 95] Hearst, M. A.: TileBars: Visualization of Term Distribution Information in Full Text Information Access, in *CHI*, pp. 59–66 (1995)
[Hu 15] Hu, S., Dou, Z., Wang, X., Sakai, T., and Wen, J.-R.: Search Result Diversification Based on Hierarchical Intents, in *CIKM*, pp. 63–72 (2015)
[Iwata 12] Iwata, M., Sakai, T., Yamamoto, T., Chen, Y., Liu, Y., Wen, J.-R., and Nishio, S.: AspecTiles: Tile-based Visualization of Diversified Web Search Results, in *SIGIR*, pp. 85–94 (2012)
[Kato 12] Kato, M. P., Sakai, T., and Tanaka, K.: Structured Query Suggestion for Specialization and Parallel Movement: Effect on Search Behaviors, in *WWW*, pp. 389–398 (2012)
[Kelly 10] Kelly, D., Cushing, A., Dostert, M., Niu, X., and Gyllstrom, K.: Effects of Popularity and Quality on the Usage of Query Suggestions During Information Search, in *CHI*, pp. 45–54 (2010)
[Liu 14] Liu, Y., Song, R., Zhang, M., Dou, Z., Yamamoto, T., Kato, M., Ohshima, H., and Zhou, K.: Overview of the NTCIR-11 IMine Task, in *NTCIR* (2014)
[Pirolli 07] Pirolli, P.: *Information Foraging Theory: Adaptive Interaction with Information*, Human Technology Interaction Series, Oxford University Press (2007)
[Prabha 07] Prabha, C., Connaway, L. S., Olszewski, L., and Jenkins, L. R.: What is Enough? Satisficing Information Needs, *Journal of Documentation*, Vol. 63, No. 1, pp. 74–89 (2007)
[Qvarfordt 13] Qvarfordt, P., Golovchinsky, G., Dunnigan, T., and Agapie, E.: Looking Ahead: Query Preview in Exploratory Search, in *SIGIR*, pp. 243–252 (2013)
[Robertson 09] Robertson, S. and Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond, *Foundations and Trends in Information Retrieval*, Vol. 3, No. 4, pp. 333–389 (2009)
[Sakai 13] Sakai, T., Dou, Z., Yamamoto, T., Liu, Y., Zhang, M., and Song, R.: Overview of the NTCIR-10 INTENT-2 Task, in *NTCIR* (2013)
[Song 11] Song, R., Zhang, M., Sakai, T., Kato, M. P., Liu, Y., Sugimoto, M., Wang, Q., and Orii, N.: Overview of the NTCIR-9 INTENT task, in *NTCIR* (2011)
[Tsukuda 13] Tsukuda, K., Sakai, T., Dou, Z., and Tanaka, K.: Estimating Intent Types for Search Result Diversification, in *AIRS*, pp. 25–37 (2013)
[White 09] White, R. W. and Roth, R. A.: Exploratory Search: Beyond the Query-Response Paradigm, *Synthesis Lectures on Information Concepts, Retrieval, and Services*, Vol. 1, No. 1, pp. 1–98 (2009)
[Wildemuth 12] Wildemuth, B. M. and Freund, L.: Assigning Search Tasks Designed to Elicit Exploratory Search Behaviors, in *HCIR*, pp. 4:1–4:10 (2012)
[Wu 14] Wu, W.-C. and Kelly, D.: Online Search Stopping Behaviors: An Investigation of Query Abandonment and Task Stopping, *Proceedings of the American Society for Information Science and Technology*, Vol. 51, No. 1, pp. 1–10 (2014)
[Zeng 04] Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., and Ma, J.: Learning to Cluster Web Search Results, in *SIGIR*, pp. 210–217 (2004)

〔担当委員：奥 健太〕

2016 年 5 月 9 日　受理

著　者　紹　介

梅本　和俊

東京大学生産技術研究所特任助教および情報通信研究機構ソーシャルビッグデータ研究連携センター研究員．2016 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主に情報検索におけるユーザ行動の分析とその応用に関する研究に従事．情報処理学会，日本データベース学会各会員．

山本　岳洋

京都大学大学院情報学研究科社会情報学専攻助教．2011 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．情報検索におけるユーザインタラクションやユーザ理解に関する研究に従事．情報処理学会，日本データベース学会，電子情報通信学会，ACM 各会員．

田中　克己（正会員）

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