AspecTiles: Tile-based Visualization of Diversified Web Search Results

Mayu Iwata[*]
Microsoft Research Asia, China
Osaka University, Japan
iwata.mayu@ist.osaka-u.ac.jp

Tetsuya Sakai
Microsoft Research Asia, China
tetsuyasakai@acm.org

Takehiro Yamamoto
Microsoft Research Asia, China
Kyoto University, Japan
tyamamot@dl.kuis.kyoto-u.ac.jp

Yu Chen
Microsoft Research Asia, China
Yu.Chen@microsoft.com

Yi Liu
Microsoft Research Asia, China
lewisliu@microsoft.com

Ji-Rong Wen
Microsoft Research Asia, China
jrwen@microsoft.com

ABSTRACT

A diversified search result for an underspecified query generally contains web pages in which there are answers that are relevant to different aspects of the query. In order to help the user locate such relevant answers, we propose a simple extension to the standard Search Engine Result Page (SERP) interface, called AspecTiles. In addition to presenting a ranked list of URLs with their titles and snippets, AspecTiles visualizes the relevance degree of a document to each aspect by means of colored squares ("tiles"). To compare AspecTiles with the standard SERP interface in terms of usefulness, we conducted a user study involving 30 search tasks designed based on the TREC web diversity task topics as well as 32 participants. Our results show that AspecTiles has some advantages in terms of search performance, user behavior, and user satisfaction. First, AspecTiles enables the user to gather relevant information significantly more efficiently than the standard SERP interface for tasks where the user considers several different aspects of the query to be important at the same time (*multi-aspect tasks*). Second, AspecTiles affects the user's information seeking behavior: with this interface, we observed significantly fewer query reformulations, shorter queries and deeper examinations of ranked lists in multi-aspect tasks. Third, participants of our user study found AspecTiles significantly more useful for finding relevant information and easy to use than the standard SERP interface. These results suggest that simple interfaces like AspecTiles can enhance the search performance and search experience of the user when their queries are underspecified.

[*]This research was conducted while the first author was an intern at Microsoft Research Asia.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
Copyright 2012 ACM 978-1-4503-1472-5/12/08 ...$10.00.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]; H.5.2 [**Interfaces**]: [Evaluation/methodology]

General Terms

Experimentation, Human Factors

Keywords

Diversified Search, Search Result, Visualization, User Study

1. INTRODUCTION

Search queries are often *ambiguous* ("office" may mean a workplace or a Microsoft product) or *underspecified* ("harry potter" may mean the books, the films or the main character) [7]. Given such queries, *search result diversification* aims to cover different search *intents* of the query in a single ranked list of retrieved documents [1]. We are particularly interested in underspecified queries as every query is arguably underspecified to some extent unless it is a perfect representation of the underlying information need. One possible way to explicitly represent these different intents would be to use *information types*, which we refer to as *aspects*[1]. For example, desired aspects of an underspecified query "elliptical trainer" (a stationary exercise machine) may include *product names*, *benefits* and *reviews*. The present study concerns what kind of interface would be effective for presenting a diversified ranked list of results obtained in response to such a query.

An underspecified query may possibly represent an information need for a specific aspect of a topic or for various aspects of a topic. For example, with the aforementioned "elliptical trainer" query, the user may be looking for *product names* of the elliptical trainer, or he may be looking for not only product names, but also *benefits* and *reviews* of elliptical trainers. Thus, we make a distinction between *single-aspect tasks* and *multi-aspect tasks* in our study. This is because the effectiveness of a search result presentation interface may depend on the variety of aspects that the user desires. Another possible viewpoint for categorizing search tasks would be the number of relevant answers the user requires (e.g. home page findings vs. recall-oriented search) [3, 17], but this study primarily concerns

[1] This usage of "aspect" is different from TREC, where it was used synonymously with "instance." [11, 12]

recall-oriented search where several answers that match a certain aspect of the query need to be collected by the user.

To help the user efficiently and effectively locate retrieved relevant answers that match the desired aspects within the Search Engine Result Page (SERP), we propose a simple extension to the standard SERP interface, which we call *AspecTiles*. Figure 1 shows a screenshot of the AspecTiles interface. As can be seen, AspecTiles visualizes which of the ranked web pages are relevant to which aspects by means of colored squares ("tiles") that are shown to the left of each document in the SERP. Moreover, the color represents the degree of relevance to an aspect. We believe that this line of research is important but missing in current search result diversification studies: even though most diversification systems rely on explicit identification of different search intents, a standard SERP interface does not tell the user which document is likely to be relevant to which of the intents, and to what degree. We hypothesized that a more informative interface such as AspecTiles will help web search engine users.

In this study, we conduct a user study involving 30 search tasks designed based on the TREC web diversity task topics as well as 32 participants to address the following Research Questions:

- RQ1: Does AspecTiles enable users to gather relevant information more effectively and efficiently when compared to the standard SERP interface?
- RQ2: Does AspecTiles affect the user's information seeking behavior? In what way?

RQ1 is the central question addressed in this paper. Compared to the standard SERP interface, AspecTiles provides the additional information of which documents are relevant to which aspects and to what degree. Thus, we expect the user to quickly scan the ranked list and locate relevant documents that match the desired aspects. Moreover, we would like to clarify under what conditions AspecTiles can be beneficial. For example, it is possible that AspecTiles may be more useful for multi-aspect tasks than for single-aspect tasks, as it can highlight the difference between a document that covers multiple aspects and one that covers only one aspect. Moreover, other factors such as the number of tiles displayed, the quality of the aspects represented as tiles and the accuracy of the estimated degree of relevance represented by the tile color may affect the outcome.

While RQ1 concerns user performance, RQ2 concerns user behavior. For example, we can hypothesize that, if the displayed aspects match well with the user's desired aspects, then the user can locate desired information without reformulating his query many times. Another possible hypothesis would be that, since AspecTiles enables the user to skip documents that are seemingly irrelevant to the desired aspects while scanning the ranked list, it may guide the user to documents that lie deeper in the ranked list.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our experimental design for evaluating AspecTiles, and Section 4 presents our results. Finally, Section 5 concludes this paper and discusses future work.

2. RELATED WORK

Information visualization takes advantage of the user's visual information processing abilities by generating graphical representations of data or concepts for presenting textual information [21], and is known to be useful for quantitative comparisons [15]. Thus, many search result visualization methods have been proposed to provide the user with a quick overview of the retrieved documents. According to Turetken and Sharda [20], web documents can be visualized based on content (e.g., title and term frequency) [4, 9, 10, 14, 16, 23], connectivity (e.g., number of incoming and outgoing links), and others (e.g., metadata like document size and the web domain) [4, 10]. Our AspecTiles falls into the first category, as it visualizes the relevance of each retrieved web page with respect to each displayed aspect.

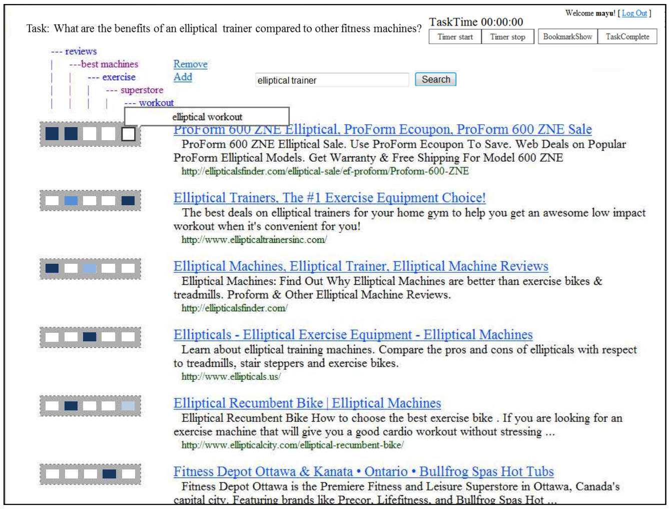
The TileBar interface proposed by Hearst [9] and the simpler HotMap interface proposed by Hoeber and Yang [14] are visually somewhat similar to AspecTiles. In TileBar, each retrieved document is represented by a rectangle, where the query terms are arranged vertically and the document length is represented horizontally. For each query term, its term frequency within each section of a document is represented in grayscale. In HotMap, a row of squares are shown to the left of each retrieved document, where each square represents a particular query term and its within-document term frequency. The user experiments by Hoeber and Yang suggested that HotMap is useful for effective and efficient information gathering. Our AspecTiles is different from TileBar and HotMap in that (a) the row of squares (which we call *tiles*[2]) represent possible aspects or intents of the query rather than the individual query terms of that query; and (b) the color of each tile represents the degree of relevance of the document to that particular aspect rather than term frequency. Thus, AspecTiles requires the backend search engine to provide a set of possible aspects given a query, and estimated per-aspect relevance scores. To our knowledge, our work is the first to study this kind of interface for the purpose of presenting diversified search results.

There are other ways to visualize search engine results, such as *glyph-based* and *graph-based* methods. Glyph-based approaches include the work by Heimonen and Jhaveri [10] who used document-shaped icons, and the work by Chau [4] who used a flower metaphor for representing statistics such as term frequency and document length. Graph-based approaches include bar charts of Reiterer *et al.* [16], and the radar charts of Yamamoto and Tanaka [23]. The radar charts were used for visualizing a fixed set of credibility criteria (accuracy, objectivity, authority, currency, and coverage). None of these studies was to do with multiple intents/aspects and search result diversification.

Brandt *et al.* [2] describe a method for dynamically presenting diversified search results. The main idea is to present a tree of web pages and dynamically recomputing it based on the nodes expanded by the user instead of showing a flat static list. While this rich presentation approach may have benefits, it is radically different from the standard SERP, and may be difficult to prevail. Moreover, its usability (e.g. the cognitive load of making the user backtrack within a tree, that of rearranging documents that the user has seen or has yet to see, etc.) is unknown. In contrast, AspecTiles lets the user go through a list just like traditional SERP, and is arguably easier for existing web search engines to adopt.

Our choice of using simple tiles (i.e. colored squares) for representing per-aspect relevance of retrieved documents is based on findings from previous research in user interface studies including the aforementioned ones. In particular, the user evaluation of HotMap [14] suggested that the simple interface offers effective and efficient information access as well as high user satisfaction. While HotMap users may quickly skip documents that are nonrelevant to the query and locate relevant documents with high query term frequencies, AspecTiles users may quickly skip documents that are nonrelevant to the *desired aspects of the query* and even locate relevant documents that cover multiple desired aspects. Fur-

[2] Note that Hearst's "tiles" refer to nonoverlapping text segments [9].



Task: What are the benefits of an elliptical trainer compared to other fitness machines?

ProForm 600 ZNE Elliptical, ProForm Ecoupon, ProForm 600 ZNE Sale
ProForm 600 ZNE Elliptical Sale. Use ProForm Ecoupon To Save. Web Deals on Popular ProForm Elliptical Models. Get Warranty & Free Shipping For Model 600 ZNE
http://ellipticalsfinder.com/elliptical-sale/ef-proform/Proform-600-ZNE

Elliptical Trainers, The #1 Exercise Equipment Choice!
The best deals on elliptical trainers for your home gym to help you get an awesome low impact workout when it's convenient for you!
http://www.ellipticaltrainersinc.com/

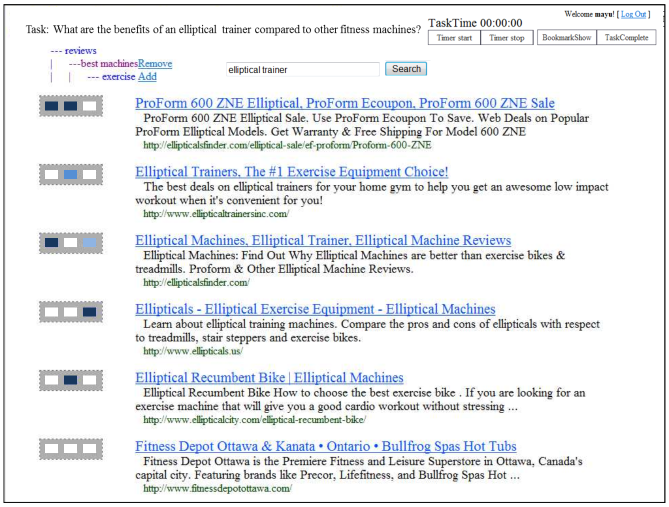
Elliptical Machines, Elliptical Trainer, Elliptical Machine Reviews
Elliptical Machines: Find Out Why Elliptical Machines are better than exercise bikes & treadmills. Proform & Other Elliptical Machine Reviews.
http://ellipticalsfinder.com/

Ellipticals - Elliptical Exercise Equipment - Elliptical Machines
Learn about elliptical training machines. Compare the pros and cons of ellipticals with respect to treadmills, stair steppers and exercise bikes.
http://www.ellipticals.us/

Elliptical Recumbent Bike | Elliptical Machines
Elliptical Recumbent Bike How to choose the best exercise bike . If you are looking for an exercise machine that will give you a good cardio workout without stressing ...
http://www.ellipticalcity.com/elliptical-recumbent-bike/

Fitness Depot Ottawa & Kanata • Ontario • Bullfrog Spas Hot Tubs
Fitness Depot Ottawa is the Premiere Fitness and Leisure Superstore in Ottawa, Canada's capital city. Featuring brands like Precor, Lifefitness, and Bullfrog Spas Hot ...

Figure 1: Screenshot of AspecTiles with five tiles (elliptical trainer).



Task: What are the benefits of an elliptical trainer compared to other fitness machines?

ProForm 600 ZNE Elliptical, ProForm Ecoupon, ProForm 600 ZNE Sale
ProForm 600 ZNE Elliptical Sale. Use ProForm Ecoupon To Save. Web Deals on Popular ProForm Elliptical Models. Get Warranty & Free Shipping For Model 600 ZNE
http://ellipticalsfinder.com/elliptical-sale/ef-proform/Proform-600-ZNE

Elliptical Trainers, The #1 Exercise Equipment Choice!
The best deals on elliptical trainers for your home gym to help you get an awesome low impact workout when it's convenient for you!
http://www.ellipticaltrainersinc.com/

Elliptical Machines, Elliptical Trainer, Elliptical Machine Reviews
Elliptical Machines: Find Out Why Elliptical Machines are better than exercise bikes & treadmills. Proform & Other Elliptical Machine Reviews.
http://ellipticalsfinder.com/

Ellipticals - Elliptical Exercise Equipment - Elliptical Machines
Learn about elliptical training machines. Compare the pros and cons of ellipticals with respect to treadmills, stair steppers and exercise bikes.
http://www.ellipticals.us/

Elliptical Recumbent Bike | Elliptical Machines
Elliptical Recumbent Bike How to choose the best exercise bike . If you are looking for an exercise machine that will give you a good cardio workout without stressing ...
http://www.ellipticalcity.com/elliptical-recumbent-bike/

Fitness Depot Ottawa & Kanata • Ontario • Bullfrog Spas Hot Tubs
Fitness Depot Ottawa is the Premiere Fitness and Leisure Superstore in Ottawa, Canada's capital city. Featuring brands like Precor, Lifefitness, and Bullfrog Spas Hot ...
http://www.fitnessdepotottawa.com/

Figure 2: Screenshot of AspecTiles with three tiles (elliptical trainer).

thermore, Xiang *et al.* [22] compared a tree view and a list view to visualize criminal data relations, and reported that users may find it difficult to use unfamiliar interfaces like the tree view. Chen and Yu [5] conducted a meta-analysis of six information visualization usability studies for IR and also reported that users perform better with simple visual interfaces than with complex ones.

Regarding the underlying search engine that SERP interfaces require, a number of effective search result diversification algorithms have been proposed recently [1, 8, 18]. Any of these algorithms could be used with our AspecTiles presentation interface, provided that an explicit set of aspects of the given query and per-aspect relevance scores for each retrieved document are available. In this study, we use as input the diversified web search system described by Dou *et al.* [8]. This system was a top performer in the Japanese Document Ranking (search result diversification) subtask of the recent NTCIR-9 INTENT task [19].

3. EXPERIMENTAL DESIGN

This section describes a user study that we designed to address the research questions mentioned in Section 1: Is AspecTiles more effective and efficient than the standard SERP? How does it affect the user's search behavior?

3.1 Interfaces

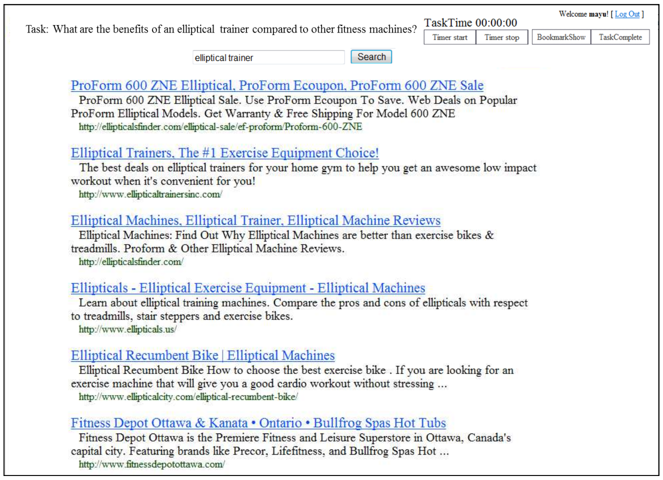
We compared three prototype interfaces in our user experiments: AspecTiles with five tiles, AspecTiles with three tiles and a baseline interface with no tiles. The number of tiles refers to the *initial* setting – participants were allowed to change the number of displayed tiles during their search tasks, as we shall explain later. We initially displayed either five or three tiles because (a) too many tiles would mean a wide left margin for the SERP and therefore waste of space; and (b) showing too much information may have adverse effects on the user performance and satisfaction. Note, for example, that current web search engines typically show no more than eight query suggestions at a time.

Figures 1, 2 and 3 show screenshots of our three interfaces for the query "elliptical trainer." All interfaces show the title, snippet and URL of each ranked document. In addition, the two AspecTiles interfaces show (a) *aspect labels* (e.g. "reviews", "best machines", "exercise" etc.) at the top of SERP; (b) a colored tile for each displayed aspect, whose color represents the per-aspect relevance of the document in five-point scale (dark blue means highly relevant and light blue means marginally relevant); and (c) a popup text when the mouse is placed over a tile, which reminds the user of the aspect label. In addition, the user can freely adjust the number of displayed aspects to any number between zero and eight during the experiment, by clicking the Add or the Remove button. For example, if a user of the three-tile AspecTiles interface finds that none of the displayed aspects is useful, he may choose to click Add to show more aspects, some of which may turn out be relevant to his needs. Whereas, the baseline interface is very similar to popular search engines such as Google and Bing.

The three experimental interfaces share the same underlying diversification algorithm. When a query is issued, its top 300 search results and *query suggestions* are obtained using Bing API[3]. In our study, we treated these query suggestions as the aspects to be displayed on the interface, as they are easy to obtain, reasonably accurate and were effective for search result diversification with Dou *et al.*'s algorithm [8] at the NTCIR-9 INTENT task [19]. To save space, the aforementioned aspect labels were obtained after removing all query terms from each query suggestion; whereas, the original query suggestion string was shown in the popup text (e.g. "reviews" for the former and "elliptical trainer reviews" for the latter). A more sophisticated method that is dedicated to finding accurate aspects (information types) for AspecTiles deserves to be explored, but this is beyond the scope of our study: we focus on the usefulness of AspecTiles in a realistic situation where the displayed aspects do not perfectly overlap with the desired aspects.

Using the aspects thus obtained, we used the algorithm by Dou *et al.* to compute per-aspect relevance scores and to rerank the SERP obtained by Bing API. Our future plan includes letting the user click the tiles (as well as the aspect labels) on the AspecTiles interface and thereby dynamically rerank documents, but this is beyond the scope of the present study. Our current focus is to examine the effect of *showing* the tiles to the user rather than that of interacting with the tiles as well as with the aspect labels (query suggestions).

[3] http://www.bing.com/toolbox/bingdeveloper/



Task: What are the benefits of an elliptical trainer compared to other fitness machines?

elliptical trainer

ProForm 600 ZNE Elliptical, ProForm Ecoupon, ProForm 600 ZNE Sale
ProForm 600 ZNE Elliptical Sale. Use ProForm Ecoupon To Save. Web Deals on Popular ProForm Elliptical Models. Get Warranty & Free Shipping For Model 600 ZNE
http://ellipticalsfinder.com/elliptical-sale/ef-proform/Proform-600-ZNE

Elliptical Trainers, The #1 Exercise Equipment Choice!
The best deals on elliptical trainers for your home gym to help you get an awesome low impact workout when it's convenient for you!
http://www.ellipticaltrainersinc.com/

Elliptical Machines, Elliptical Trainer, Elliptical Machine Reviews
Elliptical Machines: Find Out Why Elliptical Machines are better than exercise bikes & treadmills. Proform & Other Elliptical Machine Reviews.
http://ellipticalsfinder.com/

Ellipticals - Elliptical Exercise Equipment - Elliptical Machines
Learn about elliptical training machines. Compare the pros and cons of ellipticals with respect to treadmills, stair steppers and exercise bikes.
http://www.ellipticals.us/

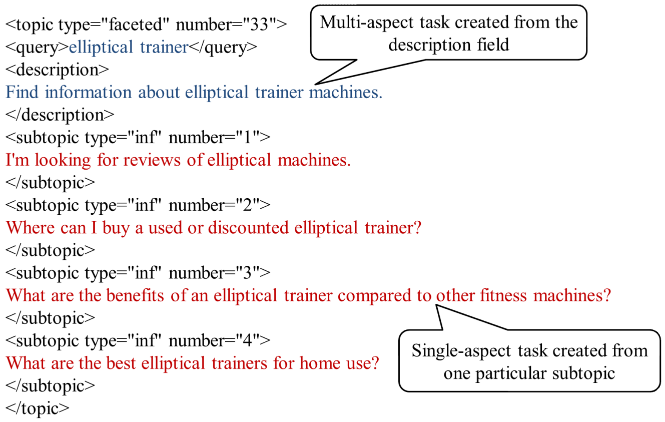
Elliptical Recumbent Bike | Elliptical Machines
Elliptical Recumbent Bike How to choose the best exercise bike . If you are looking for an exercise machine that will give you a good cardio workout without stressing ...
http://www.ellipticalcity.com/elliptical-recumbent-bike/

Fitness Depot Ottawa & Kanata • Ontario • Bullfrog Spas Hot Tubs
Fitness Depot Ottawa is the Premiere Fitness and Leisure Superstore in Ottawa, Canada's capital city. Featuring brands like Precor, Lifefitness, and Bullfrog Spas Hot ...
http://www.fitnessdepotottawa.com/

Figure 3: Screenshot of the baseline interface (elliptical trainer).

3.2 Tasks

As we hypothesized that AspecTiles may be more useful for multi-aspect tasks (e.g. Find *product names*, *benefits* and *reviews* of the elliptical trainer) than for single-aspect tasks (e.g. What are the *benefits* of an elliptical trainer?), we devised 15 multi-aspect tasks and 15 single-aspect tasks shown in Table 1. For this purpose, we carefully examined the TREC 2009 web diversity task topics and selected 15 *faceted* (i.e. underspecified) topics that had *informational* subtopics [6], so that each selected topic could be used as a resource for creating exactly one multi-aspect task and exactly one single-aspect task [4]. Figure 4 shows a selected topic: it contains a query field, a description field and several informational subtopics. From this topic, we created a multi-aspect task: "Find information about elliptical trainer machines. (reviews, benefits, product names)." The words in parenthesis correspond to the *desired aspects*, i.e., the types of information that the participants are asked to collect. In this example, these desired aspects correspond to informational subtopics 1, 3 and 4: for every multi-aspect task, we selected exactly three desired aspects. When more than three informational subtopics were available, we carefully selected three desired aspects from them so that the overlap between the selected aspects and the *displayed* aspects (i.e. query suggestions) are neither too high nor too low. As we mentioned earlier, we wanted to examine the realistic situation where the displayed aspects are of *reasonable* accuracy. In fact, the average number of informational subtopics per faceted topic was 3.8, so we used up most of the informational subtopics in the faceted topics. On the other hand, from the same topic, we created one single-aspect task: "What are the benefits of an elliptical trainer compared to other fitness machines?" Note that this is exactly subtopic 3. In summary, we prepared 15 multi-aspect tasks that require participants to look for exactly three aspects (e.g. reviews, benefits and product names), and 15 single-aspect tasks that require them to look for exactly one aspect (e.g. benefits).



```
<topic type="faceted" number="33">
<query>elliptical trainer</query>
<description>
Find information about elliptical trainer machines.
</description>
<subtopic type="inf" number="1">
I'm looking for reviews of elliptical machines.
</subtopic>
<subtopic type="inf" number="2">
Where can I buy a used or discounted elliptical trainer?
</subtopic>
<subtopic type="inf" number="3">
What are the benefits of an elliptical trainer compared to other fitness machines?
</subtopic>
<subtopic type="inf" number="4">
What are the best elliptical trainers for home use?
</subtopic>
</topic>
```

Multi-aspect task created from the description field

Single-aspect task created from one particular subtopic

Figure 4: A sample topic from the TREC 2009 web diversity task.

3.3 Procedure

We hired 32 participants for our user study. 21 of them were male and 11 were female; 24 were students in computer science; four were researchers in computer science and the other four were desk workers. The average age of the participants was 26.18 years, with a standard deviation of 4.34.

Then we formed an experimental design similar to that of the TREC interactive track [13]. Each participant performed a total of six tasks (three multi-aspect and three single-aspect): two using the five-tile AspecTiles, two using the three-tile AspecTiles and the other two using the baseline. The choice and the order of systems and tasks were randomized for each participant.

Prior to the actual user experiment, each participant went through a training task to become familiar with the interfaces and the tasks. Each training session took approximately ten minutes.

For each of the six tasks, the participants were provided with a printed task description document that contained the initial query (e.g. "elliptical trainer"), a short explanation of the topic (e.g. "a stationary exercise machine"), (for multi-aspect tasks only) the three desired aspects (e.g. "reviews," "benefits" and "product names.") and some additional information if required. For the elliptical trainer example, the short explanation indicated that it is a kind of stationary exercise machine (description obtained from Wikipedia) and a sample picture of an elliptical trainer was shown to the participants as additional information. This is because the tasks that originate from TREC are not real information needs of the participants, and therefore the participants were not necessarily familiar with all of the tasks.

In each task session, the participant used one of the aforementioned three interfaces to collect as many *answers* as possible within ten minutes (for single-aspect tasks) or fifteen minutes (for multi-aspect tasks). An answer is a relevant instance that matches one of the desired aspects (e.g. a particular product name of elliptical trainer). When the participant clicked on a URL within the interface, a special landing page window popped up, from which he could extract arbitrary part of the HTML by a mouse drag and *bookmark* it. A single bookmark can contain multiple text segments from the same webpage, as well as images and links. A single *relevant* bookmark may contain one or multiple answers, and may even cover multiple desired aspects. For example, one bookmark may contain a product name of elliptical trainer, its review as well as its picture. Note that we used a special browsing interface to facilitate collection of answers rather than a standard web browser.

[4] TREC diversity topics also have *navigational* subtopics, which we did not utilize as our focus was on collecting multiple answers that are relevant to each aspect.

Table 1: Task list.

Topic	S: single-aspect task, M: multi-aspect task (three desired aspects)
elliptical trainer	S: What are the benefits of an elliptical trainer compared to other fitness machines? M: Find information about elliptical trainer machines. (reviews, benefits, product names)
disneyland hotel	S: What hotels are near Disneyland? M: Find information about hotels near Disneyland Resort in California. (packages, reviews, coupons)
getting organized	S: Find catalogs of office supplies for organization and decluttering. M: Find tips, resources, supplies for getting organized and reducing clutter. (tips, resources, supplies)
dinosaurs	S: I'm looking for free pictures of dinosaurs. M: I want to find information about and pictures of dinosaurs. (pictures, names, games)
poker tournaments	S: Find books on tournament poker playing. M: I want to find information about live and online poker tournaments. (schedules, game names, books)
volvo	S: Where can I find Volvo semi trucks for sale (new or used)? M: I'm looking for information on Volvo cars and trucks. (reviews, dealers, product names)
rick warren	S: I want to see articles and web pages about the controversy over Rick Warren's invocation at the Obama inauguration. M: I'm looking for information on Rick Warren, the evangelical minister. (biography, articles, books)
diversity	S: What is cultural diversity? What is prejudice? M: Find information on diversity, both culturally and in the workplace. (management methods, definition, poems)
starbucks	S: Find the menu from Starbucks, with prices. M: Find information about the coffee company Starbucks. (menu, calories, coupons)
diabetes education	S: Find free diabetes education materials such as videos, pamphlets, and books. M: I'm looking for online resources to learn and teach others about diabetes. (materials, diet, classes)
atari	S: I want to read about the history of the Atari 2600 and other Atari game consoles. M: Find information about Atari, its game consoles and games. (history, classic games, arcade games)
cell phones	S: What cell phone companies offer Motorola phones? M: Find information about cell phones and cellular service providers. (provider names, prices, reviews)
hoboken	S: Find restaurants in Hoboken. M: Find information on the city of Hoboken, New Jersey. (restaurants, history, real estate)
orange county convention center	S: What hotels are near the Orange County Convention Center? M: Looking for information about the Orange County Convention Center in Orlando, Florida. (event schedules, hotels, restaurants)
the secret garden	S: Find reviews of the various TV and movie adaptations of The Secret Garden. M: Find information and reviews about The Secret Garden book by Frances Hodgson Burnett as well as movies and the musical based on it. (reviews, summary, music)

But this condition was the same across the three interfaces that we wanted to compare.

The query field of the original TREC topic served as the initial query for each task, so that each participant began his task by looking at a diversified search result. However, participants were also allowed to freely reformulate the query. As was mentioned earlier, they were allowed to change the number of displayed aspects to an arbitrary number between zero and eight in the AspecTiles interfaces during the experiment.

Participants were also allowed to finish their task sessions prior to the time limit if they felt that their collected answers were complete. After each task session, they were asked to fill out a post-task questionnaire. After completing all of the six task sessions, they were asked to fill out an exit questionnaire. The entire procedure for one participant took approximately one hour.

4. RESULTS

Section 4.1 discusses the results for RQ1 (search performance); 4.2 discusses the results for RQ2 (search behavior); and 4.3 discusses the questionnaire results.

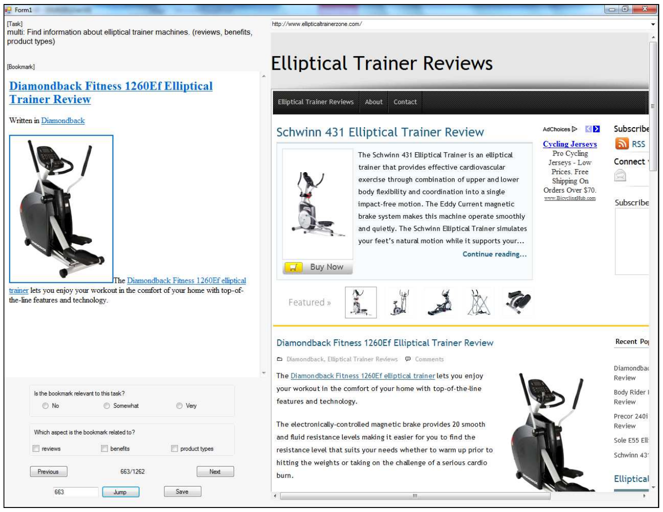
4.1 Search performance

4.1.1 *Bookmark relevance assessment*

Through 192 task sessions (32 participants, each with 6 tasks), a total of 1,262 bookmarks were obtained. In order to evaluate user performance for these task sessions, three annotators independently assessed all of these bookmarks. Figure 5 is a screenshot of the assessment tool we developed for that purpose: the left panel shows a particular bookmark (a part of an HTML file) to be assessed with radio buttons for selecting a relevance grade ("highly relevant," "somewhat relevant" or "not relevant"). For multi-aspect tasks, if a bookmark was judged relevant, then the assessors were also asked to check on one or more of the desired aspects that the bookmark contained. The inter-assessor agreement was quite high (Fleiss' kappa: 0.666). Based on the three assessment sets, we constructed two ground-truths: *strict* and *lenient* sets. In the former, the relevance grade of each bookmark reflects the lowest relevance grade for that bookmark among the three assessors, and the matched aspects of that bookmark are obtained by taking the intersection of the three assessors' selected aspects. In the latter, the relevance grade of each bookmark reflects the highest relevance grade for that bookmark among the three assessors, and the matched aspects of that bookmark are obtained by taking the union of the three assessors' selected aspects. Thus, the strict set reflects the view: "nothing is relevant unless all three assessors say that it is relevant" while the lenient set reflects the view: "anything is relevant if at least one assessor says that it is relevant."

4.1.2 *Effectiveness and Efficiency*

Using the strict and lenient sets of ground truths, we compared the user effectiveness and efficiency for AspecTiles and for the baseline interface. For each task session, let n be the number of relevant bookmarks found by a particular user. Then precision is n divided by the number of bookmarks found by that user for that task session, and recall is n divided by the total number of known relevant bookmarks for that task session. As measures of search efficiency, we also measured the time taken to obtain the *first relevant* bookmark ("Time-FirstRel"), the number of document clicks before the first relevant bookmark was obtained



[Task]
multi: Find information about elliptical trainer machines. (reviews, benefits, product types)

[Bookmark]

Diamondback Fitness 1260Ef Elliptical Trainer Review

Written in Diamondback

The Diamondback Fitness 1260Ef elliptical trainer lets you enjoy your workout in the comfort of your home with top-of-the-line features and technology.

Is the bookmark relevant to this task?
○ No ○ Somewhat ○ Very

Which aspect is the bookmark related to?
□ reviews □ benefits □ product types

663/1262

663

http://www.ellipticaltrainerzone.com/

Elliptical Trainer Reviews

Schwinn 431 Elliptical Trainer Review

The Schwinn 431 Elliptical Trainer is an elliptical trainer that provides effective cardiovascular exercise through combination of upper and lower body flexibility and coordination into a single impact-free motion. The Eddy Current magnetic brake system makes this machine operate smoothly and quietly. The Schwinn Elliptical Trainer simulates your feet's natural motion while it supports your...

Cycling Jerseys
Pro Cycling Jerseys - Low Prices. Free Shipping On Orders Over $70.
www.BicyclingHub.com

Diamondback Fitness 1260Ef Elliptical Trainer Review

Diamondback, Elliptical Trainer Reviews Comments

The Diamondback Fitness 1260Ef elliptical trainer lets you enjoy your workout in the comfort of your home with top-of-the-line features and technology.

The electronically-controlled magnetic brake provides 20 smooth and fluid resistance levels making it easier for you to find the resistance level that suits your needs whether to warm up prior to hitting the weights or taking on the challenge of a serious cardio burn.

Figure 5: Screenshot of the assessment tool. The left panel shows a bookmark (containing a picture, a product name and its review) and the right panel shows the entire page from which the bookmark was extracted.

("DocClick-FirstRel"), the average time between obtaining a relevant bookmark and obtaining the next relevant bookmark ("Time-TwoRel"), and the average number of document clicks between obtaining a relevant bookmark and obtaining the next relevant bookmark ("DocClick-TwoRel"). Clearly, all of these statistics should be low for an ideally efficient interface. For multi-aspect tasks that require three aspects, we also compute aspect coverage, i.e. the fraction of desired aspects covered by the user's bookmarks (which can range from $0/3$ to $3/3$).

Tables 2 and 3 show the user effectiveness and efficiency results with the strict and lenient sets, respectively. For both task types (single-aspect and multi-aspect), we conducted a two-factor (interface and task) ANOVA at $p \leq 0.05$ for all types of measurements. For single-aspect tasks, only the task effect was significant for Precision, Time-FirstRel, DocClick-FirstRel, Time-TwoRel, and DocClick-TwoRel; none of the effects was significant for Recall. For multi-aspect tasks, only the task effect was significant for Precision and Aspect coverage, and only the system effect was significant for Recall, Time-FirstRel, DocClick-FirstRel, Time-TwoRel, and DocClick-TwoRel. If a significant interaction was found, we also conducted a Bonferroni post-hoc test at $p \leq 0.05$ between two interfaces. In the tables, statistically significant differences with the baseline are shown in bold.

In multi-aspect tasks, it can be observed that the five-tile AspecTiles demonstrated significantly higher recall than the baseline. Moreover, although not statistically significant, it can be observed that Precision and Aspect Coverage values also tend to be higher with AspecTiles. On the other hand, it can be observed that AspecTiles is significantly better than the baseline in terms of Time-FirstRel, DocClick-FirstRel, Time-TwoRel and DocClick-TwoRel. The latter two statistics in particular show that even the three-tile AspecTiles is advantageous over the baseline. Thus, we can conclude that, for multi-aspect tasks, AspecTiles lets the user collect relevant pieces of information in shorter time using fewer clicks than the baseline, while keeping the user effectiveness at least comparable. In addition, note that, for multi-aspect tasks, the measurements for five-tile AspecTiles are consistently better than those for three-tile AspecTiles. Although the differences are not statistically significant in our experiment, this suggests that displaying tiles is beneficial.

Table 2: Effectiveness and efficiency results (strict). The average of all 32 task sessions is reported, with the standard deviation in parenthesis. Significant differences with the baseline at $p \leq 0.05$ according to the Bonferroni post-hoc test are shown in bold.

Task type		Baseline	Three-tile AspecTiles	Five-tile AspecTiles
Single	Precision	0.498 (0.387)	0.496 (0.409)	0.469 (0.395)
	Recall	0.143 (0.127)	0.154 (0.193)	0.152 (0.145)
	Time-FirstRel	187.5 (96.8)	193.4 (108.4)	183.8 (100.9)
	DocClick-FirstRel	2.500 (2.502)	1.655 (1.738)	2.258 (2.065)
	Time-TwoRel	99.8 (63.6)	89.3 (70.1)	113.4 (76.9)
	DocClick-TwoRel	2.631 (1.348)	2.234 (1.061)	2.997 (1.742)
Multi	Precision	0.684 (0.248)	0.685 (0.252)	0.717 (0.232)
	Recall	0.130 (0.088)	0.184 (0.178)	**0.214** (0.204)
	Aspect Coverage	0.644 (0.315)	0.741 (0.233)	0.759 (0.250)
	Time-FirstRel	166.9 (117.0)	130.3 (92.7)	**118.0** (120.6)
	DocClick-FirstRel	3.345 (3.199)	1.963 (1.055)	**1.714** (0.897)
	Time-TwoRel	115.5 (77.1)	**82.4** (77.0)	**69.3** (48.4)
	DocClick-TwoRel	2.808 (1.607)	**2.087** (1.270)	**1.997** (0.598)

For single-aspect tasks, we cannot claim that AspecTiles is effective as there are no significant differences observed. We attribute this mainly to the fact that for single-aspect tasks, the chance of finding the desired aspect within the displayed aspects is smaller compared to multi-aspect tasks. For example, for both the single-aspect task "What cell phone companies offer Motorola phones?" and the multi-aspect task "Find information about cell phones and cellular service providers. (provider names, prices, reviews)," the initial query automatically issued by the system was "cell phones." As a result, in both tasks, the initially displayed aspects (obtained from query suggestions in response to "cell phones") were "providers, verizon, sprint" in three-tile AspecTiles, and additionally "prepaid, cheap" in five-tile AspecTiles. While the displayed aspects such as "providers" and "cheap" are probably directly useful for the multi-aspect task as it requires "provider names" and "prices," they are probably not useful for the particular single-aspect task as "motorola" is not included in them. We manually counted the number of initially displayed aspects that are actually relevant to the desired aspects for all task sessions: it was 0.683 for single-aspect tasks and 1.113 for multi-aspect tasks on average. The next section investigates this effect of quality of the displayed aspects in more detail.

Table 3: Effectiveness and efficiency results (lenient). The average of all 32 task sessions is reported, with the standard deviation in parenthesis. Significant differences with the baseline at $p \leq 0.05$ according to the Bonferroni post-hoc test are shown in bold.

Task type		Baseline	Three-tile AspecTiles	Five-tile AspecTiles
Single	Precision	0.615 (0.425)	0.581 (0.441)	0.649 (0.383)
	Recall	0.136 (0.107)	0.148 (0.151)	0.185 (0.149)
	Time -FirstRel	176.2 (100.0)	183.9 (105.3)	155.8 (87.7)
	DocClick -FirstRel	2.188 (2.086)	2.207 (2.555)	1.871 (1.408)
	Time -TwoRel	94.1 (65.4)	107.9 (91.2)	90.8 (54.1)
	DocClick -TwoRel	2.461 (1.354)	2.758 (2.328)	2.285 (1.170)
Multi	Precision	0.772 (0.245)	0.802 (0.285)	0.849 (0.221)
	Recall	0.127 (0.081)	0.173 (0.103)	**0.213** (0.194)
	Aspect Coverage	0.733 (0.296)	0.828 (0.262)	0.828 (0.246)
	Time -FirstRel	165.3 (117.4)	124.3 (87.8)	**114.0** (118.0)
	DocClick -FirstRel	3.333 (3.188)	1.926 (1.072)	**1.643** (0.870)
	Time -TwoRel	110.6 (76.0)	**72.9** (62.5)	**66.6** (47.3)
	DocClick -TwoRel	2.749 (1.615)	**1.987** (1.211)	**1.920** (0.564)

4.1.3 *Effect of Quality of Displayed Aspects and Relevance Estimation Accuracy*

AspecTiles requires as input (a) a set of aspects to be displayed in response to a query; and (b) per-aspect relevance score for each retrieved document. In this study, we chose to use query suggestions for (a), and a state-of-the-art diversified web search system [8] for (b), so that we can study the advantages of AspecTiles in a practical situation. Thus, in our study, both the quality of the displayed aspects and the estimated per-aspect relevance score (shown as colors of the tiles) are far from perfect. In this section, we investigate their effects on user effectiveness.

To investigate the effect of displayed aspect quality on user effectiveness, we computed precision, recall and (for multi-aspect tasks only) aspect coverage of the bookmarks at the *query level* rather than the task session level as was done in Section 4.1.2. This is because participants can issue multiple queries during a task session, and the displayed aspects (and therefore their quality) change accordingly within the task session. For each issued query, let n be the number of relevant bookmarks found by a particular user for that query. Then *per-query* precision is n divided by the number of bookmarks found by that user for that query, *per-query* recall is n divided by the total number of known relevant bookmarks for that query. We also compute *per-query* aspect coverage, i.e. the fraction of desired aspects covered by the user's bookmarks for that query. The quality of each set of displayed aspects was computed automatically: if a displayed aspect had at least one overlapping word with the task sentence, it was regarded as relevant; the quality of the displayed aspects set was defined as the fraction of such relevant displayed aspects.

Table 4: Spearman's rank correlation between user effectiveness (strict) and the quality of displayed aspects / relevance estimation accuracy.

Task type		Displayed aspect quality	Relevance estimation accuracy
Single	Precision	0.334	0.496
	Recall	0.300	0.490
Multi	Precision	0.434	0.506
	Recall	0.324	0.516
	Aspect Coverage	0.369	0.506

On the other hand, to investigate the effect of relevance estimation accuracy, we computed precision, recall and (for multi-aspect tasks only) aspect coverage of the bookmarks *for every clicked document*. This is because the quality of estimated per-aspect relevance score for each clicked document change accordingly within the task session. For each clicked document, let n be the number of relevant bookmarks found by a particular user on that document. Then *per-document* precision is n divided by the number of bookmarks found by that user within a particular document, *per-document* recall is n divided by the total number of known relevant bookmarks within that document. We also compute *per-document* aspect coverage, i.e. the fraction of desired aspects covered by the user's bookmarks within that document. For example, if two bookmarks were saved from a clicked document and only one of them was relevant, the precision for this document is 0.5. For computing these metrics, the first author of this paper manually examined all 1,483 clicked documents obtained in our experiments, and identified *false alarms*: a false alarm is a document-aspect pair that AspectTiles presented as relevant even though in fact the document content suggested otherwise. (Whereas, she did not check for *misses*: relevant document-aspect pairs which AspecTiles failed to present as relevant.)

Tables 4 and 5 show the Spearman's rank correlation values between user effectiveness and the aspect quality / relevance estimation accuracy with the strict and lenient sets, respectively. Recall that the user effectiveness values were computed per query for the displayed aspect quality and per document for the relevance estimation accuracy. It can be observed that user effectiveness (precision, recall and aspect coverage) is indeed correlated with the quality of displayed aspects and with the accuracy of estimated per-aspect relevance scores. The correlation is especially high for the latter. These results suggest that AspecTiles would be even more effective if the underlying search system can provide more appropriate selection of aspects to be displayed, as well as accurate per-aspect relevance scores. Note that showing innacurate relevance scores (in the form of colors) may seriously mislead the user: if the system fails to indicate that a document is relevant from a certain aspect, the user may skip the document; conversely, if the system shows a false positive for a certain aspect, the user may be forced to visit a document that is not useful to him.

4.2 Search Behavior

Now we discuss RQ2: how does AspecTiles affect the user's information seeking behavior? For this purpose, we recorded the participants's every operation performed on the three interfaces during the aforementioned experiments.

Table 5: Spearman's rank correlation between user effectiveness (lenient) and the quality of displayed aspects / relevance estimation accuracy.

Task type		Displayed aspect quality	Relevance estimation accuracy
Single	Precision	0.320	0.522
	Recall	0.287	0.506
Multi	Precision	0.411	0.555
	Recall	0.392	0.542
	Aspect Coverage	0.327	0.523

4.2.1 Search Behavior Measurements

Table 6 summarizes the number of query reformulations, the average query lengths (including the initial query and the queries reformulated by the participants), the average rank of clicked documents ("Rank-All"), the average rank of clicked documents that contain at least one relevant bookmark in the strict set ("Rank-Rel (strict)"), and the average rank of clicked documents that contain at least one relevant bookmark in the lenient set ("Rank-Rel (lenient)"). As before, for both task types (single-aspect and multi-aspect), we conducted a two-factor (interface and task) ANOVA at $p \leq 0.05$ for all types of measurements. For multi-aspect tasks, the interface effect was significant for all measures; for single-aspect tasks, none of the effects was significant. We thus conducted a Bonferroni post-hoc test at $p \leq 0.05$ between two interfaces for multi-aspect tasks; significant differences with the baseline are shown in bold. It can be observed that, when the participants used AspecTiles for multi-aspect tasks, they performed significantly fewer query reformulations, issued shorter queries, and clicked documents that are deeper in the ranked list, compared to the baseline interface. Though not statistically significant, similar trends can be observed for single-aspect tasks, except for the query length. We also examined other measurements such as the number of clicked documents and the time to complete the task session, but did not find any clear trends.

The above significant differences between AspecTiles and the baseline interface are quite intuitive. With AspecTiles, users perform fewer query reformulations probably because the displayed aspects serve as an alternative to forming new queries. For example, if the displayed aspects represent more specific information needs than the current query, then the user may not have to add a new query term to the current query to explicitly specialize his need. This is also reflected in the significant differences in the average query length. Moreover, with AspecTiles, users dig deeper in the ranked lists, probably because AspecTiles lets the user quickly skip documents that appear to be irrelevant to the desired aspects. That is, provided that the choice of displayed aspects and the per-aspect relevance shown in colors are both sufficiently accurate, AspecTiles may be able to let the user reach the desired documents efficiently. Moreover, note that five-tile AspecTiles outperforms three-tile AspecTiles in all of these measurements, though the differences are not statistically significant.

4.2.2 Search Behavior Sample

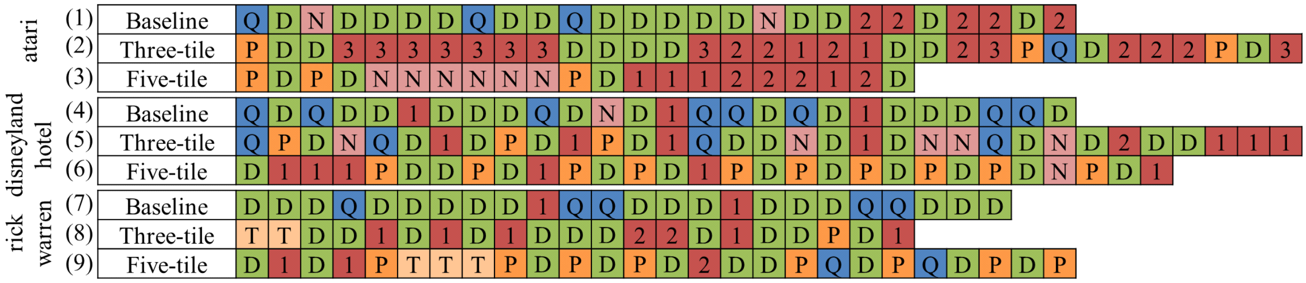
Section 4.2.1 discussed the summary statistics that represent user behavior. We now discuss the actual operation sequences of the participants. Figure 6 shows some typical user operations hand-picked from our multi-aspect tasks; similar data for single-aspect tasks are omitted due to lack of space. Each box represents a particular user operation on the interface, and a row of boxes represent a particular participant's task session over a timeline. The operations shown are: "select a highly/somewhat relevant bookmark of the lenient set" (number of aspects covered by the bookmark is shown); "select a Nonrelevant answer of the lenient set" (indicated by **N**)); "reformulating a Query" (**Q**); "click a Document" (**D**); "put mouse over a tile to show the Popup label" (**P**); and "increase number of Tiles" (**T**).

Table 6: Number of query reformulations, query length, and the average rank of clicked documents. The average and interquartile range values are reported. Significant differences with the baseline at $p \leq 0.05$ according to the Bonferroni post-hoc test are shown in bold.

Task type		Baseline	Three-tile AspecTiles	Five-tile AspecTiles
Single	Query reform	2.313 (0-8)	1.724 (0-5)	1.548 (0-5)
	Query length	2.585 (1-7)	2.532 (1-5)	2.646 (1-8)
	Rank-All	7.502 (1-46)	8.899 (1-49)	8.285 (1-47)
	Rank-Rel (strict)	6.968 (1-27)	10.025 (1-39)	8.511 (1-47)
	Rank-Rel (lenient)	7.132 (1-27)	9.929 (1-39)	9.155 (1-47)
Multi	Query reform	3.900 (0-13)	2.370 (0-10)	**1.552** (0-9)
	Query length	2.741 (1-7)	2.347 (1-5)	**2.189** (1-5)
	Rank-All	7.229 (1-70)	**10.785** (1-81)	**13.152** (1-95)
	Rank-Rel (strict)	7.294 (1-40)	**12.426** (1-61)	**13.725** (1-86)
	Rank-Rel (lenient)	7.282 (1-40)	**12.108** (1-61)	**13.347** (1-86)

Figure 6 confirms some of our aforementioned findings more qualitatively, and offers some additional insights into user behavior. For example, by comparing the rows of five-tile AspecTiles with those of the baseline, it can be observed not only that the latter contains more query reformulation operations (**Q**'s), but that participants often reformulated queries at the very beginning of the search interactions with the baseline interface (See lines (1) and (4)). With the baseline interface, the number of sessions beginning with an immediate query reformulation was 17 for single-aspect tasks and 12 for multi-aspect tasks; in contrast, with five-tile AspecTiles, the same statistic was 8 for single-aspect tasks and 4 for multi-aspect tasks. Moreover, it can be observed that the number of clicked documents before finding the first relevant bookmark (**D**'s to the left of the first square with a number) tends to be smaller with AspecTiles, which is again in line with our quantitative results. Furthermore, it can be observed that, with AspecTiles, participants tend to find a larger number of relevant bookmarks and relevant aspects, although these trends were not statistically significant in our quantitative results.

Next, we discuss typical usage patterns of AspecTiles, which we have observed physically during the user experiments and also from the operation logs that we analyzed postmortem. An AspecTile user typically begins a session by checking the aspect labels of the initial search result. If the displayed aspects seem appropriate, he utilizes the tiles, i.e. place a mouse over a tile and see the aspect label as a popup text, and clicks the document. In Figure 6, this can



atari	(1)	Baseline	Q	D	N	D	D	D	D	Q	D	D	Q	D	D	D	D	D	N	D	D	2	2	D	2	2	D	2							
	(2)	Three-tile	P	D	D	3	3	3	3	3	3	3	D	D	D	D	3	2	2	1	2	1	D	D	2	3	P	Q	D	2	2	2	P	D	3
	(3)	Five-tile	P	D	P	D	N	N	N	N	N	N	P	D	1	1	1	2	2	2	1	2	D												
disneyland hotel	(4)	Baseline	Q	D	Q	D	D	1	D	D	D	Q	D	N	D	1	Q	Q	D	Q	D	1	D	D	D	Q	Q	D							
	(5)	Three-tile	Q	P	D	N	Q	D	1	D	P	D	1	P	D	1	Q	D	D	N	D	1	D	N	N	Q	D	N	D	2	D	D	1	1	1
	(6)	Five-tile	D	1	1	1	P	D	D	P	D	1	P	D	P	D	1	P	D	P	D	P	D	P	D	P	D	N	P	D	1				
rick warren	(7)	Baseline	D	D	D	Q	D	D	D	D	D	1	Q	Q	D	D	D	1	D	D	D	Q	Q	D	D	D									
	(8)	Three-tile	T	T	D	D	1	D	1	D	1	D	D	D	2	2	D	1	D	D	P	D	1												
	(9)	Five-tile	D	1	D	1	P	T	T	T	P	D	P	D	P	D	2	D	D	P	Q	D	P	Q	D	P	D	P							

Figure 6: The participants' operation sequences in nine task sessions. The operations shown are: "select a highly/somewhat relevant bookmark of the lenient set" (number of aspects covered by the bookmark is shown); "select a Nonrelevant answer of the lenient set" (indicated by N)); "reformulating a Query" (Q); "click a Document" (D); "put mouse over a tile to show the Popup label" P; "increase number of Tiles" T.

be seen as a repetition of **P**'s and **D**'s (See lines (2), (3) and (6)). On the other hand, if the displayed aspects look unsatisfactory, the user often increases tiles and reformulate queries, as represented by **T**'s and **Q**'s in the figure (See lines (5) and (8)). For example, line (8) represents a multi-aspect task for "I'm looking for information on Rick Warren, the evangelical minister. (biography, articles, books)," where the initially displayed aspect labels were "controversy, obama, sermons." As the user judged that these displayed aspects did not directly match the three desired aspects, he immediately increased the number of displayed aspects (two **T**'s at the beginning of the session). Again, this example suggests that the quality of the displayed aspects is important. Moreover, Figure 6 line (3) contains many nonrelevant bookmarks (**N**'s), and this was because the document clicked by the user had a high estimated relevance to the aspect "(atari) classic games" even though it in fact discussed many other types of games besides Atari. Thus the user was misled by the color displayed by AspecTiles. Again, this highlights the importance of providing accurate relevance per-aspect estimates to the user.

As an additional and inconclusive remark, it appeared that while participants with high computer skills (e.g. researchers) tended to actively reformulate queries without referring to the tiles, those with relatively lower computer skills tended to rely on the tiles. While it is possible that AspecTiles may be more helpful to the latter class of users, we will have to design a new experiment with a larger set of participants to pursue this new research question.

4.3 User Satisfaction

Finally, we discuss the results of the exit questionnaire, which we summarized in Table 7. The participant's answer to each question was on a five point scale: from 1 (strongly disagree) to 5 (strongly agree). The Mann-Whitney's U test at $p \leq 0.05$ was used to compare AspecTiles with the baseline. Note that only Question 6 was asked separately for three-tile and five-tile AspecTiles. It can be observed that the participants found AspecTiles useful for finding relevant information, easy to use and that they are willing to use it again, compared to the baseline (Q1-Q3). Moreover, participants seemed to prefer the five-tile version to the three-tile version (Q6), although the difference is not statistically significant. Recall that the user effectiveness was indeed generally higher with five-tile AspecTiles than with the three-tile version.

We also received some additional feedback from participants. Five participants remarked that they wanted to select or create their own aspects to display; other five remarked that reranking the search result by a selected aspect would be very useful. In summary, AspecTiles was generally popular with the participants, and some of them wanted even more advanced features in it.

For the results of the post-task questionnaire, we asked the participants to answer four questions: "Are you familiar with this topic?," "Was it easy to do the search on this topic?," "Are you satisfied with your search results?" and "Did you have enough time to do an effective search?," but we could not find any clear trends among the three interfaces.

5. CONCLUSIONS

In this study, we proposed a simple extension to the standard SERP interface for presenting a diversified search result for an underspecified query. AspecTiles shows a row of tiles to the left of each retrieved document, where each tile represents a particular aspect of the query and the per-aspect relevance degree of that document. To compare AspecTiles with the standard SERP interface in terms of usefulness, we conducted a user study involving 30 search tasks designed based on the TREC web diversity task topics as well as 32 participants. Our main findings are:

- In multi-aspect tasks, participants were significantly more efficient with AspecTiles than with the baseline interface, in terms of time to find the first relevant bookmark, number of documents clicked before finding the first relevant bookmark, the average time between finding two relevant bookmarks, and the average number of document clicks between finding two relevant bookmarks. On the other hand, the search effectiveness with AspecTiles was at least as high as that with the baseline: in fact, the five-tile AspecTiles significantly outperformed the baseline in terms of recall as well.

- In multi-aspect tasks, AspecTiles users used significantly fewer query reformulations, shorter queries, and dug deeper in the ranked list, compared to users of the baseline interface.

- Participants of our user study found AspecTiles significantly more useful for finding relevant information, easier to use, and were more willing to use it again than the baseline interface.

These results suggest that simple interfaces like AspecTiles can enhance the search performance and search experience of the user when their queries are underspecified. While we did not obtain statistically significant results for single-aspect tasks, the general trend suggests that AspecTiles may help these cases as well.

Table 7: Exit questionnaire results. The average and standard deviation values are reported. Significant differences with the baseline at $p \leq 0.05$ according to the Mann-Whitney's U test are shown in bold.

Questions	Baseline	Three-tile AspecTiles	Five-tile AspecTiles
Q1: Did you find the interface useful for finding relevant information?	2.875 (0.941)	**3.718** (0.581)	
Q2: Was it easy to use the interface?	3.156 (0.919)	**3.562** (0.800)	
Q3: Do you want to use the interface again?	2.875 (0.906)	**3.593** (0.837)	
Q4: Was the color information intuitive?	-	3.218 (1.099)	
Q5: Were the aspect labels appropriate?	-	3.812 (0.997)	
Q6: Was the number of displayed aspects appropriate?	-	3.437 (0.877)	3.562 (0.913)

As our future work, we plan to enable the aspect-based reranking feature that was mentioned in Section 4.3, to enable both more efficient and more exploratory search. Moreover, as our results show that the quality of displayed aspects and the accuracy of per-aspect relevance estimates are important for maximizing the benefit of AspecTiles, we plan to improve the backend search system. Finally, as we discussed in Section 4.2.2, we would like to identify the class of users for which interfaces like AspecTiles would be most helpful.

6. ADDITIONAL AUTHOR

Shojiro Nishio
(Osaka University, Japan, email: nishio@ist.osaka-u.ac.jp)

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Leong. Diversifying Search Results, In *Proc. of WSDM 2009*, pages 5–14, 2009.

[2] C. Brandt, T. Joachims, Y. Yue, and J. Bank. Dynamic Ranked Retrieval, In *Proc. of WSDM 2011*, pages 247–256, 2011.

[3] A. Broder. A Taxonomy of Web Search, *ACM SIGIR Forum*, 36(2):3–10, 2002.

[4] M. Chau. Visualizing Web Search Results Using Glyphs: Design and Evaluation of a Flower Metaphor, *ACM Transactions on Management Information Systems*, Vol. 2, No. 1, Article 2, 2011.

[5] C. Chen and Y. Yu. Empirical Studies of Information Visualization: A Meta-Analysis, *Int'l Journal of Human-Computer Studies*, 53(5):851–866, 2000.

[6] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track, In *Proc. of TREC 2009*, 2010.

[7] C. L. Clarke, M. Kolla, and O. Vechtomova. An Effectiveness Measure for Ambiguous and Underspecified Queries, In *Proc. of ICTIR 2009*, pages 188–199, 2009.

[8] Z. Dou, S. Hu, K. Chen, R. Song, and J. -R. Wen. Multi-Dimensional Search Result Diversification, In *Proc. of WSDM 2011*, pages 475–484, 2011.

[9] M. A. Hearst. TileBars: Visualization of Term Distribution Information in Full Text Information Access, In *Proc. of CHI 1995*, pages 59–66, 1995.

[10] T. Heimonen and N. Jhaveri. Visualizing Query Occurrence in Search Result Lists, In *Proc. of IV 2005*, pages 877–882, 2005.

[11] W. Hersh, A. M. Cohen, P. Roberts, and H. K. Rekapalli. TREC 2006 Genomics Track Overview, In *Proc. of TREC 2006*, pages 52–78, 2006.

[12] W. Hersh and P. Over. Interactivity at the Text Retrieval Conference (TREC), *Information Processing and Management*, 37(3):365–367, 2001.

[13] W. Hersh and P. Over. TREC-9 Interactive Track Report, In *Proc. of TREC-9*, pages 41–50, 1999.

[14] O. Hoeber and X. D. Yang. A Comparative User Study of Web Search Interfaces: HotMap, Concept Highlighter, and Google, In *Proc. of WI 2006*, pages 866–874, 2006.

[15] J. Mackinlay. Automating the Design of Graphical Presentations of Relational Information, *ACM Transactions on Graphics*, 5(2):110–141, 1986.

[16] H. Reiterer, G. Tullius. and T. M. Mann. INSYDER: A Content-based Visual-Information-Seeking System for the Web, *Int'l Journal on Digital Libraries*, 5(1):25–41, 2005.

[17] T. Sakai. Evaluation with Informational and Navigational Intents, In *Proc. of WWW 2012*, pages 499–508, 2012.

[18] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting Query Reformulations for Web Search Result Diversification, In *Proc. of WWW 2010*, pages 881–890, 2010.

[19] 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 *Proc. of NTCIR-9*, pages 82–105, 2011.

[20] O. Turetken and R. Sharda. Visualization of Web Spaces: State of the Art and Future Directions, *ACM SIGMIS Database*, 38(3):51–81, 2007.

[21] C. Ware. Information Visualization: Perception for Design, *Morgan Kaufmann*, 2000.

[22] Y. Xiang, M. Chau, H. Atabakhsh, and H. Chen. Visualizing Criminal Relationships: Comparison of a Hyperbolic Tree and a Hierarchical List, *Decision Support Systems*, 41(1):69–83, 2005.

[23] Y. Yamamoto and K. Tanaka. Enhancing Credibility Judgment of Web Search Results, In *Proc. of CHI 2011*, pages 1235–1244, 2011.