ScentBar: A Query Suggestion Interface Visualizing the Amount of Missed Relevant Information for Intrinsically Diverse Search

Kazutoshi Umemoto*, Takehiro Yamamoto, and Katsumi Tanaka
Kyoto University, Yoshida Honmachi, Sakyo, Kyoto, 606-8501, Japan
{umemoto,tyamamot,tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT

For intrinsically diverse tasks, in which 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 intrinsically diverse tasks, we propose ScentBar, a query suggestion interface visualizing the amount of important information that a user potentially misses collecting from the search results of 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) ScentBar users 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 intrinsically diverse tasks more efficient, unless inaccurate estimates of missed information are visualized.

Keywords

query suggestion interface; intrinsic diversity; search stopping

1. INTRODUCTION

Searchers often issue more than one query and browse multiple documents in exploratory search tasks [31]. These tasks can be characterized as open-ended and multi-faceted, requiring searchers to collect multiple relevant documents [32]. Many people conduct exploratory searches in the medical and health domains [4]. Take, as an example, a searcher who investigates the effect of smoking through Web searches in order to decide whether to continue smoking or not. This search topic contains various aspects such as diseases caused by smoking, increases in the price of cigarettes, and how smoking affects mental health. To fully understand this topic and make appropriate decisions, the searcher needs to collect extensive information covering different aspects. Such tasks are called *intrinsically diverse* search tasks in the literature [25].

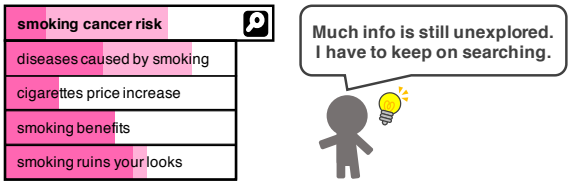


Figure 1: Our concept of visualizing missed information.

There are several problems common to intrinsically diverse tasks. First, searchers often cannot easily come up with effective queries for collecting documents that cover diverse aspects. They have to issue more queries to complete the tasks if search engines return few documents relevant to unexplored aspects. Second, deciding when to stop searching is also difficult for searchers. Quitting the tasks too early without in-depth exploration prevents searchers from finding essential information; on the other hand, if they have already obtained extensive information, continuing these tasks for too long wastes time and effort, as they will not be able to acquire much more gain. The above problems are primarily due to the fact that (1) searchers may not know what aspects exist in the search topic and how important they are and (2) they cannot guess how much important information on each aspect is available on the Web and how much of the information remains unexplored.

The present study proposes a query suggestion interface, which we call ScentBar, with the goal of helping searchers discover unexplored aspects and find the appropriate timing for search stopping in intrinsically diverse tasks. As shown in Figure 1, ScentBar visualizes, for both the search query and suggestion queries, the amount of missed information[1] important to the current search topic in the form of a stacked bar chart so that users can grasp their search progress visually. 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. Gain is formalized as a set-wise metric based on aspect importance, aspect novelty, and per-aspect docu-

*Currently at the University of Tokyo and at the National Institute of Information and Communications Technology

SIGIR '16, July 17–21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07...$15.00

DOI: http://dx.doi.org/10.1145/2911451.2911546

[1] We use the phrase "missed information" to refer to information that has been retrieved by the system but that the searcher misses collecting, similar to the definition by Mansourian and Ford [20].

ment relevance and is estimated by using a state-of-the-art algorithm [30] for subtopic mining and search result diversification.

We conducted a user study involving 24 participants to investigate how ScentBar affected their search strategies and search outcomes. Post-hoc analyses revealed that ScentBar had the following advantages when the gain estimation algorithm worked reasonably: (1) ScentBar users 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 intrinsically diverse tasks more efficient, unless inaccurate estimates of missed information are visualized.

The main contributions of this paper are as follows.

- We formalized an intent-aware metric for evaluating the gain that searchers acquire from a set of their collected documents and provided the explanation of its relation to other metrics for search result diversification.
- We proposed ScentBar, which visualizes the amount of missed information estimated as the additional gain searchers can obtain, and conducted a user study involving 24 participants to investigate the effect of ScentBar on their search strategies and outcomes (*e.g.*, query formulation and search stopping).

2. RELATED WORK

Existing studies related to this work cover the following research areas: (1) search interfaces showing information scent, (2) understanding and modeling searchers' stopping behavior, and (3) subtopic mining for search result diversification.

2.1 Search Interfaces

The concept of *information scent* was introduced in Information Foraging Theory [22], which explains the information seeking behavior of human beings by making an analogy to the food foraging behavior of animals. In this theory, information scent indicates proximal cues from which searchers perceive the value of distal information sources [6]. As summarized in Hearst's book [12], considerable research has been conducted on developing search interfaces that show users information scent [11, 13, 17] with the objective of gaining a better understanding of users' search behavior and/or making their searches more effective.

To help searchers make quick relevance assessments, Hearst [11] proposed the TileBars interface visualizing the query term occurrence as a rectangle for each search result, where the horizontal bar represents the document length and the vertical one represents the query terms. Iwata *et al.* [15] also proposed a tile-based visualization interface, called AspecTiles, which was designed to help users issue queries with multiple aspects. Zha *et al*. [37] proposed an interface that suggests queries with their representative images so that users can efficiently convey their specific search intents. Aiming to help searchers understand inter-query relationships, Kato *et al*. [16] proposed SParQS, which presents query suggestions classified into automatically generated categories.

One of the studies closest to ours is the query preview interface proposed by Qvarfordt *et al*. [24]. Their interface visualizes three kinds of information on the search results of the current search query in the form of a stacked bar chart: (1) the number of newly retrieved search results, (2) the number of re-retrieved but not clicked ones, and (3) the number of re-retrieved ones that have already been clicked by the searcher. While both interfaces provide query-level proximal cues, ScentBar is designed to help searchers conduct intrinsically diverse tasks, where the click information alone is, we think, less informative as they need to collect extensive information covering a variety of different aspects. Therefore, given the complexity of intrinsically diverse tasks, we consider aspect-level factors when formalizing the amount of missed information.

2.2 Search Stopping

Much effort has been devoted to understanding how users decide when to stop searching [20, 23, 33, 35] and to modeling searchers' stopping behavior [18]. Most studies have been based on interviews that clarify the qualitative characteristics of search stopping. Prabha *et al*. [23], for example, interviewed people in academia to analyze how much information is enough to meet their information needs. They reported that study participants had qualitative criteria for search stopping, including whether they feel that sufficient information has been collected (the sense of "good enough" [35]).

Toms and Freund's work [29] is one of the few quantitative studies on search stopping. They analyzed the characteristic actions that preceded people stopping their searches. Wu *et al*. [34] conducted a user study in which they varied the number and distribution of relevant search results and found that these factors had different effects on when searchers left the search engine results pages (SERPs). There have also been attempts to mathematically model search stopping behavior [2]. Maxwell *et al*. [21] very recently investigated different search stopping rules to find which one approximated actual search behavior most closely.

Dostert and Kelly [8] demonstrated through a user study the gap between cognitive and actual recall: while the study participants believed that they had found about 51–60% of relevant documents when they stopped their search tasks, the actual recall they obtained was less than 10% on average. These results suggest that searchers have difficulty in accurately estimating the amount of relevant information they have (or have not) found.

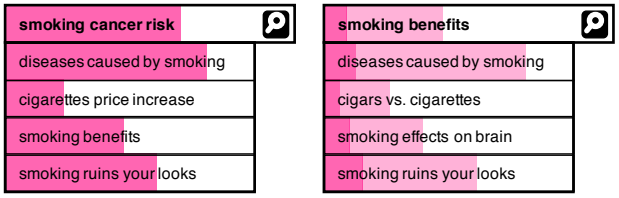
2.3 Subtopic Mining

Researchers have proposed various search result diversification algorithms [1, 9, 27] to satisfy different user intents behind ambiguous or underspecified queries. Most of these algorithms depend on subtopic mining, which is a technique for mining the intents underlying a given query, in order to find a set of documents that cover as many important intents as possible. Approaches to subtopic mining can be classified into two categories: subtopics that are modeled explicitly or implicitly [10]. Maximal Marginal Relevance (MMR) proposed by Carbonell and Goldstein [3] is based on the implicit modeling of subtopics; it iteratively selects a document that is relevant to a given query and dissimilar from already selected documents as the next element of a diversified document list. Approaches belonging to explicit subtopic modeling include IA-Select proposed by Agrawal *et al*. [1], who defined the problem of search result diversification as discovering a set of documents covering important subtopics.

Tsukuda *et al*. [30] proposed one of the state-of-the-art algorithms for explicitly mining subtopics. Their algorithm utilizes different resources to collect subtopics for a given query and clusters the resulting subtopics to identify intents behind the query. We use their algorithm to identify aspects underlying intrinsically diverse tasks. The resulting aspects are used to estimate the amount of missed information for individual queries. We also use the MMR algorithm to diversify search results returned by a search engine.

3. PROPOSED INTERFACE

In this section, we illustrate how ScentBar works through an example and describe the research questions addressed in this work.



(a) At beginning of search task (b) At end of search task

Figure 2: How ScentBar works. The state of missed information after the initial search is shown in Figure 1.

3.1 How It Works

When a user is typing a search query, ScentBar visualizes the amount of *missed information* for both the search query and suggestion queries in the form of a stacked bar chart. To be precise, missed information for a query represents information that (1) can be obtained from the search results of the query, (2) is important to the search topic, and (3) the user has not yet obtained.

Take the search topic described in Section 1 as an example again. Figure 2a illustrates the state of ScentBar visualization when a user starts his/her search task with the query "smoking cancer risk" to learn the effect of smoking. At the beginning of this task, the amount of missed information for a query can be interpreted as the total amount of important information that the user can obtain from the search results of this query. For example, the user can infer from Figure 2a that the queries "smoking cancer risk" and "diseases caused by smoking" return search results containing more important information than those such as "cigarettes price increase".

When the user has obtained sufficient important information from the returned documents, the amount of missed information for the search query decreases, as shown in Figure 1. The length of the bar for the search query (indicated in deep pink) is much shorter than its initial length (indicated in light pink), which suggests this query contains only a small amount of missed information. Note that the bar for the suggestion query "diseases caused by smoking" also has a short length compared to the initial state. This is because the suggestion query shares with the search query a certain amount of important information. In contrast, the visualized bars for other suggestion queries (*e.g.*, "smoking ruins your looks") remain almost unchanged from the initial ones, which suggests great amounts of important information are still unexplored for these queries.

When the user has exhaustively collected important information on the search topic from a variety of angles through several query reformulations, the amount of missed information considerably decreases for any query related to the topic. Figure 2b illustrates the state of ScentBar visualization at the end of the search task. As suggested in the figure, there is little important information left for either the search query or the suggestion queries.

3.2 Research Questions

In the present study, we investigate the effect of displaying missed information on searchers' strategies and outcomes. More specifically, we address the five research questions listed below.

In intrinsically diverse tasks, searchers are more likely to issue multiple queries to exhaustively collect relevant information from different angles. Thus, we frame the following two research questions on individual searches (*i.e.*, after a query is issued before another one is issued or the session ends):

RQ1 *How does ScentBar affect users' decisions on when to stop the current search?*

RQ2 *How does ScentBar affect users' decisions on which query to use for the next search?*

We expect our visualization enables searchers to utilize more rational strategies in individual searches. For example, ScentBar users may be able to stop the current search after collecting a sufficient amount of important information by monitoring the bar visualization for the search query. Missed information may also affect searchers' query formulation strategies; ScentBar users may use a more effective query for the subsequent search by comparing the visualized bars for suggestion queries.

The remaining three research questions relate to the overall search sessions of intrinsically diverse tasks:

RQ3 *How does ScentBar affect the temporal change in gain that users acquire through their search process?*

RQ4 *How does ScentBar affect users' decisions on when to stop their task sessions?*

RQ5 *How does ScentBar affect the relationship between the effort that users expend and the gain that they obtain?*

Missed information may affect searchers' strategies and outcomes at the session level as well as at the query level. For example, ScentBar users may acquire high gain at any point in their sessions. Analogous to **RQ1**, ScentBar may also enable users to rationally decide when to stop the task sessions[2]. It can also be hypothesized that ScentBar makes users' search processes more cost-effective: they may be able to collect a sufficient amount of important information from a variety of angles without expending much effort.

4. MISSED INFORMATION ESTIMATION

In this section, we first introduce *gain*, a metric for evaluating search outcomes in intrinsically diverse tasks, and then define the amount of *missed information* by using the gain metric. We then describe our algorithm of estimating the gain components.

4.1 Gain

Let us consider how searchers can obtain gain in intrinsically diverse tasks. As described in Section 1, these tasks require searchers to collect extensive information covering a variety of different aspects. Considering this characteristic, it would be natural to assume that the gain they obtain through their searches is independent of their browsing order of documents. Thus, we formalize gain as a set-wise metric, whose relation to other metrics is discussed in Section 4.2. We also derive from this characteristic the following requirements that the gain metric should satisfy:

Importance Documents relevant to a central aspect of the search topic produce higher gain than those relevant to a peripheral one.

Relevance Highly relevant documents produce higher gain than partially relevant ones.

Novelty Documents relevant to an unexplored aspect produce higher gain than those relevant to a fully explored aspect.

First, we decompose the gain metric from the topic level to the aspect level in an intent-aware manner [1]. More specifically, given a topic t, we formulate the topic-level gain $\text{Gain-IA}_t(D)$ that can be obtained from a set of documents $D = \{d_1, d_2, \dots\}$ as

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

where A_t is a set of aspects for t, $\Pr(a \mid t)$ is a probability mass function representing the importance of an aspect a to t, and $\text{Gain}_a(D)$ is the per-aspect gain that can be obtained from the documents D with respect to a. To satisfy the first requirement, the above formula puts greater value on the per-aspect gain for highly important aspects when calculating the topic-level gain.

[2] As in the literature [21], we refer to stopping behavior regarding **RQ1** as *query stopping* and that regarding **RQ4** as *session stopping*. *Search stopping* is used as the generic phrase for these two.

Next, the per-aspect gain is calculated by

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

where $\mathrm{Rel}_a(d) \in [0, 1]$ denotes the degree of relevance of a document d to the aspect a. To satisfy the second requirement, this formula sums up the per-aspect relevance degrees of individual documents and returns a high value if D contains many documents that are highly relevant to a. Note that $\mathrm{Gain}_a(\emptyset) = 0$.

The other term $\mathrm{Disc}_a(\cdot)$ in the above formula is a function designed to satisfy the third requirement; it discounts the degree of per-aspect document relevance $\mathrm{Rel}_a(d_i)$ when the aspect a is well covered by a set of documents $\{d_1, \ldots, d_{i-1}\}$ that have already been browsed. We define the discount function as

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

When D' contains many documents that are highly relevant to the aspect a, the above formula returns a value of nearly zero. In this case, newly browsed documents relevant to a receive high discounts from this formula and therefore contribute little to the per-aspect gain for a. Note that $\mathrm{Disc}_a(\emptyset) = 1$.

4.2 Relation of Gain to Others

Gain-IA, the intent-aware gain metric defined above, apparently looks similar to the intent-aware version of Expected Reciprocal Rank (ERR-IA) [5], which is an evaluation metric for diversified search results. The only difference between ERR-IA and Gain-IA is that the latter does not have any reciprocal rank factor. As ERR-IA evaluates the effectiveness of *ordered* search results, it discounts the value of lower-ranked documents. Unlike ERR-IA, our objective is to evaluate the gain that searchers can obtain from *unordered* documents, separately from their expended effort. Thus, we designed Gain-IA to be a set-wise metric that is affected by neither the rank of nor the browsing order of documents. This is obvious from the fact that Gain-IA has the following equivalent form:

$$\text{Gain-IA}_t(D) = \sum_{a \in A_t} \Pr(a \mid t) \cdot \left[1 - \prod_{d \in D} (1 - \mathrm{Rel}_a(d))\right] ,$$

which is derived from the following lemma[3]:

LEMMA 1. $\mathrm{Gain}_a(D) = 1 - \mathrm{Disc}_a(D)$.

Note that the rewritten Gain-IA is exactly the same as the objective function formulated by Agrawal *et al.* [1] for search result diversification. Assuming that $\mathrm{Rel}_a(d)$ represents the probability that a document d satisfies a searcher who believes an aspect a is as important to a topic t as $\Pr(a \mid t)$, this function can be viewed as the probability that the searcher gains satisfaction with the topic t by browsing a set of documents D.

4.3 Missed Information

Using the above-mentioned gain metric, 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. Let D_u be a set of documents that a user u has already browsed in the current task on a topic t and D_q^K be a set of top-K documents returned for a query q. Then, $\mathrm{MI}_{u,t}(q)$, the amount of missed information for q, is defined as

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

Under situations where the above formula behaves ideally, a high value of $\mathrm{MI}_{u,t}(q)$ indicates that unclicked documents $D_q^K \setminus D_u$ contains a large amount of information important to the topic t that the user u misses collecting from the search results of the query q.

[3] Due to space limitation, we omit the proof of Lemma 1, which can be accomplished by mathematical induction.

4.4 Gain Components

The following three components are required to calculate the gain and the amount of missed information on a topic t: (1) A_t, a set of aspects for t; (2) $\Pr(a \mid t)$, a probability representing the importance of an aspect a to t; and (3) $\mathrm{Rel}_a(d)$, the degree of relevance of a document d to a. As for the parameter K (*i.e.*, the number of search results to be fetched), which is also required to calculate the amount of missed information, we report the value used in our experiment in Section 5. We describe below our algorithm of estimating these three components. Note that, if the topic is known beforehand, these components can be estimated before ScentBar users conduct their searches.

As overviewed in Section 2.3, existing approaches to explicitly mining subtopics have the same task of estimating the above components. Thus, we decided to use one of the state-of-the-art explicit subtopic mining algorithms developed by Tsukuda *et al.* [30] to perform this estimation task. Their algorithm, which is easy to implement, achieved the second-best performance among 14 submissions in the Subtopic Mining subtask of the NTCIR-10 INTENT-2 task [26]. In what follows, we describe the outline of the algorithm and clarify some differences from theirs[4]. As described hereinafter, this algorithm relies on a search engine to fetch search results for given queries. The search engine for the algorithm is identical to the one for our search interface (see Section 5 for more details).

Topic Aspects. To estimate a set of aspects A_t for a given topic t, we first mine a set of subtopics S_t for t using three resources, following Tsukuda *et al.* [30]. The first resource consists of query suggestions (*i.e.*, related queries and auto-completion queries) returned by Web search engines in response to the topic query t. Each suggested query is regarded as a subtopic of t. The second resource is query logs, from which queries starting from t are extracted as subtopics. The last resource is the search results for the topic query. The clustering algorithm proposed by Zeng *et al.* [36] is applied to the search results for extracting key phrases as subtopics from individual clusters. Once S_t is mined in this way, Ward's method is applied to S_t for obtaining a set of subtopic clusters $\mathcal{C}_t$, each of which is regarded as belonging to an aspect of the topic t.

As described in Section 5, the search topics used in our experiment were selected from tasks in past NTCIR workshops. Thus, we utilized the resources distributed to the task participants to mine subtopics for these topics.

Aspect Importance. To estimate the importance probability of an aspect of a topic t, we first estimate $\mathrm{Imp}_t(s)$, the importance of a subtopic s to t, by using $\mathrm{Imp}_t(s) = \sum_{d \in D_s^N \cap D_t^N} 1/\mathrm{Rank}_t(d)$, as with Tsukuda *et al.* [30], where $\mathrm{Rank}_t(d)$ denotes the rank of a document d returned for the topic query t. Next, we select a representative subtopic a from each cluster $C \in \mathcal{C}_t$ by using $a = \arg\max_{s \in C} \mathrm{Imp}_t(s)$. The selected subtopic a is regarded as a member of the aspect set A_t for the topic t. Finally, we estimate the importance probability $\Pr(a \mid t)$ of each aspect $a \in A_t$ to t by

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

Per-Aspect Document Relevance. The algorithm proposed by Tsukuda *et al.* [30] performs the relevance estimation only for top-N documents returned for each aspect query $a \in A_t$. This is probably because their objective is to obtain diversified search results comprising about ten documents. However, searchers are expected to browse many more documents in intrinsically diverse tasks. To obtain more accurate estimates of missed information,

[4] We use parameter values reported in the original paper [30].

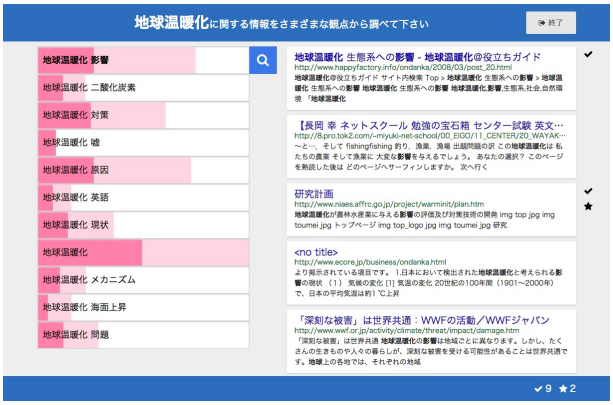


Figure 3: Screenshot of ScentBar (topic: "global warming").

the relevance should be estimated for as many related documents as possible. Thus, we expand the target of the relevance estimation with respect to an aspect a to $D_a = \bigcup_{s \in C_a} D_s^N$, where $C_a \in \mathcal{C}_t$ is a cluster of subtopics belonging to a. In light of the subtopic importance, we estimate $\mathrm{Rel}_a(d)$, the relevance degree of a document $d \in D_a$ to an aspect $a \in A_t$, by

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

where $\mathrm{Rel}_s(d)$ is d's relevance degree to a subtopic s and is estimated by $\mathrm{Rel}_s(d) = 1/\sqrt{\mathrm{Rank}_s(d)}$, as with Tsukuda *et al.* [30].

5. EXPERIMENTAL DESIGN

We conducted a user study in a laboratory setting to investigate the research questions listed in Section 3.2. The details of the experimental design are described below.

5.1 Interfaces

We designed the type of interfaces was to be a within-subjects factor in our user study, where we compared two search interfaces: (1) *w/ scent*, *i.e.*, the proposed interface ScentBar, which visualizes missed information for individual queries; and (2) *w/o scent*, *i.e.*, a baseline interface without missed information visualization.

Figure 3 shows the screenshot of ScentBar for the sample search topic "global warming". Note that the baseline interface has the same appearance as ScentBar except for the query visualization. The brief description of the current search topic is shown at the top of these interfaces. While a user is typing a query into the search box (at the upper left), the interfaces obtain at most ten auto-completion queries through the Google Suggest API[5] and show them at the bottom of the search box. ScentBar also visualizes missed information for these queries in the form of a stacked bar chart. The suggested auto-completion queries remain at the same place even after the user issues the current query to our search system. This behavior allows the user to easily reformulate his/her search query by clicking the bar of a suggestion query. The top K search results returned for the query are shown on the right side of the interfaces. When the user clicks the title of a search result, a document viewer is displayed within the interfaces and the user can browse the landing page on the viewer. The interfaces also have a feature for bookmarking Web pages displayed on the document viewer and present the information on which search results have already been clicked and bookmarked for the current query (on the right side of each search result) and on how many documents have already been clicked and bookmarked through the current search task (at the bottom of the interfaces).

[5] http://www.google.com/complete/search

5.2 Search System

Both interfaces described above used the same internal search system to fetch search results in response to queries. We built the search system on Apache Solr and indexed documents in ClueWeb09-JA, the Japanese portion of the ClueWeb09 collection[6], into this system. We utilized the Okapi BM25 algorithm with default Solr parameters (*i.e.*, $k_1 = 1.2$ and $b = 0.75$) to rank documents.

When testing our implemented interfaces, however, we found that near-duplicate documents sometimes occupied high positions in the search results, which would prevent users from exhaustively collecting information from multiple aspects. To deal with this problem, we decided to diversify the search results by applying the MMR algorithm [3] to R_q^K, the top-K documents retrieved by BM25 for a query q. More precisely, MMR selects d_k, a document ranked at the k-th position, by using

$$d_k = \underset{d \in R_q^K \setminus S^{k-1}}{\arg\max} \left[\lambda \cdot \mathrm{Rel}_q^{\mathrm{BM25}}(d) - (1-\lambda) \max_{d' \in S^{k-1}} \mathrm{Sim}(d, d') \right],$$

where S^{k-1} denotes a set of $k-1$ documents that MMR has already selected and $\mathrm{Rel}_q^{\mathrm{BM25}}(d)$ returns a document d's relevance score[7] that BM25 calculates for a query q. As for $\mathrm{Sim}(\cdot,\cdot)$, we used the cosine similarity between tf-idf vectors, each of which was constructed from a set of terms appearing in the title and snippet areas.

In this way, the resulting list of top-K diversified documents D_q^K was returned to the experimental interfaces in response to a query q. The control parameter was set to $\lambda = 0.3$, following Tsukuda *et al.* [30] and Dou *et al.* [9]. The number of search results to be fetched was set to $K = 100$, following Qvarfordt *et al.* [24].

5.3 Search Topics

As the topic source, we decided to use the NTCIR INTENT-1 [28], INTENT-2 [26], and IMine-1 [19] tasks, where the aspects of each topic and their imporance are provided for evaluating search result diversification. We selected four topics that satisfied the characteristics of intrinsically diverse tasks (*i.e.*, having diverse documents relevant to different aspects). These topics are listed in Table 1. Two of them were selected from the medical domain, in which people often conduct exploratory searches to collect diverse information from multiple sources [4], while the remaining two were from non-medical domains. Table 1 also shows a subset of *oracle* aspects for these topics, which were provided by the NTCIR task organizers. As shown in this table, there are a number of different aspects in these topics (the mean number of oracle aspects per topic was 10.50), and multiple relevant documents may be needed to cover each aspect.

We used these topics to design four types of intrinsically diverse search tasks. More specifically, the following description was given to the participants of our user study for the search task on a topic t:

> *You were given the assignment of submitting a thorough report on the topic t. To fully understand t, collect relevant information on this topic from a number of different aspects that you think is important. You may end this search task when you feel there is little important information left.*

5.4 Procedure

We recruited study participants via the recruiting Web site of our university. As a result, we had 24 participants (4 females and 20 males) in our user study: 16 undergraduate students, 6 graduate students, and 2 researchers. Their mean age was 23.29 (SD = 3.67)

[6] http://www.lemurproject.org/clueweb09.php

[7] We normalized $\mathrm{Rel}_q^{\mathrm{BM25}}(\cdot)$ to $[0, 1]$ by using the BM25 score of a document ranked at the first position for a query q.

Table 1: Search topics (with oracle aspects) used in our user study (translated from Japanese).

ID	Topic	Aspects
T1	**symptoms of diabetes**	causes, early symptoms, terminal illness, subjective symptoms, prevention, complicating diseases, drowsiness, ...
T2	**clinical depression**	symptoms, work, society, communication, treatment
T3	**dress codes for wedding ceremony**	males, females, families and relatives, hairstyles, after-party, invited guests, wedding reception, ...
T4	**dinosaurs**	fossils, extinction, pictorial books, species, museums, Jurassic period, images, ...

and their fields of study included informatics (six participants), engineering (four), law (four), science (three), *etc*. At the end of the experiment, they received bookstore gift cards (equivalent to $16 per person) as a reward for participating. Prior to the actual tasks, participants were assigned a training task (on the topic of global warming) to get familiar with the tasks and interfaces. During the training task, we provided participants with detailed instructions on how to use the interfaces and on what they were expected to do in our experiment. The training task took approximately 15 minutes.

A demographic questionnaire was administered at the beginning of the experiment. After filling out the questionnaire, participants worked on the four tasks one by one (two tasks with ScentBar and two with the baseline interface). The order of interfaces and tasks was determined with a Graeco-Latin square to remove the ordering effect of these variables. All tasks followed the same procedure. First, participants read the task description and completed a pre-task questionnaire asking about their knowledge on the topic and their estimated difficulty of the task. Next, they started the task with the assigned interface. To make the search process as realistic as possible, they were allowed to freely formulate search queries and were asked to use the bookmark feature provided by both interfaces when they found documents useful for completing the task. We did not allow them to follow links in the landing page because we could not guarantee that the linked pages were stored in ClueWeb09-JA, the document corpus used by our search system. After completing the task, participants filled out a post-task questionnaire asking about task-level feedback (*e.g.*, their experienced difficulty). Finally, we administered an exit questionnaire at the end of the experiment to gather feedback on both interfaces.

As mentioned above, we asked the participants to finish each task when they felt they had exhaustively collected relevant information from various aspects of the search topic (*i.e.*, each task had no time limit). We did not tell them the total number of tasks so as to avoid its effect on their stopping decisions, since they knew in advance from the recruitment information that the maximum amount of time required for our experiment was two hours. Instead, we instructed them that there were a certain number of tasks and that the experiment would finish when they had completed all tasks or two hours had passed from the beginning of the experiment. Note that, in a pilot study, every task was completed within 20 minutes. In fact, all participants completed the whole experiment within two hours.

6. RESULTS

This section reports the results of our user study and attempts to answer the research questions listed in Section 3.2. As normality was not guaranteed for the experimental data, we used non-parametric significance tests in our post-hoc analyses. Significant effects are reported on the significance level $\alpha = 0.05$.

6.1 Ground-Truth Data

To evaluate gain that participants obtained, we prepared the ground-truth data for each search topic used in our user study. As described in Section 4.4, the following three components are required to calculate the gain and the amount of missed information: (1) a set of aspects for the search topic, (2) the importance of each aspect, and (3) per-aspect document relevance. As these topics were selected from the NTCIR tasks, we used oracle data provided by the task organizers as for the first two components.

Table 2: Correlations between oracle and estimated gains. The top three values for each measure are shown in bold face.

Topic	Pearson's r	Spearman's ρ	Kendall's τ
T1	**0.834**	**0.851**	**0.683**
T2	**0.845**	**0.860**	**0.678**
T3	**0.824**	**0.862**	**0.713**
T4	0.710	0.702	0.529

Relevance Assessment. To prepare the remaining ground-truth data (*i.e.*, per-aspect document relevance), three assessors judged the relevance of each document that had been browsed by participants. A four-grade scale was used to assess the relevance of a document for each oracle aspect: irrelevant ($= 0$), partially relevant ($= 1$), highly relevant ($= 2$), and perfect ($= 3$). The following criteria were shared between assessors to make the assessed relevance as consistent and reliable as possible: documents covering (1) less than 30%, (2) at least 30% but less than 60%, (3) at least 60% but less than 90%, and (4) more than 90% of the information on a certain aspect should be respectively labeled as (1) irrelevant, (2) partially relevant, (3) highly relevant, and (4) perfect for that aspect. The resulting relevance grade $g_{a,d} \in \{0, 1, 2, 3\}$ of a document d for an aspect a was converted into the oracle relevance degree by $\mathrm{Rel}^*_a(d) = (2^{g_{a,d}} - 1)/2^3$. This conversion is often used when calculating evaluation metrics, such as ERR [5], and guarantees that the resulting relevance degree is any of $\{0, 0.125, 0.375, 0.875\}$.

Accuracy of Gain Estimation. The amount of missed information is defined as the additional gain, which is estimated by the algorithm mentioned in Section 4.4. If the estimated gain is far different from the oracle one, our interface ends up showing searchers missed information that is unrealistic and unreliable, which may have a negative effect on their searches. Thus, we decided to investigate the gap between the gain estimated by our algorithm and that calculated using the ground-truth data, before conducting post-hoc analyses. The former is referred to as *estimated gain* and the latter as *oracle gain* hereinafter.

To this end, we calculated the estimated and oracle gain scores at each point where participants browsed documents and measured correlations for pairs of gain scores using Pearson's r, Spearman's ρ, and Kendall's τ. Note that strong correlation for a topic indicates high consistency between the change in estimated gain and that in oracle gain, which suggests the gain algorithm works reasonably for that topic. Table 2 shows the values of these three correlation measures calculated for each topic.

As shown in this table, quite high correlation was observed for all topics except T4 ("dinosaurs"), which suggests that ScentBar users who conducted searches on the topic T4 might have experienced unexpected changes in visualized missed information. When analyzing the possible causes of the inaccurate gain estimation for T4, we found that aspects mined by our algorithm as important to this topic included "merchandising sales" and "skeleton models selling". As can be seen in Table 1, none of these were contained in the oracle aspects for this topic. We also found that many documents retrieved by the query "dinosaurs" at high positions tended to be

Table 3: Mean change (with SD) in oracle missed information at the query level. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.

	All topics		HC topics		LC topic	
	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)

Table 4: Mean (with SD) of oracle missed information on query issuing. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.

	All topics		HC topics		LC topic	
	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)

less relevant to the oracle aspects. As described in Section 4.4, the gain estimation algorithm relies on aspect mining and document ranking. Thus, the aspect mismatch and poor document ranking would cause the estimated gain to be inaccurate for the topic T4.

As noted above, inaccurate gain estimation could have an undesirable effect on participants' fulfilling search tasks; they might get confused by seeing unreliable missed information and quit searching at an inappropriate point. To investigate the effect of the accuracy of the estimated gain, the analyses reported in the remainder of this section were conducted at the following three topic levels: (1) all topics; (2) *HC topics*, which comprise T1, T2, and T3, where the estimated gain had high correlation with the oracle one; and (3) *LC topic*, which comprises T4, where low correlation was observed between the estimated gain and the oracle one.

6.2 Query Stopping

To answer **RQ1**, we investigated the effect of missed information on query stopping (*i.e.*, the point at which a user stops examining search results for the current query). We hypothesized that ScentBar users would stop the current search once they had exhaustively collected relevant documents from the search results. If this is the case, query stopping should occur when the amount of missed information for the query greatly decreases from its initial value. Thus, for each query issued by participants, we calculated the change in the amount of oracle missed information between the beginning of the task and the query stopping point. A large change indicates that the searcher obtained a great amount of relevant information from the current search.

Table 3 shows the mean change (with SD) in the amount of oracle missed information at the query level. As for all topics, the amount of oracle missed information decreased greatly when participants used ScentBar. The difference from the baseline interface was shown to be significant by the Mann-Whitney U test ($p = 0.001$). The greater difference was observed for the HC topics ($p < 0.001$). In contrast, the change in oracle missed information was quite small for the LC topic ($p = 0.683$). These results indicate that when gain was estimated reasonably accurately, ScentBar users collected a greater amount of relevant information through individual searches than the baseline interface users, which supports our hypothesis on query stopping. When inaccurate estimates were presented to participants, however, ScentBar had no effect on the amount of collected information at the query level.

6.3 Query Selection

To answer **RQ2**, we investigated the effect of missed information on users' query selection decisions in the subsequent search. We hypothesized that ScentBar would help users issue queries whose search results contained much missed relevant information. To test this hypothesis, we calculated, for each query issued by participants, the amount of oracle missed information on query issuing.

Table 4 shows the mean (with SD) of the amount of oracle missed information on query issuing. On average, queries with large amounts of oracle missed information were issued through ScentBar rather than through the baseline interface for all topics. The Mann-Whitney U test revealed a significant difference between these interfaces ($p = 0.002$). While a clearer trend was observed for the HC topics ($p < 0.001$), the difference was insignificant for the LC topic ($p = 0.521$). These results indicate that when the gain was estimated reasonably, ScentBar users issued queries more effective for exhaustively collecting relevant information than the baseline interface users. This supports our hypothesis on query selection. When the gain estimation was inaccurate, however, queries issued through the two interfaces differed little from one another in terms of the amount of oracle missed information on query issuing.

6.4 Gain Change

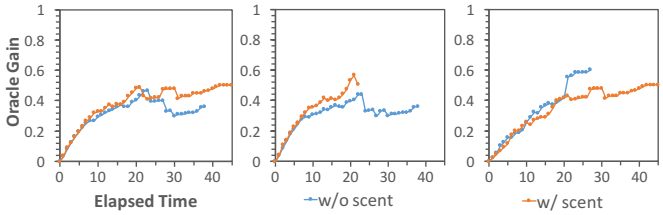
To answer **RQ3**, we investigated the effect of missed information on the change in gain over time. More specifically, we split participants' task sessions into one-minute segments and calculated the (cumulative) oracle gain they had obtained at each cutoff time point. Note that when calculating the gain at a time point, we excluded participants who had completed the task before that time.

Figure 4 shows the gain curves averaged over the individual interfaces. The two gain curves for all topics (Figure 4a) behaved similarly until seven minutes had passed from the task initiation. After that, the oracle gain for ScentBar outperformed that for the baseline interface at nearly every time point. The difference between the interfaces widened for the HC topics (Figure 4b) while the gain curve for ScentBar underperformed that for the baseline interface at most time points (Figure 4c). These results indicate that when the gain estimation was accurate enough, ScentBar users obtained as much gain at an early time of their task sessions as, and higher gain after that time than, the baseline interface users. In contrast, when ScentBar presented inaccurate estimates to participants, they obtained less gain than the baseline interface users who spent the same amount of time or more searching.

6.5 Session Stopping

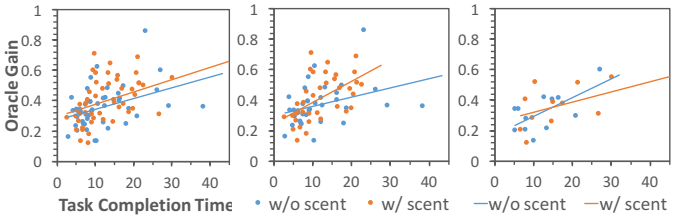
To answer **RQ4**, we investigated the effect of missed information on session stopping (*i.e*, the point at which a user completed the current search task). As in the case of our analysis on query stopping (reported in Section 6.2), we hypothesized that ScentBar users would complete the current task after exhaustively collecting information relevant to the search topic. If this is the case, session stopping should occur when the amount of oracle missed information for most queries related the task greatly decreases from the initial values. Thus, for each query issued by participants, we calculated the change in the amount of oracle missed information between the beginning and end of the task. In addition, we also calculated the (total) oracle gain obtained from the documents that individual participants had collected through their task sessions.

In Table 5, the mean change (with SD) in the amount of oracle missed information at the session level is shown at the top, and the mean (with SD) of the oracle gain obtained through task sessions is at the bottom. As for the former measure, the amount of oracle missed information decreased greatly for all topics when participants used ScentBar. The Mann-Whitney U test revealed a significant difference from the baseline interface ($p = 0.002$). The larger difference was observed for the HC topics ($p < 0.001$), while there was no significant difference for the LC topic ($p = 0.845$). As for



(a) All topics (b) HC topics (c) LC topic

Figure 4: Mean change in oracle gain over time.



(a) All topics (b) HC topics (c) LC topics

Figure 5: Relationship between effort and gain.

Table 5: Top: mean change (with SD) in oracle missed information at the session level. Bottom: mean (with SD) of oracle gain obtained through task sessions. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.

	All topics		HC topics		LC topic	
	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)

the latter measure, the mean oracle gain for ScentBar was higher for all topics than that for the baseline interface. While the same trend was observed for both the HC and LC topics, we could not find any significant differences between the two interfaces ($p = 0.097$ for all topics, $p = 0.191$ for the HC topics, and $p = 0.410$ for the LC topic). These results partially support our hypothesis on session stopping: when the gain estimation algorithm worked reasonably, ScentBar users completed their tasks after collecting greater amounts of relevant information from the SERPs of their issued queries. However, the resulting total gain obtained through ScentBar did not show statistically significant improvement compared to the baseline interface, which implies some users of ScentBar might complete their tasks earlier than others who used the baseline interface. The analysis related to this issue is reported in Section 6.6.

6.6 Relationship Between Effort and Gain

To answer **RQ5**, we investigated the effect of missed information on the relationship between expended effort and acquired gain. In the present study, the time that participants spent completing the tasks was regarded as the effort that they expended. We plotted scatter charts in which the x-axis represented the task completion time (in minutes) and the y-axis represented the oracle gain that participants obtained through their task sessions. We also performed a linear regression analysis to better understand the relationship between the effort (modeled as the explanatory variable) and the gain (as the response variable).

Figure 5 shows the scatter charts and regression lines of the effort-gain data. As can be seen in Figure 5a, little difference existed between the two interfaces for all topics. In fact, the slope β of both regression lines was almost the same: $\beta = 0.008$ ($R^2 = 0.187$) for ScentBar and $\beta = 0.007$ ($R^2 = 0.158$) for the baseline interface. As for the HC topics (Figure 5b), the great improvement was observed for ScentBar ($\beta = 0.014, R^2 = 0.281$), while the slope was nearly unchanged for the baseline interface ($\beta = 0.006, R^2 = 0.120$). In contrast, for the LC topic (Figure 5c), the slope of the regression line for ScentBar ($\beta = 0.007, R^2 = 0.310$) was lower than that for the baseline interface ($\beta = 0.012, R^2 = 0.416$). These results indicate that when the gain estimation was accurate enough, ScentBar users obtained higher gain per unit time than the baseline users. However, the former users could not obtain gain as effective as the latter ones when ScentBar displayed inaccurate estimates. Figure 4 also demonstrates that some ScentBar users completed tasks earlier without expending much effort while others spent longer to obtain much more gain.

Table 6: Results of search interaction analysis. The mean (with SD) values are reported. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.

		All topics		HC topics		LC topic	
		w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
search behavior	**#Queries**	10.417 (6.523)	11.563 (10.886)	9.444 (6.381)	8.944 (7.131)	13.333 (6.315)	19.417 (15.963)
	%Sugg Queries	0.376 (0.230)	0.480 (0.275)	0.355 (0.238)	**0.505** (0.279)	0.437 (0.201)	0.405 (0.259)
	#Clicks LastRel-Q	0.156 (0.494)	0.164 (0.511)	0.165 (0.543)	0.175 (0.465)	0.133 (0.342)	0.142 (0.596)
	#Clicks LastRel-S	0.208 (0.617)	0.563 (1.610)	0.278 (0.701)	0.389 (0.871)	0.000 (0.000)	1.083 (2.875)
	#Docs	22.333 (14.961)	25.729 (15.284)	22.667 (13.980)	24.972 (15.242)	21.333 (18.252)	28.000 (15.857)
search outcomes	**Precision**	0.714 (0.126)	0.723 (0.106)	0.685 (0.114)	0.729 (0.111)	0.801 (0.124)	0.705 (0.091)
	Recall	0.078 (0.050)	**0.093** (0.057)	0.078 (0.047)	0.095 (0.061)	0.076 (0.060)	0.088 (0.043
	Aspect Coverage	0.718 (0.154)	0.757 (0.157)	0.731 (0.162)	0.756 (0.162)	0.677 (0.125)	0.760 (0.146)

6.7 Search Interaction

We analyzed participants' interaction data to investigate how ScentBar affects their search behavior in intrinsically diverse tasks. To investigate its effect on query selection, we calculated the number of queries issued by participants (#Queries) and the fraction of suggestion queries selected by them (%SuggQueries). As document-related measures, we calculated the number of, precision of, and recall of documents browsed by participants (#Docs, Precision, and Recall, respectively) and the fraction of aspects covered by their browsed documents (AspectCoverage). We also calculated the number of clicks after the last relevant document both at query and session levels (#ClicksLastRel-Q and -S) for analyzing search stopping behavior. While Precision, Recall, and AspectCoverage are interaction metrics related to search outcomes, the other metrics are related to search behavior.

Table 6 shows the results of our analysis based on these measures. On average, more than ten queries were issued through both interfaces for all topics. While the difference in #Queries between the interfaces was not significant according to the Mann-Whitney U test, ScentBar users issued significantly more suggestion queries in the HC topics ($p = 0.017$). We also calculated some query complexity measures (*e.g.*, query length), but did not find any clear trends. As for the document-related measures, ScentBar users, on average, browsed more documents covering more aspects, although the difference was not statistically significant. Remarkably, ScentBar users obtained significantly higher recall than baseline interface users ($p = 0.048$), which demonstrates the advantage of ScentBar because recall is one of the most desired properties in intrinsically diverse tasks. Last but not least, no significant difference

was observed for the number of irrelevant document clicks just before search stopping. This result was contrary to our expectation that ScentBar could prevent searchers from continuing tasks when no additional gain would be obtained. We discuss this issue in more detail in Section 7.

6.8 Questionnaires

Finally, we report the results of the post-task and exit questionnaires, which are summarized in Table 7. Note that the per-interface results of the exit questionnaire for the HC and LC topics in the table were calculated from the ratings given by participants who addressed those topics with that interface.

The results for the questions Q1 and Q2 indicate that most participants faced difficulty exhaustively collecting relevant information in the LC topic. The question Q3 suggests that the low quality of search results presented by the search system would be a possible cause of this difficulty. As unreliable missed information was displayed in this topic (Q10 and Q11), ScentBar did not guide participants in effective decision-making on search stopping (Q7 and Q9). In fact, The Mann-Whitney U test revealed that the baseline interface was rated significantly better than ScentBar for the question Q2 ($p = 0.039$), which indicates that the inaccurate estimation had harmful effects on the subjective efficiency for ScentBar, as well as the objective one (Section 6.6). The results for the questions Q7–Q11 indicate that the ratings for missed information improved in the HC topics. As for usability, ScentBar was much more preferred by participants than the baseline interface (Q12–Q14). The differences between the two interfaces for these questions were shown to be significant by the Wilcoxon signed-rank test ($p < 0.001$ for Q12, $p < 0.001$ for Q13, and $p = 0.004$ for Q14).

7. DISCUSSION

Our user study uncovered how ScentBar affected users' search strategies and search outcomes. This section discusses the implications of our experimental results and the limitations of this work.

7.1 Implications

Our results demonstrated that ScentBar had the following advantages when the gain estimation algorithm worked reasonably: (1) ScentBar users stopped individual searches after collecting much relevant information from the SERPs (Section 6.2); (2) they issued search queries whose search results contained much missed information (Section 6.3); (3) they obtained high gain, particularly at the late stage of their sessions (Section 6.4); and (4) they obtained high gain per unit time (Section 6.6). Our post-hoc analysis of search interaction data revealed that query suggestion was more frequently used by ScentBar users (Section 6.7), which implies that suggesting queries with missed information affected searchers' query formulation strategies and thus enabled them to conduct intrinsically diverse tasks more efficiently.

The search interaction analysis also revealed that displaying missed information did not affect the number of irrelevant document clicks just before search stopping. A possible explanation for this phenomenon is that seeing visualized missed information made participants more cautious about making decisions on search stopping. In fact, ScentBar users spent, on average, longer completing the tasks (13.6 minutes) than the baseline interface users (11.8 minutes), although the difference was not statistically significant. These results suggest that additional features would be needed for supporting searchers' decision-making on search stopping. While ScentBar suggests which queries have much missed information, it does not provide any clues about which search results users should assess to obtain additional gain. Possible solutions to this issue include visualizing the per-aspect relevance for each search result, like AspecTiles [15], and/or re-ranking search results containing much missed information at high positions, both of which can be incorporated with ScentBar. Another approach could be notifying searchers of the effort they have expended so far as well as their missed information. This could help searchers more rationally decide whether they should conduct additional searches to obtain more gain or stop searching to avoid wasting time. Techniques for quantifying search cost (*e.g.*, [2]) would be useful for implementing this idea.

Another finding from our user study is that ScentBar worsened rather than improved search performance when missed information was estimated inaccurately. More specifically, the gain per unit time obtained by ScentBar users was lower than that by the baseline interface users. This is probably because unconvincing visualization of missed information and its unexpected change prevented searchers from making rational decisions on query selection and search stopping. Our results suggest that the accuracy of gain estimation is critical for ScentBar to produce intended effects.

7.2 Limitations

This work has several limitations that we should acknowledge. First, while we formalized gain as an intent-aware metric, as with other evaluation metrics like ERR-IA [5], this approach has a drawback in that the content overlap among documents is not taken into account when the gain is evaluated. Thus, this metric has the potential to give a high score to documents with high overlap. While we tried to prevent participants from collecting overlapped documents by presenting diversified search results to them, this is an ad-hoc solution, not an optimal one. Ideally speaking, gain should be modeled on the basis of more fine-grained units, such as nuggets [7]. Thus, for more accurate evaluation, our gain formalization and estimation algorithm would require further improvement by, for example, modeling and mining aspects hierarchically [14].

Second, we used only four search topics to evaluate the effectiveness of ScentBar in intrinsically diverse tasks. Limiting the number of topics was necessary given that the interfaces were designed to be the within-subject factor in our user study (*i.e.*, every participant used both interfaces) and that we could not use a predefined length of task execution time as our research questions included the effect of ScentBar on session stopping. To validate the generality of our claim about the effectiveness of ScentBar, we would need to conduct additional experiments with more search topics and more diverse participants in the future.

Third, we selected a simple interface as the baseline in our user study while various interfaces showing information scent have been proposed. As outlined in Section 2.1, the query preview interface proposed by Qvarfordt *et al*. [24] provides a similar functionality as ours: their interface visualizes the number of (un)clicked search results for the current query. To estimate its effectiveness in intrinsically diverse tasks, we measured correlations between the number of unclicked search results and the oracle missed information scent in a manner similar to the gain accuracy estimation (Section 6.1). The correlation coefficient was 0.334 even in the best topic. Given the fact that ScentBar worsened the search performance for the LC topic, this suggests that such simple visualization could be less effective for intrinsically diverse tasks, while direct comparison in a user study setting would be needed to gain more insights.

8. CONCLUSIONS

In this paper, we proposed ScentBar, a query suggestion interface visualizing the amount of missed information for individual queries. We defined the amount of missed information for a query as the additional gain that can be obtained from unclicked search

Table 7: Mean rating scores (with SD) of post-task and exit questionnaires in which five-point scale was used for each question. Significant differences from the baseline interface at $p < 0.05$ are shown in bold face.

		Question	All topics		HC topics		LC topic	
			w/o scent	w/ scent	w/o scent	w/ scent	w/o scent	w/ scent
post-task	Q1	Was it difficult to exhaustively collect relevant information?	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	Did you exhaustively collect relevant information in an efficient manner?	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	Were you satisfied with the quality of the search results?	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	Were you satisfied with the quality of the suggestion queries?	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	Were you satisfied with the information you collected?	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	How much relevant information do you think was left unexplored?	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	Was missed information useful for your decision on query stopping?	—	3.04 (1.35)	—	3.28 (1.32)	—	2.33 (1.23)
	Q8	Was missed information useful for your decision on query selection?	—	3.33 (1.24)	—	3.39 (1.27)	—	3.17 (1.19)
	Q9	Was missed information useful for your decision on session stopping?	—	3.15 (1.32)	—	3.44 (1.32)	—	2.25 (0.87)
exit	Q10	Was the amount of missed information convincing?	—	2.96 (0.91)	—	3.08 (1.00)	—	2.83 (0.83)
	Q11	Did missed information change as you expected?	—	2.50 (0.83)	—	2.58 (0.90)	—	2.42 (0.79)
	Q12	Was this interface useful for exhaustively collecting relevant information?	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	Was this interface easy to use?	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	Do you want to use this interface again?	2.50 (1.14)	**3.50** (1.06)	2.67 (1.30)	**3.42** (1.24)	2.33 (0.98)	3.58 (1.44)

results of the query. Results of our user study involving 24 participants showed the following advantages of our interface for search topics where gain was estimated reasonably accurately: (1) ScentBar users stopped SERP examination after collecting more 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 unless inaccurate estimates of missed information were visualized, ScentBar enabled users to utilize effective query formulation strategies, while it had little noticeable effect on their search stopping behavior.

Future directions of this work include exploring ways for utilizing missed information. While ScentBar visualizes what is left for browsing (*i.e.*, the visualization decreases as searchers obtain gain), showing how much searchers have browsed might have a different effect on their search behavior. Providing the visualization in different places and/or at different timing would also be worth exploring. Missed information could also be incorporated with query suggestion algorithms to rank suggestion queries from which searchers would be able to obtain much additional gain at high positions.

9. ACKNOWLEDGMENTS

This work was supported in part by JSPS Grants-in-Aid for Scientific Research (Nos. 13J06404, 15H01718, and 16K16156).

10. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.

[2] L. Azzopardi and G. Zuccon. An analysis of theories of search and search behavior. In *ICTIR*, pages 81–90, 2015.

[3] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336, 1998.

[4] M.-A. Cartright, R. W. White, and E. Horvitz. Intentions and attention in exploratory health search. In *SIGIR*, pages 65–74, 2011.

[5] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, pages 621–630, 2009.

[6] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *CHI*, pages 490–497, 2001.

[7] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pages 659–666, 2008.

[8] M. Dostert and D. Kelly. Users' stopping behaviors and estimates of recall. In *SIGIR*, pages 820–821, 2009.

[9] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *WSDM*, pages 475–484, 2011.

[10] J. He, V. Hollink, and A. de Vries. Combining implicit and explicit topic representations for result diversification. In *SIGIR*, pages 851–860, 2012.

[11] M. A. Hearst. TileBars: Visualization of term distribution information in full text information access. In *CHI*, pages 59–66, 1995.

[12] M. A. Hearst. *Search User Interfaces*. Cambridge University Press, 2009.

[13] O. Hoeber and X. D. Yang. A comparative user study of web search interfaces: Hotmap, concept highlighter, and google. In *WI*, pages 866–874, 2006.

[14] S. Hu, Z. Dou, X. Wang, T. Sakai, and J.-R. Wen. Search result diversification based on hierarchical intents. In *CIKM*, pages 63–72, 2015.

[15] M. Iwata, T. Sakai, T. Yamamoto, Y. Chen, Y. Liu, J.-R. Wen, and S. Nishio. AspecTiles: Tile-based visualization of diversified web search results. In *SIGIR*, pages 85–94, 2012.

[16] M. P. Kato, T. Sakai, and K. Tanaka. Structured query suggestion for specialization and parallel movement: Effect on search behaviors. In *WWW*, pages 389–398, 2012.

[17] D. Kelly, A. Cushing, M. Dostert, X. Niu, and K. Gyllstrom. Effects of popularity and quality on the usage of query suggestions during information search. In *CHI*, pages 45–54, 2010.

[18] D. Kraft and T. Lee. Stopping rules and their effect on expected search length. *IPM*, 15(1):47–58, 1979.

[19] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. Kato, H. Ohshima, and K. Zhou. Overview of the NTCIR-11 IMine task. In *NTCIR-11*, 2014.

[20] Y. Mansourian and N. Ford. Search persistence and failure on the web: a "bounded rationality" and "satisficing" analysis. *JDoc*, 63(5):680–701, 2007.

[21] D. Maxwell, L. Azzopardi, K. Järvelin, and H. Keskustalo. Searching and stopping: An analysis of stopping rules and strategies. In *CIKM*, pages 313–322, 2015.

[22] P. Pirolli. *Information Foraging Theory: Adaptive Interaction with Information*. Human Technology Interaction Series. Oxford University Press, 2007.

[23] C. Prabha, L. S. Connaway, L. Olszewski, and L. R. Jenkins. What is enough? satisficing information needs. *JDoc*, 63(1):74–89, 2007.

[24] P. Qvarfordt, G. Golovchinsky, T. Dunnigan, and E. Agapie. Looking ahead: Query preview in exploratory search. In *SIGIR*, pages 243–252, 2013.

[25] K. Raman, P. N. Bennett, and K. Collins-Thompson. Toward whole-session relevance: Exploring intrinsic diversity in web search. In *SIGIR*, pages 463–472, 2013.

[26] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the NTCIR-10 INTENT-2 task. In *NTCIR-10*, 2013.

[27] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *WWW*, pages 881–890, 2010.

[28] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT task. In *NTCIR-9*, 2011.

[29] E. G. Toms and L. Freund. Predicting stopping behaviour: A preliminary analysis. In *SIGIR*, pages 750–751, 2009.

[30] K. Tsukuda, T. Sakai, Z. Dou, and K. Tanaka. Estimating intent types for search result diversification. In *AIRS*, pages 25–37, 2013.

[31] R. W. White and R. A. Roth. Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

[32] B. M. Wildemuth and L. Freund. Assigning search tasks designed to elicit exploratory search behaviors. In *HCIR*, pages 4:1–4:10, 2012.

[33] W.-C. Wu and D. Kelly. Online search stopping behaviors: An investigation of query abandonment and task stopping. In *ASIS&T Annual Meeting*, 2014.

[34] W.-C. Wu, D. Kelly, and A. Sud. Using information scent and need for cognition to understand online search behavior. In *SIGIR*, pages 557–566, 2014.

[35] L. Zach. When is "enough" enough? modeling the information-seeking and stopping behavior of senior arts administrators. *JASIST*, 56(1):23–35, 2005.

[36] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *SIGIR*, pages 210–217, 2004.

[37] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *MM*, pages 15–24, 2009.