A Comparative Study of Query-biased and Non-redundant Snippets for Structured Search on Mobile Devices

Nikita Spirin, Alexander Kotov, Karrie Karahalios, Vassil Mladenov, Pavel Izhutov

1UIUC, Urbana, IL; 2Wayne State University, Detroit, MI; 3Stanford University, Stanford, CA
Department of Computer Science1, 2 and Graduate School of Business3
{spirin2,kkarahal,mladeno2}@illinois.edu, kotov@wayne.edu, izhutov@stanford.edu

ABSTRACT
To investigate what kind of snippets are better suited for structured search on mobile devices, we built an experimental mobile search application and conducted a task-oriented interactive user study with 36 participants. Four different versions of a search engine result page (SERP) were compared by varying the snippet type (query-biased vs. non-redundant) and the snippet length (two vs. four lines per result). We adopted a within-subjects experiment design and made each participant do four realistic search tasks using different versions of the application. During the study sessions, we collected search logs, “think-aloud” comments, and post-task surveys. Each session was finalized with an interview. We found that with non-redundant snippets the participants were able to complete the tasks faster and find more relevant results. Most participants preferred non-redundant snippets and wanted to see more information about each result on the SERP for any snippet type. Yet, the participants felt that the version with query-biased snippets was easier to use. We conclude with a set of practical design recommendations.

Keywords
Mobile Search; Structured Search; Snippet; User Study

1. INTRODUCTION
The ideal search system presents the best results at the top. However, due to complexities of the natural language existing systems cannot guarantee perfect retrieval results. Therefore, relevant results might appear in any position on a SERP. To address this issue, search engines “cooperate” with their users via intelligent search user interfaces (SUI). On the SERP, each result is represented as a title/name and accompanied by a snippet, which summarizes the key information about the result and helps users assess relevance.

The de facto method to generate search snippets is based on the extraction of sentences or attributes containing query terms [8, 17, 19–21]. Such query-biased snippets signal relevance by letting users see the usage of the query terms in context. Initially introduced for full-text search [17], this method is used nowadays for all sorts of search applications.

However, this approach has limitations when both queries and documents are structured and the matching is exact rather than probabilistic (e.g., structured/faceted search [18] provided by e-commerce or social networking sites). In this case, query-biased snippets don’t provide any new information as they merely repeat the constraints on attribute values expressed in the query. Therefore, all results on the SERP look the same and a user might have a hard time deciding which results to view/click on. We illustrate this limitation in Figure 1. On the left, there is a SERP for structured search in LinkedIn’s Android mobile app, which provides non-discriminative snippets and exhibits high query-snippet redundancy. On the right, there is the same SERP with the redundant information removed. As we can see, the original seemingly information-rich interface on the left actually contains a very minimal amount of new information as indicated by the blank areas on the right. Moreover, in the case of structured search, we observe a query-snippet duality, i.e., the longer the query, the more redundant and uninformative are the snippets. This might demotivate users from submitting longer queries, which are shown to be more effective [1].

Thus, researchers proposed multiple approaches to generate non-redundant snippets for structured search. The main idea behind these approaches is that we can utilize space on the SERP more effectively by showing information that helps differentiate the results, and, hence, complementary to the query. For example, there are methods to generate informative [3, 4] and diverse [10, 12] snippets for database tuples, for XML documents [8], and for job postings [14].

To sum up, there is a disconnect between the theory and practice. On the one hand, there are powerful methods to generate non-redundant, informative, and discriminative snippets for structured documents [3, 4, 8, 10, 12, 14]. On the other hand, query-biased snippets have shown its utility for full-text search [17, 20, 21] and are used ubiquitously (e.g., Figure 1 and http://imgur.com/a/aEqEq). We formulated two research questions to resolve this disconnect.

- **RQ1:** Which snippets do users prefer in the case of structured search on mobile devices? (subjective)
- **RQ2:** Which snippets make users more productive and effective when performing structured search on mobile devices? (objective)

We created an experimental mobile search application (Section 2) and conducted a task-oriented interactive user study (Section 3) to answer these questions and determine the best
2. EXPERIMENTAL SYSTEM

2.1 Experimental Collection

As the collection, we used Indeed’s Resume Search production index, which contains over 21,000,000 real resumes. Our application programmatically interacted with Indeed by sending carefully tuned HTTP requests and parsing the responses. For example, to retrieve only resumes from “New York”, we appended the in-New-York-NY?co=US parameter string to the URL, got the HTML version of a SERP from Indeed, and post-processed it to generate our own SERP.

2.2 Structured Search User Interface

Our app consisted of four different pages. The Settings Page (where users could enter their anonymized userID), select the task type and the version of the SERP), the Filters Page (where users could specify search criteria), and the Details Page (where the detailed information about the particular search result was shown) were the same in all versions of the experimental system. However, our app had four different versions of the SERP shown in Figures 2(a-d):

- QB2, which showed query-biased snippets with two attribute lines per result (Figure 2(a));
- NR2, which showed non-redundant snippets with two attribute lines per result (Figure 2(b));
- QB4, which showed query-biased snippets with four attribute lines per result (Figure 2(c));
- NR4, which showed non-redundant snippets with four attribute lines per result (Figure 2(d)).

To submit a new query, a user had to open the Filters Page and set constraints/filters on the attribute values using drop-down menus. Then, the user could either submit a modified query or go back to the original SERP keeping the previous query. The user could filter results by using eight different attributes: Job Title, Company, University, Degree, Location, Skills, Years of Experience, and Major. The SUI was designed in such a way that only one attribute value could be selected at a time. We decided to adopt this strategy to get the maximum control over the collected experimental data. Finally, there was a special Favorites filter, which overrides all other filters and returns only “starred” results (added to favorites). The search criteria (query constraints) were always shown as tags/“breadcrumbs” [7] in the navigation bar. The navigation bar stayed at the top of the SERP as users scrolled down. For query-biased snippets it merely showed the current query, while for non-redundant snippets it had a more important role — it served as a part of the snippet shared across all results (in the exact match scenario all results share a subset of attribute values).

Query-biased snippets were generated as follows. First, we took all the attributes specified in a query and sorted them by priority (a global score for attribute importance that could be derived based on the frequency of filters usage or a user study). Then, we appended all attributes that weren’t mentioned in the query also sorted by priority. Finally, we took as many attributes from the top of this concatenated attribute list as there were lines in a specific snippet version. For non-redundant snippets we did the same but considered only the attributes not used in the query.

3. USER STUDY METHODOLOGY

We used a simulated work task situation approach [2, 9], i.e. we analyzed the behavior of participants seeking to accomplish real tasks by interacting with the experimental search application. The study consisted of two stages.

In the first stage (12 participants), we compared four different versions of the SERP by varying the snippet type and the snippet length. Adopting a within-subjects study design, we made each participant perform four different search tasks using four different versions of the system. Each participant performed only one randomly assigned task using each version. The task version order was randomized following the Greek-Latin square experiment design [16] to account for fatigue, task difficulty, and learning effects. The goal was to quickly eliminate two least favorable versions.

In the second stage (24 participants), we compared only the best two versions from the first stage. This time, each participant performed two out of four tasks using each version. It allowed us to increase the number of measurements.
per participant-version combination to two and make the analysis more rigorous. The goal of the second stage was to actually answer our research questions and collect extensive data to make valid conclusions. Again, we used the Greek-Latin square design as a randomization protocol.

3.1 Participants

The participants were required to be at least 18 years old, use a social networking site at least once a week, search for people online at least once a month, and have a mobile device. Of the 36 participants that we recruited, 32 were UIUC students with diverse majors (e.g. Computer Science, Business, Linguistics, etc.) and four were working professionals. 20 participants were female. More than 80% of the participants searched for people online at least once a week and more than 30% did it every day. All participants were active Facebook (log in at least a few times per week) and LinkedIn (60% log in at least once a week) users.

3.2 Tasks

We designed four people search tasks simulating a real need: (1) find a person to ask for career advice, (2) find a keynote speaker for a conference, (3) help a recruiter find candidates to hire, (4) find a collaborator for a project. For instance, the wording for Task 2 was “Suppose you are a part of the organizing committee for a conference on _______ and your task is to find a keynote speaker. You understand that people are busy and not everyone will agree/have time to accept your invitation. Your strategy is to find five candidates, message them, and hope that one of them will accept the invitation. A search tool is provided to assist you in this task. You are also required to rank and evaluate the selected candidates after you finish the task.”. The participants needed to fill the gaps in the task templates based on their background. We did it to further increase relevance of the tasks to the participants and make them as realistic as possible. We asked the participants to pick a different topic for each task to avoid carry-over effects across tasks. According to our post-task surveys (Table 3, Q5), the participants found the tasks as realistic (4.06 on a 5-point Likert scale).

3.3 Experimental Procedure

Each participant in our user study was asked to work on four tasks using different versions of the experimental search system. We allocated 10 minutes per task, which is typically enough for users to carefully explore the results [6, 21]. The participants were asked to “think-aloud” as they worked on the tasks, and the experimenter took the detailed notes.

After each task, a participant was asked to complete a short post-task survey to capture her/his subjective satisfaction with the specific version of the system, which s/he just tested. The survey contained nine multiple choice questions (eight 5-point Likert scales [13] and one 5-point semantic differential). We also asked each participant to evaluate the quality of selected search results on a 5-point Likert scale. Each result was judged independently, and, hence, more than one result could receive the same relevance score.

All tasks were completed using our app with no interventions by the experimenter. The study was conducted in the same room to decrease the influence of external factors on study results. Since we interacted with the live search index, we grouped all user study sessions as close as possible time-wise (12 days starting March 24th, 2016) to ensure minimal influence of changes in Indeed’s index on the study results.

At the end of each study session, we conducted a 15-20 minute semi-structured interview by asking open-ended questions motivating the participants to express their thoughts about four (two, during Stage 2 of the study) versions of the SERP. The participants could see all four (two) versions and provide suggestions on how to improve the application. To understand the way participants interacted with the search system, we logged all possible actions performed by them, such as page views, clicks on the search results, additions/removals from favorites, and page scrolls. For each event, we always logged the following metadata: task type, snippet type, timestamp, current query (combination of constraints on attribute values), and anonymized userID, which allowed us to measure system effectiveness and efficiency. We describe the metrics/dimensions that we used to compare different system versions in Section 5.
4. USER STUDY RESULTS: STAGE ONE
The key conclusion from 12 user study sessions in Stage 1 is that the participants mostly noticed the difference in the snippet length and not in the snippet type and favored longer snippets (11 out of 12 participants). Eight out of 12 participants placed NR4 in either first or second place; three placed QR4 in either first or second place. When scored on a scale from 1 to 4 (lower is better), shorter snippets obtained mean scores of 3.5 (QB2) and 2.8 (NR2), whereas longer snippets obtained mean scores of 2.2 (QB4) and 1.5 (NR4).

Surprisingly, we found that the extra scrolling cost caused by longer snippets didn’t have a negative impact on participants’ preferences. Many participants said that they either didn’t notice it or preferred to have more information on the SERP to make more informed click decisions, like in [22]. It is also worth noting from this stage of the study that most of the participants preferred non-redundant snippets over query-biased snippets given the fixed snippet length.

5. USER STUDY RESULTS: STAGE TWO
Based on the results from Stage 1, we limited the number of snippet versions to two and only kept the ones that were more favored by the participants, i.e. in Stage 2 we compared query-biased snippets spanning 4 lines (QB4) and non-redundant snippets spanning 4 lines (NR4). Both objective and subjective data was used to compare these versions. The objective evaluation was based on search user interaction logs, and the subjective evaluation was based on user post-task survey responses and post-study interviews.

5.1 Subjective System Preference
At the end of each session, we asked the participants “Which version do you like most?” The majority (17 out of 24) preferred NR4, which is statistically significant according to the exact two-tailed binomial test (p = 0.0433).

Next, we probed the participants further and asked them to explain the reasons behind their decision (“Why do you like this version more? What makes it better than the other version?”). In responding to these open-ended questions, the participants provided many rich and diverse opinions.

The participants who preferred NR4 said that it: (a) shows new non-redundant information (15 out of 17 participants); (b) helps discriminate results (11 out of 17 participants); (c) shows more relevant attributes (6 out of 17 participants); (d) helps accomplish a task faster (3 out of 17 participants); (e) requires less scrolling (3 out of 17 participants); (f) returns more relevant results (1 out of 17 participants).

The participants who preferred QB4 paid more attention to other aspects of the SERP. They liked the version with query-biased snippets because it: (a) has a more regular layout (6 out of 7 participants); (b) shows more relevant attributes (5 out of 7 participants); (c) is more predictable and reassuring (5 out of 7 participants); (d) demands less cognitive load/effort (5 out of 7 participants); (e) provides a good balance of selected and new information (2 out of 7 participants); (f) works faster (2 out of 7 participants); (g) returns more relevant results (1 out of 7 participants); (h) forces to check individual profiles (1 out of 7 participants).

To summarize, almost everyone who preferred NR4 noted the lack of redundancy. On the other hand, rather than focusing on SERP informativeness, the participants who favored QB4 liked the consistent layout of the SERP. This key finding is also supported by numerous other subjective data that we collected during the study (Table 3). For example, in the post-task surveys, people gave slightly higher scores to NR4 (Q6) but felt that QB4 was easier to use (Q1) and that the search process was less stressful (Q9).

5.2 System Effectiveness
Rather than using predefined topics typical for information retrieval system evaluation [9], we used templated tasks to closer simulate real life situations and motivate participants to engage in exploratory search, where the objective is usually subtle and might evolve during the search session [11]. Therefore, only the participants themselves could evaluate the quality of selected candidates. To this end, after each task we asked the participants to provide independent relevance judgments to the selected candidates on a scale from 1 to 5 (5 is the highest), like in [6, 22]. Then, we computed the average relevance score over the five selected candidates. We also computed the average relevance score for the candidates selected at a specific position in the session, i.e. if some candidate was the first discovered candidate among five selected candidates, then his/her relevance score was aggregated under the First group. The results are presented in a global comparative Table 3.

From the second part of Table 3, it follows that NR4 helps people find more relevant candidates than QB4. The data also suggests that candidates discovered at a specific position in the session have a higher relevance in the case of NR4.

The post-task survey responses also indicate that NR4 helps find more relevant results. The participants rated NR4 as more helpful for deciding who to contact (Q3), felt that the information on the SERP was more useful (Q5 and Q7), and, overall, rated NR4 slightly higher (Q6).

5.3 System Efficiency
In addition to the analysis of system effectiveness, we also used query log data and subjective post-task responses to compare the efficiency of different systems. The primary measure to assess the efficiency of different systems was time. In particular, we computed the average time required by each participant to finish a task and found that NR4 led to faster task completion times in seconds (µ = 389.04; σ = 118.30) than QB4 (µ = 414.77; σ = 122.10).

We performed further analysis to understand why the participants were able to complete the tasks faster with NR4. We discovered that the average time between consecutive queries within a session was smaller for NR4 (µ = 56.77; σ = 34.88) than for QB4 (µ = 63.80; σ = 39.59) and that the average number of queries per session was also smaller for NR4 (µ = 4.96; σ = 3.90) than for QB4 (µ = 5.91; σ = 5.60). We continued our investigation to understand why the time between queries was shorter and why the participants submitted less queries when using NR4. We discovered that the participants made less clicks on the SERP (visited fewer number of detailed profiles) in the case of NR4 (µ = 16.00; σ = 4.60) than in the case of QB4 (µ = 18.51; σ = 5.80). This difference is statistically significant according to the non-parametric two-tailed Mann-Whitney test (p = 0.0368). However, time to the first SERP click was longer for NR4 (µ = 13.17; σ = 6.96) compared to QB4 (µ = 11.36; σ = 4.71), which suggests that the participants engaged more with the SERP and made more informed decisions by looking at the snippets. There are several other...
findings supporting the hypothesis that the participants engaged with the SERP more in the case of NR4. While not statistically significant, the participants added (µ = 1.44; σ = 2.77) and removed (µ = 0.65; σ = 1.40) from favorites on the SERP more often when using NR4 than when using QB4 (µ = 1.00; σ = 2.13) and removed (µ = 0.43; σ = 1.11) when using QB4. In the surveys, the utility of NR4 was also rated higher (Q5, Q7) compared to QB4.

6. DISCUSSION AND IMPLICATIONS

In this section, we discuss the implications from our findings for the design of more useful future structured SUIs.

6.1 Eliminate Redundancy

Our key finding is that users do notice redundancy in the search user interface and prefer non-redundant snippets. At the same time, non-redundant snippets cannot be used when users specify multiple constraints per attribute (e.g. “Uber or Lyft or YellowCabNYCTaxi”) or when an attribute has an array type (e.g. skills “Java, Scala, Hadoop, Spark”) because attribute values will vary from one result to another. Therefore, we suggest using non-redundant snippets when the query formulation interface doesn’t allow people to specify multiple constraints per attribute (e.g. Zappos mobile app) or when only one value per facet is selected. Query-biased snippets should be used when the query implies that the attribute values might vary from one result to another.

We also discovered that users mostly add people to favorites from the detailed profile page and, hence, the stars that occupy a significant part of the SERP space could be eliminated, too. We propose allowing users to (un)bookmark candidates by swiping to the left/right rather than by clicking on stars. Since the swipe is an interaction pattern, no space on the SERP is required to provide this functionality.

The freed space might be used to show some additional information to help users make more informed click decisions. From the interviews, we learned that in the case of people search the searchers not only try to assess candidates based on their relevance to the query/task but also use other factors such as social similarity, status, approachability, seniority. Therefore, to simplify this process for the users, search engines might explicitly show on the SERP the similarity score between a searcher and a candidate result or a simple data visualization with the scores along several dimensions [6]. Plus, these scores can be made more interpretable by showing explanatory comments next to each number, e.g. “87% (both you and John Smith went to MIT).”

From Stage 1, we concluded that users don’t notice the scrolling cost and prefer having more information about each result on the SERP. Therefore, we suggest using long non-

<table>
<thead>
<tr>
<th>Metric \ Snippet Version</th>
<th>Query-biased, 4</th>
<th>Non-redundant, 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Part I: Post-task Subjective Survey Responses</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1: The system is easy to use</td>
<td>4.06 ± 0.59</td>
<td>3.91 ± 0.79</td>
</tr>
<tr>
<td>Q2: The system provides me relevant candidates</td>
<td>3.98 ± 0.98</td>
<td>3.98 ± 0.86</td>
</tr>
<tr>
<td>Q3: The system helps me decide who to contact</td>
<td>3.92 ± 0.71</td>
<td>4.13 ± 0.64</td>
</tr>
<tr>
<td>Q4: The system helps me find relevant candidates efficiently</td>
<td>3.65 ± 0.93</td>
<td>3.69 ± 0.88</td>
</tr>
<tr>
<td>Q5: The display of each profile on the SERP is useful</td>
<td>3.96 ± 0.92</td>
<td>4.04 ± 0.77</td>
</tr>
<tr>
<td>Q6: Overall, I am satisfied with the system in this task</td>
<td>3.71 ± 0.68</td>
<td>3.73 ± 0.68</td>
</tr>
<tr>
<td>Q7: Summaries/attributes presented for each result on the SERP are useful</td>
<td>3.94 ± 0.93</td>
<td>4.08 ± 0.82</td>
</tr>
<tr>
<td>Q8: I can see myself doing this task in the real world</td>
<td>4.06 ± 0.81</td>
<td>4.06 ± 0.78</td>
</tr>
<tr>
<td>Q9: The search process is (stressful / relaxing)</td>
<td>3.40 ± 0.87</td>
<td>3.35 ± 0.93</td>
</tr>
</tbody>
</table>

**Part II: Post-task Subjective Result Relevance Judgments**

<table>
<thead>
<tr>
<th></th>
<th>Query-biased, 4</th>
<th>Non-redundant, 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected result relevance</td>
<td>4.00 ± 0.96</td>
<td>4.07 ± 0.92</td>
</tr>
<tr>
<td>First selected result relevance</td>
<td>4.23 ± 0.81</td>
<td>4.23 ± 0.93</td>
</tr>
<tr>
<td>Second selected result relevance</td>
<td>4.02 ± 0.93</td>
<td>4.04 ± 0.82</td>
</tr>
<tr>
<td>Third selected result relevance</td>
<td>3.81 ± 0.87</td>
<td>3.88 ± 1.10</td>
</tr>
<tr>
<td>Fourth selected result relevance</td>
<td>3.83 ± 1.17</td>
<td>3.90 ± 0.99</td>
</tr>
<tr>
<td>Fifth selected result relevance</td>
<td>4.10 ± 0.97</td>
<td>4.31 ± 0.69</td>
</tr>
</tbody>
</table>

**Part III: Query Log Data**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Query-biased, 4</th>
<th>Non-redundant, 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query length</td>
<td>3.03 ± 0.97</td>
<td>3.14 ± 0.95</td>
</tr>
<tr>
<td># of queries per task session</td>
<td>5.91 ± 5.60</td>
<td>4.96 ± 3.90</td>
</tr>
<tr>
<td>Time between consecutive queries within a task session (seconds)</td>
<td>63.80 ± 39.59</td>
<td>56.77 ± 34.88</td>
</tr>
<tr>
<td>Time to complete a task (seconds)</td>
<td>414.77 ± 122.10</td>
<td>395.04 ± 115.30</td>
</tr>
<tr>
<td>Time to the first SERP click after submitting a query (seconds)</td>
<td>11.36 ± 4.71</td>
<td>13.17 ± 6.96</td>
</tr>
<tr>
<td># of SERP clicks (profile views) per task session</td>
<td>18.51 ± 5.80</td>
<td>16.00 ± 4.60</td>
</tr>
<tr>
<td># of profiles added to favorites per task session (SERP)</td>
<td>1.00 ± 2.13</td>
<td>1.44 ± 2.77</td>
</tr>
<tr>
<td># of profiles removed from favorites per task session (SERP)</td>
<td>0.43 ± 1.11</td>
<td>0.65 ± 1.40</td>
</tr>
<tr>
<td># of profiles added to favorites per task session (detailed profile page)</td>
<td>4.85 ± 2.11</td>
<td>4.48 ± 2.03</td>
</tr>
<tr>
<td># of profiles removed from favorites per task session (detailed profile page)</td>
<td>0.45 ± 0.78</td>
<td>0.39 ± 0.74</td>
</tr>
</tbody>
</table>

**Part IV: Post-study Subjective Preference (Interview Data)**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Query-biased, 4</th>
<th>Non-redundant, 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of participants favoring a specific version (out of 24)</td>
<td>29.2% (7/24)</td>
<td>70.8% (17/24)</td>
</tr>
</tbody>
</table>

Table 1: Metrics for two versions of the experimental application used in the formal study (Stage 2): query-biased snippets with four lines per result and non-redundant snippets with four lines per result. We use *** to mark statistically significant differences according to the two-tailed Mann-Whitney test (p < 0.05).
redundant snippets. At the same time, we warn the readers that there must be at least a few results per page to enable comparison. This is because several of our participants explicitly mentioned that they used the results above and below the result in question to decide whether to ‘star’ it.

6.2 Provide More Control

Many participants mentioned that they would love to have more control over the search process, e.g. the ability to specify what attributes to show on the SERP or set importance weights for attributes to manipulate result ranking, like in [6]. The participants also mentioned that, despite having a good ranking algorithm, they would prefer to have more ranking algorithm transparency and be able to resort the attributes based on well-defined criteria, e.g. years of work experience and age. Therefore, we suggest exposing the seams and providing more control, which might be especially beneficial for advanced users. Interestingly, some of the participants even wanted to reorder the sections of the resume depending on the task. The algorithmic transparency might lead to many positive outcomes for the users [5] and establish a higher level of trust between the users and the systems.

From the study, we also learned that some users prefer to have input confirmation, while others seek more information per pixel. We think that a balanced solution might be designed. One idea is to mix non-redundant and query-biased snippets by always showing one query-biased attribute and filling in the rest of the lines with non-redundant information. Alternatively, this decision might be “delegated” to the users such that they could select a snippet type in the search settings. Finally, one can predict the user type algorithmically based on search interaction data and show the snippets which are more appropriate for them, e.g. the elderly users might benefit from a less dynamic interface with query-biased snippets, while the younger users actively using digital products and searching online might see a more information-rich SERP with non-redundant snippets.

7. CONCLUSIONS AND FUTURE WORK

This work serves as the first real evidence discovered via an interactive user-centric study that structured search engines should eliminate SUI redundancy. We found that the participants liked non-redundant snippets more when performing structured search on mobile devices. Our qualitative and quantitative data indicates that the system with non-redundant snippets was more effective and efficient. Non-redundant snippets led to faster task completion times, helped find more relevant results, and made the participants do more informed SERP click decisions. The participants engaged with the SERP more when using the system with non-redundant snippets. On the other hand, the participants who favored query-biased snippets paid more attention to the visual features of the SUI and felt that the system was easier to use. We also learned that the participants favored longer snippets without respect to the snippet type.

Despite our efforts to design a rigorous study, it still has several limitations. First, we focused on people search, and that’s why our findings are limited only to this search vertical. In the future, we will examine other structured search verticals, such as travel, e-commerce, or jobs. Second, the majority of our participants were students. We plan to expand the pool of participants and investigate whether there is some interaction between a snippet type and a level of professional experience. Third, we used subjective judgments to evaluate system effectiveness since there was no ground-truth labels. A live A/B test should help understand what kind of snippet leads to better search outcomes for end-users.

8. ACKNOWLEDGEMENTS

We thank Nick Craswell, Susan Dumais, Jiawei Han, Dawid Hawking, Daniel Tunkelang, ChengXiang Zhai, all participants of the WSDM2016 DC, and anonymous CIKM reviewers for thoughtful discussions and critical comments.

9. REFERENCES