A Study of Snippet Length and Informativeness

Behaviour, Performance and User Experience

David Maxwell
School of Computing Science
University of Glasgow
Glasgow, Scotland
d.maxwell.1@research.gla.ac.uk

Leif Azzopardi
Computer & Information Sciences
University of Strathclyde
Glasgow, Scotland
leif.azzopardi@strath.ac.uk

Yashar Moshfeghi
School of Computing Science
University of Glasgow
Glasgow, Scotland
Yashar.Moshfeghi@glasgow.ac.uk

ABSTRACT

The design and presentation of a *Search Engine Results Page (SERP)* has been subject to much research. With many contemporary aspects of the SERP now under scrutiny, work still remains in investigating more traditional SERP components, such as the *result summary*. Prior studies have examined a variety of different aspects of result summaries, but in this paper we investigate the influence of result summary length on search behaviour, performance and user experience. To this end, we designed and conducted a within-subjects experiment using the *TREC AQUAINT* news collection with 53 participants. Using *Kullback-Leibler distance* as a measure of *information gain*, we examined result summaries of different lengths and selected four conditions where the change in information gain was the greatest: *(i)* title only; *(ii)* title plus one snippet; *(iii)* title plus two snippets; and *(iv)* title plus four snippets. Findings show that participants broadly preferred longer result summaries, as they were perceived to be more informative. However, their performance in terms of correctly identifying relevant documents was similar across all four conditions. Furthermore, while the participants felt that longer summaries were more informative, empirical observations suggest otherwise; while participants were more likely to click on relevant items given longer summaries, they also were more likely to click on non-relevant items. This shows that longer is not necessarily better, though participants perceived that to be the case – and second, they reveal a positive relationship between the length and informativeness of summaries and their attractiveness (i.e. clickthrough rates). These findings show that there are tensions between perception and performance when designing result summaries that need to be taken into account.

CCS CONCEPTS

•Information systems →Search interfaces;

ACM Reference format:
David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. 2017. A Study of Snippet Length and Informativeness. In *Proceedings of SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan, ,* 10 pages.
DOI: http://dx.doi.org/10.1145/3077136.3080824

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080824

1 INTRODUCTION

Interactive Information Retrieval (IIR) is a complex, non-trivial process where a searcher undertakes a variety of different actions during a search session [16]. Core to their experience and success is the *Search Engine Results Page (SERP)*, with its presentation and design over the years having been subject to much research. With more complex components now becoming commonplace in modern day Web search engines (such as the *information card* [5, 36] or *social annotations* [35]), much work however still remains on examining how more traditional SERP components (such as *result summaries*) are designed and presented to end users.

Result summaries have traditionally been viewed as the *'ten blue links'* with the corresponding URL of the associated document, and one or more textual *snippets* of *keywords-in-context* from the document itself, approximately 130-150 characters (or two lines) in length [15]. Numerous researchers have explored result summaries in a variety of different ways, such as: examining their length [11, 19, 38]; the use of thumbnails [47, 52]; their attractiveness [9, 14]; and the generation of *query-biased snippets* [41, 48]. The performance of users has broadly been evaluated in a limited fashion (e.g. by examining task completion times). In this work, we are interested in how the length and information content of result summaries affects SERP interactions and a user's ability to select relevant over non-relevant items. Prior research has demonstrated that longer result summaries tend to lower completion times for informational tasks (where users need to find one relevant document) [11], but does this hold in other contexts, specifically for ad-hoc retrieval, where users need to find *several* relevant items? Furthermore, how does the length and information associated with longer result summaries affect the user's ability to discern the relevant from the non-relevant?

This work therefore serves as an investigation into the effects of search behaviour and search performance when we vary *(i)* result summary snippet lengths, and by doing so *(ii)* the information content within the summaries. To this end, a within-subjects crowdsourced experiment ($n = 53$) was designed and conducted. Under ad-hoc topic retrieval, participants used four different search interfaces, each with a different size of result summary. Findings allow us to address the two main research questions of this study. ***RQ1*** How does the value of information gain represented as snippet length affect behaviour, performance and user experience? ***RQ2*** Does information gain – again represented as snippet length – affect the decision making ability and accuracy (identifying relevant documents) of users? We hypothesise that longer and more informative snippets will enable users to make better quality decisions (i.e. higher degrees of accurately identifying relevant content).

2 RELATED WORK

As previously mentioned, the design and presentation of SERPs has been examined in depth. Researchers have examined various aspects of SERPs, and how the designs of such aspects influence the behaviour of users. Here, we provide a summary of the various aspects that have been investigated. Specifically, we focus upon: the layout of SERPs; the size of SERPs; how snippet text is generated; and how much text should be presented within each result summary – the latter being the main focus of this work.

2.1 SERP Layouts and Presentation

Early works regarding the presentation of result summaries [6, 12] examined approaches to automatically categorise result summaries for users, similar to the categorisation approach employed by early search engines. Chen and Dumais [6] developed an experimental system that automatically categorised result summaries on-the-fly as they were generated. For a query, associated categories were then listed as verticals, with associated document titles provided underneath each category header. Traditional result summaries were then made available when hovering over a document title. Subjects of a user study found the interface easier to use than the traditional 'ten blue links' approach - they were 50% faster at finding information displayed in categories. This work was then extended by Dumais et al. [12], where they explored the use of hover text to present additional details about search results based upon user interaction. Searching was found to be slower with hover text, perhaps due to the fact that explicit decisions about when to seek additional information (or not to) were required.

Alternatives to the traditional, linear list of result summaries have also been trialled (like grid-based layouts [8, 20, 40]). For example, Krammerer and Beinhaur [20] examined differences in user behaviour when interacting with a standard list interface, compared against a tabular interface (title, snippet and URL stacked horizontally in three columns for each result), and a grid-based layout (result summaries placed in three columns). Users of the grid layout spent more time examining result summaries. The approach demonstrated promise in overcoming issues such as *position bias* [10], as observed by Joachims et al. [17].

Marcos et al. [34] performed an eye-tracking user study examining the effect of user behaviour while interacting with SERPs – and whether the *richness* of result summaries provided on a SERP (i.e. result summaries enriched with metadata from corresponding pages) impacted upon the user's search experience. Enriched summaries were found to help capture a user's attention. Including both textual and visual representations of a document when presenting results could have a positive effect on relevance assessment and query reformulation [18]. Enriched summaries were also examined by Ali et al. [2] in the context of navigational tasks. Striking a good balance between textual and visual cues were shown to better support user tasks, and search completion time.

2.2 Generating Snippet Text

Users can be provided with an insight by result summaries as to whether a document is likely to be relevant or not [14]. Consequently, research has gone into examining different kinds of snippets, and how long a snippet should be. Work initially focused upon how these summaries should be generated [30, 31, 39, 48, 51]. These early works proposed the idea of summarising documents with respect to the query (query-biased summaries) or keywords-in-context – as opposed to simply extracting the representative or lead sentences from the document [29]. Tombros and Sanderson [48] showed that subjects of their study were likely to identify relevant documents more accurately when using query-biased summaries, compared to summaries simply generated from the first few sentences of a given document. Query-biased summaries have also been recently shown to be preferred on mobile devices [45].

When constructing snippets using query-biased summaries, Rose et al. [41] found that a user's perceptions of the result's quality were influenced by the snippets. If snippets contained truncated sentences or many fragmented sentences *(text choppiness)*, users perceived the quality of the results more negatively, regardless of length. Kanungo and Orr [21] found that poor readability also impacts upon how the resultant snippets are perceived. They maintain that readability is a crucial presentation attribute that needs to be considered when generating a query-biased summary. Clarke et al. [9] analysed thousands of pairs of snippets where result *A* appeared before result *B*, but result *B* received more clicks than result *A*. As an example, they found results with snippets which were very short (or missing entirely) had fewer query terms, were not as readable, and attracted fewer clicks. This led to the formulation of several heuristics relating to document surrogate features, designed to emphasise the relationship between the associated page and generated snippet. Heuristics included: *(i)* ensuring that all query terms in the generated snippet (where possible); *(ii)* withholding the repeating of query terms in the snippet if they were present in the page's title; and *(iii)* displaying (shortened) readable URLs.

Recent work has examined the generation of snippets from more complex angles – from manipulating underlying indexes [4, 49] to language modelling [14, 32], as well as using user search data to improve the generation process [1, 42]. Previous generation approaches also may not consider what parts of a document searchers actually find useful. Ageev et al. [1] incorporated into a new model post-click searcher behaviour data, such as mouse cursor movements and scrolling over documents, producing *behaviour-biased snippets*. Results showed a marked improvement over a strong text-based snippet generation baseline. Temporal aspects have also been considered – Svore et al. [46] conducted a user study, showing that users preferred snippet text with *trending* content in snippets when searching for trending queries, but not so for general queries.

2.3 Results per Page

Today, a multitude of devices are capable of accessing the *World Wide Web (WWW)* – along with a multitude of different screen resolutions and aspect ratios. The question of how many result summaries should be displayed per page – *or results per page (RPP)* – therefore becomes hugely important, yet increasingly difficult to answer. Examining behavioural effects on mobile devices when interacting with SERPs has attracted much research as of late (e.g. [24–26]), and with each device capable of displaying a different number of results *above-the-fold*, recent research has shown that the RPP value can influence the behaviour of searchers [17, 25]. Understanding this behaviour can help guide and inform those charged with designing contemporary user interfaces.

In a Google industry report, Linden [33] however stated that users desired more than 10RPP, despite the fact that increasing the RPP yielded a 20% drop in traffic; it was hypothesised that this was due to the extra time required to dispatch the longer SERPs. This drop however be attributed to other reasons. Oulasvirta et al. [37] discusses the *paradox of choice* [43] in the context of search, where more options (results) – particularly if highly relevant – will lead to poorer choice and degrade user satisfaction. In terms of user satisfaction, modern search engines can therefore be a victim of their own success, presenting users with *choice overload*. Oulasvirta et al. [37] found that presenting users with a six-item search result list was associated with higher degrees of satisfaction, confidence with choices and perceived carefulness than an a list of 24 items.

Kelly and Azzopardi [22] broadly agreed with the findings by Oulasvirta et al. [37]. Here, the authors conducted a between-subjects study with three conditions, where subjects were assigned to one of three interfaces - the baseline interface, showing 10RPP (the 'ten blue links'), and two interfaces displaying 3RPP and 6RPP respectively. Their findings showed that individuals using the 3RPP and 6RPP interfaces spent significantly longer examining top-ranking results and were more likely to click on higher ranked documents than those on the 10RPP interface. Findings also suggested that subjects using the interfaces showing fewer RPP found it comparatively easier to find relevant content than those using the 10RPP interface. However, no significant difference was found between the number of relevant items found across the interfaces. Currently, 10RPP is still considered the *de-facto* standard [15].

2.4 Snippet Lengths: Longer or Shorter?

Snippet lengths have been examined in a variety of ways. A user study by Paek et al. [38] compared a user's preference and usability against three different interfaces for displaying result summaries. With question answering tasks, the interfaces: displayed a *normal* SERP (i.e. a two line snippet for each summary, with a clickable link); an *instant* interface, where an expanded snippet was displayed upon clicking it; and a *dynamic* interface, where hovering the cursor would trigger the expanded snippet. The instant view was shown to allow users to complete the given tasks in less time than the normal baseline, with half of participants preferring this approach.

Seminal work by Cutrell and Guan [11] explored the effect of different snippet lengths (*short*: 1 line, *medium*: 2-3 lines; and *long*: 6-7 lines). They found that longer snippets significantly improved performance for *informational tasks* (e.g. `'Find the address for Newark Airport.'`). Users performed better for informational queries as snippet length increased. This work was followed up by Kaisser et al. [19]. They conducted two experiments that estimated the preferred snippet length according to answer type (e.g. finding a person, time, or place), and comparing the results of the preferred snippet lengths to users' preferences to see if this could be predicted. The preferred snippet length was shown to depend upon the type of answer expected, with greater user satisfaction shown for the snippet length predicted by their technique.

More contemporary work has begun to examine what snippet sizes are appropriate for mobile devices. Given smaller screen sizes, this is important – snippet text considered acceptable on a computer screen may involve considerable scrolling/swiping on a smaller screen. Kim et al. [27] found that subjects using longer snippets on mobile devices exhibited longer search times and similar search accuracy under informational tasks[1]. Longer reading times and frequent scrolling/swiping (with more viewport movements) were exhibited. Longer snippets did not therefore appear to be very useful on a small screen – an *instant* or *dynamic* snippet approach (as per Paek et al. [38]) may be useful for mobile search, too.

The presentation of result summaries has a strong effect on the ability of a user to judge relevancy [14]. Relevant documents may be overlooked due to uninformative summaries – but conversely, non-relevant documents may be examined due to a misleading summary. However, longer summaries also increase the examination cost, so there is likely a trade-off between informativeness/accuracy and length/cost. The current, widely accepted standard for result summaries are two query-based snippets/lines [15]. This work examines whether increasing and decreasing the length (and consequently the informativeness) of result summary snippets affects user accuracy and costs of relevance decisions in the context of ad-hoc topic search, where multiple relevant documents are sought.

3 EXPERIMENTAL METHOD

To address our two key research questions outlined in Section 1, we conducted a within-subjects experiment. This allowed us to explore the influence of snippet length and snippet informativeness on search behaviours, performance and user experience. Subjects used four different search interfaces, each of which varied the way in which result summaries were presented to them.

To decide the length and informativeness of the result summaries, we performed a preliminary analysis to determine the average length (in words) and informativeness (as calculated by the *Kullback-Leibler distance* [28] to measure *information gain*, or *relative entropy*) of result summaries with the title and varying numbers of snippet fragments (0–10). The closer the entropy value is to zero, the more information gained. Figure 1 plots the number of words, the information gain, and the information gain per word[2]. It is clear from the plot that a higher level of information gain was present in longer snippets. However, as the length increases with each additional snippet fragment added, the informativeness per word decreased. Consequently, for this study, we selected the four different interface conditions in the region where informativeness had the highest change, i.e. from zero to four. The conditions we selected for the study were therefore:

- ***T0*** where only the title for each result summary were presented;
- ***T1*** where for each result summary, a title and one query-biased snippet fragment were presented;
- ***T2*** where a title and two query-biased snippet fragments were presented; and
- ***T4*** where a title and four query-biased snippet fragments were presented,

where our independent variable is snippet informativeness, controlled by the length. Figure 2 provides an example of the different

[1] The tasks considered by Kim et al. [27] were similar to those defined by Cutrell and Guan [11], where a single relevant document was sought.

[2] To obtain these values, we submitted over 300 queries from a previous study (refer to Azzopardi et al. [3]) conducted on similar topics and on the same collection to the search system that we used.

result summaries in each condition. The remainder of this section details our methodology for this experiment, including a discussion of: the corpus, topics and system used (Subsection 3.1); how we generated snippets (Subsection 3.2); the behaviours we logged (Subsection 3.3); how we obtained the opinions of subjects regarding their experience (Subsection 3.4); and further details on our study, including measures taken for quality control (Subsection 3.5).

3.1 Corpus, Search Topics and System

For this experiment, we used the TREC AQUAINT test collection. Using a traditional test collection provided us with the ability to easily evaluate the performance of subjects. The collection contains over one million newspaper articles from the period 1996-2000. Articles were gathered from three newswires: the *Associated Press (AP)*; the *New York Times (NYT)*; and *Xinhua*.

We then selected a total of five topics from the *TREC 2005 Robust Track*, as detailed by Voorhees [50]. The topics selected were: № 341 *(Airport Security)*; № 347 *(Wildlife Extinction)*; № 367 *(Piracy)*; № 408 *(Tropical Storms)*; and № 435 *(Curbing Population Growth)*. We selected topic № 367 as a practice topic so that subjects could familiarise themselves with the system. These topics were chosen based upon evidence from a previous user study with a similar setup, where it was shown that the topics were of similar difficulty [23]. For each subject, the remaining four topics were assigned to an interface (one of ***T0***, ***T1***, ***T2*** or ***T4***) using a Latin-square rotation.

To ground the search tasks, subjects of the experiment were instructed to imagine that they were newspaper reporters, and were required to gather documents to write stories about the provided topics. Subjects were told to find as many relevant documents as they could during the allotted time, which was 10 minutes per topic – hereafter referred to as a *search session*. With the traditional components of a SERP, such as the query box and result summaries present (refer to Figure 3), subjects were instructed to mark documents they considered relevant by clicking on the `'Mark as Relevant'` button within the document view – accessed by clicking on a result summary he or she thought was relevant. Coupled with a two minute period to familiarise themselves with the system (using topic № 367), subjects spent approximately 45-50 minutes undertaking the complete experiment when pre- and post-task surveys were accounted for.

For the underlying search engine, we used the *Whoosh Information Retrieval (IR)* toolkit [3]. We used BM25 as the retrieval algorithm ($b = 0.75$), but with an implicit `AND`ing of query terms to restrict the set of retrieved documents to only those that contained all the query terms provided. This was chosen as most search systems implicitly `AND` terms together.

3.2 Snippet Generation

For interfaces ***T2*** and ***T4***, each result summary presented to the subjects required one or more textual snippets from the corresponding document. These snippet fragments were query-biased [48], and were generated by scoring sentences according to BM25 and selecting fragments from those sentences. Fragments were then extracted from the ordered series of sentences, by identifying query terms within those sentences with a window of 40 characters from either side of the term. Figure 2 provides a complete, rendered example of the result summaries generated by each of the four interfaces. Each result summary contains a document title, a newswire source (acting as a replacement for a document URL), and, if required, one or more textual snippets.

[3]Whoosh can be accessed at https://pypi.python.org/pypi/Whoosh/.

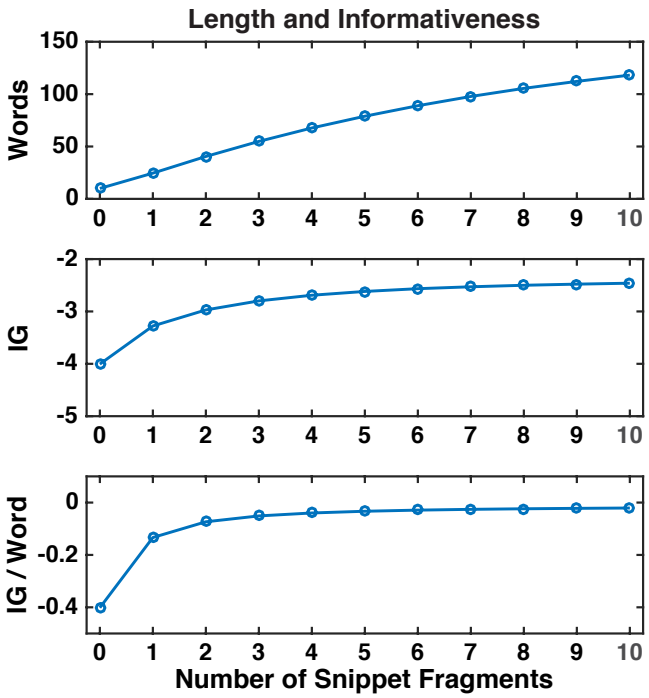

Figure 1: Plots showing the length (in words), informativeness (in information gain, *IG*) and the information gain *(IG)* per word for title, plus 0 to 10 snippets. The closer the value is to zero, the more information that is gained.

3.3 Behaviours Logged

In order for us to address our research questions, our experimental system was required to log a variety of behavioural attributes for each subject as they performed the variety of actions that take place during a search session. Search behaviours were operationalised over three types of measures: *(i)* interaction, *(ii)* performance, and *(iii)* the time spent undertaking various search activities. All behavioural data was extracted from the log data produced by our system, and from the TREC 2005 Robust Track QRELs [50]. All data was recorded with the interface and topic combination used by the subject at the given time.

Interaction measures included the number of queries issued, the number of documents viewed, the number of SERPs viewed, and the greatest depths in the SERPs to which subjects clicked on – and hovered over – result summaries.

Performance measures included a count of the documents marked as relevant by the subject, the number of documents marked that were also TREC relevant – as well as TREC non-relevant, and $P@k$

Figure 2: Examples of the result summaries generated by each of the four interfaces used in this study. The same document above is used – with the circle denoting what interface is being shown (of *T0*, *T1*, *T2* or *T4*). Each of the result summaries consists of a title (in blue, underlined), none, one or more snippet fragments (in black, with fragments separated by ellipses), and a newswire source (in green).

measurements for the performance of the subject's issued queries for a range of rankings.

Time-Based measures included the time spent issuing queries, examining SERPs – as well as examining result summaries[4] – and the time spent examining documents. All of these times added together yielded the total search session time, which elapsed once 10 minutes had been reached.

From this raw data, we could then produce summaries of a search session, producing summarising measures such as the number of documents examined by searchers per query that they issued. We could also calculate from the log data probabilities of interaction, including a given subject's probability of clicking a result summary link, given that it was TREC relevant ($P(C|R)$) or TREC non-relevant ($P(C|N)$) – or the probability of marking a document that was clicked, given it was either TREC relevant ($P(M|R)$) or TREC non-relevant ($P(M|N)$). Actions such as hover depth over result summaries were inferred from the movement of the mouse cursor, which in prior studies has been shown to correlate strongly with the user's gaze on the screen [7, 44].

3.4 Capturing User Experiences

To capture user experiences, we asked subjects to complete both pre- and post-task surveys for each of the four interface conditions.

[4]Result summary times were approximated by dividing the total recorded SERP time by the number of snippets hovered over with the mouse cursor. We believe this is a reasonable assumption to make – the timings of hover events proved to be unreliable due to occasional network latency issues beyond our control.

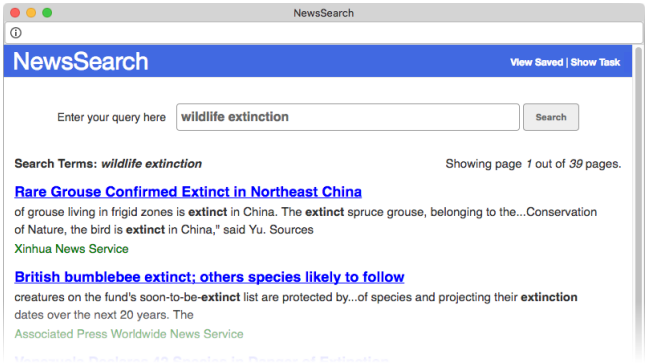

Figure 3: Screenshot of the experimental search interface, showing the SERP view, complete with query box (with query 'wildlife extinction') and the associated result summaries. In this example screenshot, interface *T2* – presenting two snippets per result summary – is shown.

Pre-task surveys consisted of five questions, each of which was on a seven-point Likert scale (7 – *strongly agree* to 1 – *strongly disagree*). Subjects were sought for their opinions on their: *(i)* prior knowledge of the topic; *(ii)* the relevancy of the topic to their lives; *(iii)* their desire to learn about the topic; *(iv)* whether they had searched on this topic before; and *(v)* the perceived difficulty to search for information on the topic.

The same Likert scale was used for post-task surveys, where subjects were asked to judge the following statements: ***(clarity)*** – the result summaries were clear and concise; ***(confidence)*** – the result summaries increased my confidence in my decisions; ***(informativeness)*** – the result summaries were informative; ***(relevance)*** – the results summaries help me judge the relevance of the document; ***(readable)*** – the result summaries were readable; and ***(size)*** – the result summaries were an appropriate size and length.

At the end of the experiment, subjects completed an exit survey. From five questions, they were asked to pick which of the four interfaces was the closest fit to their experience. We sought opinions on what interface: ***(most informative)*** – yielded the most informative result summaries; ***(least helpful)*** – provided the most unhelpful summaries; ***(easiest)*** – provided the easiest to understand summaries; ***(least useful)*** – provided the least useful result summaries; and ***(most preferred)*** – the subject's preferred choice for the tasks that they undertook.

3.5 Crowdsourced Subjects & Quality Control

As highlighted by Zuccon et al. [54], crowdsourcing provides an alternative means for capturing user interactions and search behaviours from traditional lab-based user studies. Greater volumes of data can be obtained from more heterogeneous workers at a lower cost – all within a shorter timeframe. Of course, pitfalls of a crowdsourced approach include the possibility of workers completing tasks as efficiently as possible, or submitting their tasks without performing the requested operations [13]. Despite these issues, it has been shown that there is little difference in the quality between crowdsourced and lab-based studies [54]. Nevertheless, quality control is a major component of a well-executed crowdsourced experiment [5]. Here, we detail our subjects and precautions taken.

The study was run over the *Amazon Mechanical Turk (MTurk)* platform. Workers from the platform performed a single *Human Intelligence Task (HIT)*, which corresponded to the entire experiment. Due to the expected length of completion for the study (45-50 minutes), subjects who completed the study in full were reimbursed for their time with US$9; a typically larger sum (and HIT duration) than most crowdsourced experiments. A total of 60 subjects took part in the experiment, which was run between July and August, 2016. However, seven subjects were omitted due to quality control constraints (see below). In all, of the 53 subjects who satisfied the expected conditions of the experiment, 28 were male, with 25 female. The average age of our subjects was 33.8 years (min = 22; max = 48; $stdev$ = 7.0), with 19 of the subjects possessing a bachelor's degree or higher, and all expressing a high degree of search literacy, with all subjects stating that they conducted at least five searches for information online per week. With 53 subjects, each searching over four topics, this meant a total of 212 search sessions were logged.

We examined extra precautionary measures to ensure the integrity of the log data that was recorded. Precautions were taken from several angles. First, workers were only permitted to begin the experiment on the MTurk platform that: *(i)* were from the United States, and were native English speakers; *(ii)* had a HIT acceptance rate of at least 95%; and *(iii)* had at least 1000 HITs approved. Requiring *(ii)* and *(iii)* reduced the likelihood of recruiting individuals who would not complete the study in a satisfactory manner. Recruits were forewarned about the length of the HIT, which was considerably longer than other crowdsourced experiments.

We also ensured that the computer the subject was attempting the experiment on had a sufficiently large screen resolution (1024x768 or greater) so as to display all of the experimental interface on screen. With the experiment being conducted in a Web browser popup window of a fixed size, we wanted to ensure that all subjects would be able to see the same number of results on a SERP within the popup window's viewport. As the experiment was conducted via a Web browser, we wanted to ensure that only the controls provided by the experimental apparatus were used, meaning that the popup window had all other browser controls disabled to the best of our ability (i.e. history navigation, etc.). The experimental system was tested on several major Web browsers, across different operating systems. This gave us confidence that a similar experience would be had across different system configurations.

We also implemented a series of log post-processing scripts after completion of the study to further identify and capture individuals who did not perform the tasks as instructed. It was from here that we identified the seven subjects that did not complete the search tasks in a satisfactory way – spending less than three of the ten minutes searching. These subjects were excluded from the study, reducing the number of subjects reported from 60 to 53. Finally, results are reported based upon the first 360 seconds as some of the remaining subjects didn't fully use all 600 seconds.

4 RESULTS

Both search behaviour and user experience measures were analysed by each interface. To evaluate these data, ANOVAs were conducted using the interfaces as factors; main effects were examined with $\alpha = 0.05$. Bonferroni tests were used for post-hoc analysis. It should be noted that the error bars as shown in the plots for Figures 4 and 5 refer to the *standard error*.

Table 1: Characters, words and *Information Gain (IG)* across each of the four interface conditions. An ANOVA test reveals significant differences, with follow-up tests (refer to Section 4) showing that each condition is significantly different to others. There are clearly diminishing returns in information gain as snippet length increases. An IG value closer to zero denotes a higher level of IG. In the table, *IG/W.* denotes *IG per word.*

	T0	*T1*	*T2*	*T4*
Words	6.58±0.01*	25.21±0.06*	44.29±0.10*	77.06±0.13*
Chars.	37.37±0.05*	103.29±0.17*	168.36±0.23*	284.78±0.31*
IG	-6.35±0.01*	-3.59±0.00*	-3.00±0.00*	-2.67±0.00*
IG/W.	-1.17±0.00*	-0.18±0.00*	-0.08±0.00*	-0.04±0.00*

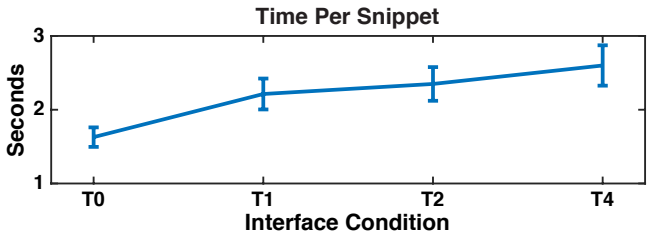

Figure 4: Plot showing the mean time spent examining result summaries across each of the four interfaces examined. Note the increasing mean examination time as the snippet length increases, from *T0→T4*.

To check whether the interfaces were different with respect to snippet length and information gain, we performed an analysis of the observed result summaries. Table 1 summarises the number of words and characters that result summaries contained on average. As expected, the table shows an increasing trend in words and characters as snippet lengths increase. Information gain for each snippet was then calculated using the Kullback-Leibler distance [28] to measure information gain (e.g. relative entropy). Statistical testing showed that the differences between snippet length ($F(3, 208) = 1.2x10^5, p < 0.001$) and information gain ($F(3, 208 = 2.6x10^5, p < 0.001$)) were significant. Follow up tests revealed that this was the case over all four interfaces, indicating that our conditions were different on these dimensions. These findings provide some justification for our choices for the number of snippet fragments present for each interface – a diminishing increase in information gain after four snippets suggested that there wouldn't be much point generating anything longer.

4.1 Search Behaviours

Interactions. Table 2 presents the mean (and standard deviations) of the number of queries issued, the number of SERPs viewed per query, documents clicked per query, and the click depth per query over each of the four interfaces examined. Across the four different interfaces, there were no significant differences reported between

Table 2: Summary table of both interaction and performance measures over each of the four interfaces evaluated. For each measure examined, no significant differences are reported across the four interfaces.

	T0	*T1*	*T2*	*T4*
Number of Queries	3.72± 0.34	3.19± 0.35	3.30± 0.35	3.28± 0.31
Number of SERP Pages per Query	2.87± 0.29	2.69± 0.23	2.43± 0.13	2.40± 0.20
Number of Docs Clicked per Query	4.23± 0.55	4.83± 0.54	5.14± 0.66	4.76± 0.62
Depth per Query	24.47± 2.96	22.87± 2.47	20.02± 1.46	19.40± 2.04
P@10	0.25± 0.02	0.23± 0.02	0.27± 0.02	0.25± 0.03
Number of Documents Marked Relevant	6.68± 0.66	7.00± 0.63	6.49± 0.58	7.60± 0.79
Number of TREC Rels Found	2.58± 0.34	2.28± 0.25	2.47± 0.28	2.66± 0.32
Number of Unjudged Docs Marked Relevant	1.85± 0.32	2.08± 0.29	1.98± 0.24	1.68± 0.32

Table 3: Summary table of times over each of the four interfaces evaluated. Significant differences exist between T0 and T4 (identified by the *, where $\alpha = 0.05$) on a follow-up Bonferroni test.

	T0	*T1*	*T2*	*T4*
Time per Query	8.29± 0.57	7.99± 0.57	9.42± 0.79	8.12± 0.48
Time per Document	17.31± 2.12	22.82± 6.03	17.19± 1.86	18.99± 2.13
Time per Result Summary*	1.63 ± 0.13*	2.21± 0.21	2.35± 0.23	2.60 ± 0.27*

any of these measures. The number of queries issued follows a slight downward trend as the length of result summaries increases (3.72 ± 0.34 for ***T0*** to 3.28 ± 0.31 for ***T4***), as too does the number of SERPs examined, and the number of documents examined per query. The depth to which subjects went to per query however follows a downward trend – as the length of snippets increases, subjects were likely to go to shallower depths when examining result summaries (24.47 ± 2.96 for ***T0*** to 19.4 ± 2.04 for ***T4***).

Interaction probabilities all showed an increasing trend as snippet length increased over the four interfaces, as shown in Table 4. Although no significant differences were observed over the four interfaces and the different probabilities examined, trends across all probabilities show an increase as the snippet length increases. An increase of both the probability of clicking result summaries on the SERP ($P(C)$) and marking the associated documents ($P(M)$) as relevant were observed. When these probabilities are examined in more detail by separating the result summaries clicked and documents marked by their TREC relevancy (through use of TREC QRELs), we see increasing trends for clicking and marking – both for TREC relevant ($P(C|R)$ and $P(M|R)$ for clicking and marking, respectively) and TREC non-relevant documents ($P(C|N)$ and $P(M|N)$). This interesting finding shows that an increase in snippet length does not necessarily improve the accuracy of subjects – simply the likelihood that they would consider documents as relevant.

Performance. Table 2 also reports a series of performance measures over the four conditions, averaged over the four topics examined. We report the mean performance of the queries issued with $P@10$, the number of documents marked relevant, and the number of documents marked relevant that were TREC relevant. Like the interaction measures above, no significant differences were observed over the four interfaces for each of the performance measures examined. The performance of queries issued by subjects was very similar across all four conditions ($P@10 \approx 0.25$), along with the number of documents identified by subjects as relevant (6.49 ± 0.58 for ***T2*** to 7.6 ± 0.79 for ***T4***), and the count of documents marked that were actually TREC relevant (2.28 ± 0.25 for ***T1*** to 2.66 ± 0.32 for ***T4***). We also examined the number of documents marked that were not assessed (unjudged) by the TREC assessors, in case one interface surfaced more novel documents. On average, subjects marked two such documents, but again there was no significant differences between interfaces.

Time-Based Measures. Table 3 reports a series of selected interaction times over each of the four evaluated interfaces. We include: the mean total query time per subject, per interface; the mean time per query; the mean time spent examining documents per query; and the mean time spent examining result summaries per query. No significant differences were found between the mean total query time, the time per query and the time per document. However, a significant difference did exist for the time spent per result summary. A clear upward trend in the time spent examining snippets can be seen in Figure 4 as result summaries progressively got longer, from 1.63 ± 0.13 for ***T0*** to 2.6 ± 0.27 for ***T4***, which was significantly different ($F(3, 208) = 3.6, p = 0.014$). A follow-up Bonferroni test showed that the significant difference existed between ***T0*** and ***T4***. This suggests that as result summary length increases, the amount of time spent examining result summaries also increases (an intuitive result). This also complies with trends observed regarding examination depths. When the length of result summaries increased, subjects were likely to examine result summaries to shallower depths.

Table 4: Table illustrating a summary of interaction probabilities over each of the four interfaces evaluated. Note the increasing trends for each probability from *T0* → *T4* (short to long snippets). Refer to Section 4.1 for an explanation of what each probability represents.

	T0	*T1*	*T2*	*T4*
$P(C)$	0.20± 0.02	0.25± 0.02	0.26± 0.03	0.28± 0.03
$P(C\|R)$	0.28± 0.03	0.34± 0.03	0.35± 0.03	0.40± 0.04
$P(C\|N)$	0.18± 0.02	0.23± 0.02	0.25± 0.03	0.24± 0.03
$P(M)$	0.61± 0.04	0.68± 0.04	0.65± 0.03	0.71± 0.03
$P(M\|R)$	0.66± 0.06	0.69± 0.05	0.67± 0.05	0.66± 0.05
$P(M\|N)$	0.55± 0.04	0.65± 0.04	0.58± 0.04	0.67± 0.04

Table 5: Summary table of the recorded observations for the post-task survey, indicating the preferences of subjects over six criteria and the four interfaces, where $*$ indicates that T0 was significantly different from the other conditions. In the table, *Conf.* represents *Confidence*, *Read.* represents *Readability*, *Inform.* represents *Informativeness*, and *Rel.* represents *Relevancy*.

	T0	T1	T2	T4
Clarity	4.16± 0.27*	5.00± 0.21	5.06± 0.24	5.40± 0.20
Conf.	3.71± 0.26*	4.66± 0.26	4.75± 0.24	5.06± 0.25
Read.	5.18± 0.31*	6.32± 0.17	6.46± 0.14	6.36± 0.14
Inform.	4.20± 0.30*	5.38± 0.24	5.27± 0.24	5.62± 0.20
Rel.	3.84± 0.28*	4.89± 0.25	5.08± 0.24	5.36± 0.20
Size	4.00± 0.31*	4.94± 0.25	5.21± 0.22	5.36± 0.19

Table 6: Table presenting responses from the exit survey completed by subjects. The survey is discussed in Section 3.4.

	T0	T1	T2	T4
Most Informative	1	4	**20**	**29**
Least helpful	**46**	5	1	2
Easiest	4	4	**24**	**22**
Least Useful	**49**	4	0	1
Most Preferred	3	5	**20**	**26**

4.2 User Experience

Task Evaluations. Table 5 presents the mean set of results from subjects across the four interfaces, which were answered upon completion of each search task. The survey questions are detailed in Section 3.4. Using the seven-point Likert scale for their responses (with 7 indicating *strongly agree*, and 1 indicating *strongly disagree*), significant differences were found in all question responses (**clarity** $F(3, 208) = 5.22, p = 0.001$, **confidence** $F(3, 208) = 5.3, p = 0.001$, **readable** $F(3, 208) = 9.25, p < 0.001$, **informative** $F(3, 208) = 5.22, p = 0.001$, **relevance** $F(3, 208) = 6.44, p < 0.001$, and **size** $F(3, 208) = 7.28, p < 0.001$). Follow-up Bonferroni tests however showed that the significant difference existed only between ***T0*** and the remaining three interfaces, ***T1***, ***T2*** and ***T4***. A series of discernible trends can be observed throughout the responses, with subjects regarding longer snippets as more concise, and a higher degree of clarity (4.16±0.27 for ***T0*** to 5.4±0.2 for ***T4***). This perceived clarity also made subjects feel more confident that the longer result summaries helped them make better decisions as to whether they were relevant to the given topic – interaction results presented above however differ from this, where the overall probability of marking documents increased, regardless of the document/topic TREC relevancy judgement. Other notable trends observed from the results included an increase in how informative subjects perceived the result summaries to be – again, with longer summaries proving more informative. Subjects also reported a general increase in satisfaction of the length of the presented result summaries/snippets – although, as mentioned, no significant difference existed between the three interfaces that generated snippets (***T1***, ***T2*** and ***T4***).

System Evaluations. Upon completion of the study, subjects completed the exit survey as detailed in Section 3.4. Responses from the subjects are presented in Table 6. From the results, subjects found result summaries of longer lengths (i.e. those generated by interfaces ***T2*** and ***T4***) to be the most informative, and those generated by ***T0*** – without snippets – to be the least helpful and useful. The longer result summaries were also consistently favoured by subjects, who preferred them over the result summaries generated by interfaces ***T0*** and ***T1***. Subjects also found the result summaries of longer length easier to use to satisfy the given information need.

From the results, it is therefore clear that a majority of subjects preferred longer result summaries to be presented on SERPs, generated by interfaces ***T2*** and ***T4***. Figure 5 provides summary plots, showing general trends across the four interfaces, examining observed interactions and reported experiences.

5 DISCUSSION AND FUTURE WORK

In this paper, we investigated the influence of result summary length on search behaviour and performance. Using the Kullback-Leibler distance [28] as a measure of information gain, we examined result summaries of different lengths, selected a series of snippet lengths where there was a significant difference in information gain between them, which yielded the configurations for our four experimental conditions, ***T0***, ***T1***, ***T2*** and ***T4***. We conducted a crowdsourced user study comprising of 53 subjects, each of whom undertook four search tasks, using each of the four interfaces.

Our work was focused around addressing our two research questions, which explored **RQ1** how the value of information gain (represented by snippet length) affected search behaviour and user experience; and **RQ2** whether information gain affected the decision making ability and accuracy of users. Addressing **RQ1** first in terms of search behaviour, there was little difference – but we did observe the following trends: as summary length increases, participants: issued fewer queries; examined fewer pages; but clicked more documents, i.e. they spent more of their time assessing documents at higher ranks. Second, our results show that in terms

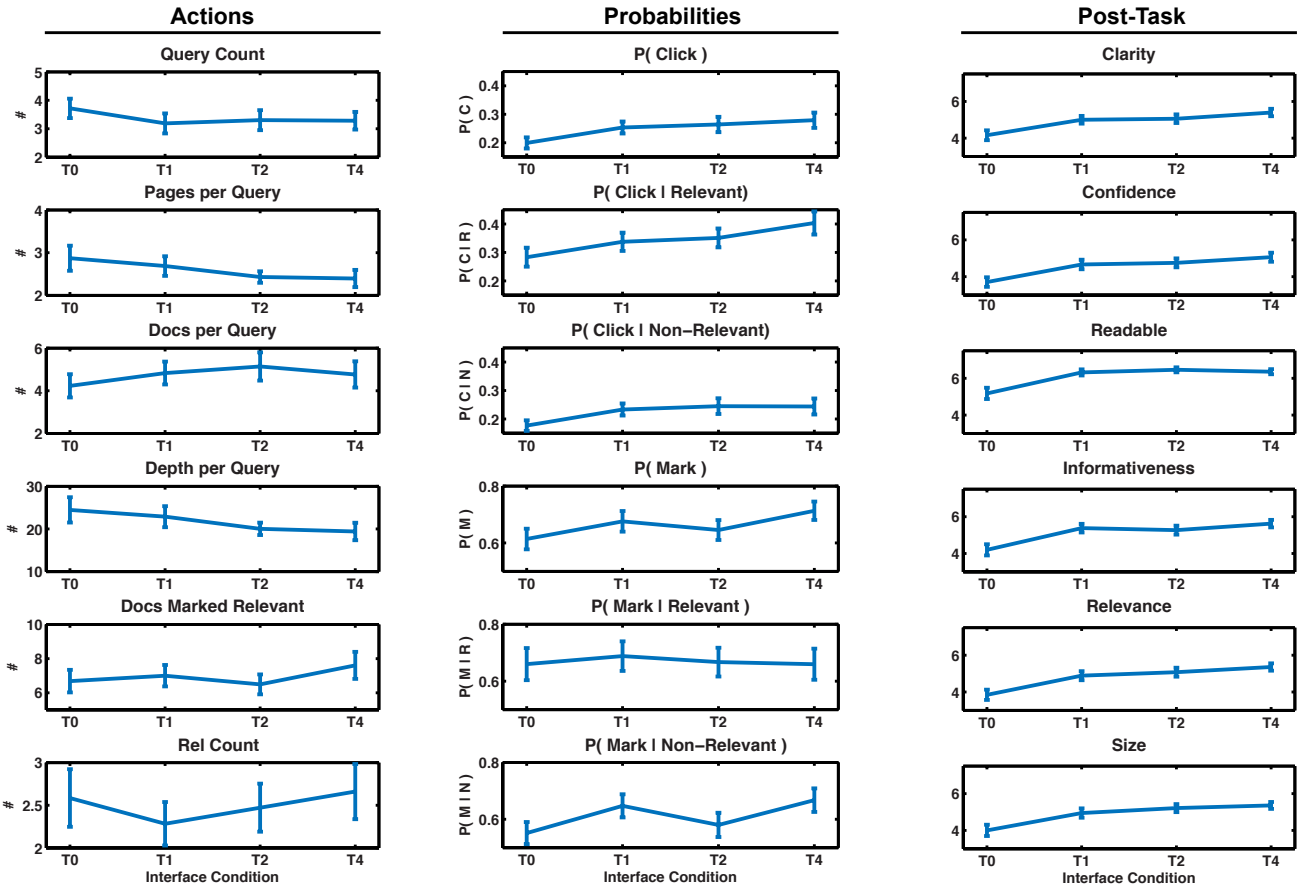

Figure 5: Plots, showing a variety of measures and survey results from subjects across the four interfaces examined. From left to right: actions associated with the subjects' search behaviours and performance; probabilities of interaction; and post-task survey responses, using a seven-point Likert scale (7 - *strongly agree* to 1 - *strongly disagree*).

of experience, subjects broadly preferred longer summaries. The participants felt that longer summaries were more clear, informative, readable – and interestingly – gave them more confidence in their relevance decisions. With respect to **RQ2**, we again observed little difference in subjects' decision making abilities and accuracy between the four interfaces. While subjects perceived longer snippets to help them infer relevance more accurately, our empirical evidence shows otherwise. In fact, it would appear that longer result summaries were more attractive, increasing the information scent of the SERP [53]. This may account for the increase in clicks on the early results, without the benefits, however: accuracy of our subjects did not improve with longer snippets; nor did they find more relevant documents. Increased confidence in the result summaries (from ***T0*** → ***T4***) may have led to a more relaxed approach at marking content as relevant – as can be seen by increasing click and mark probabilities for both relevant and non-relevant content. It is also possible that the *paradox of choice* [37] could play a role in shaping a user's preferences. For example, in the condition with longer result summaries (***T4***), users viewed fewer results/choices than on other conditions. This may have contributed to their feelings of greater satisfaction and increased confidence in their decisions.

These novel findings provide new insights into how users interact with result summaries in terms of their experiences and search behaviours. Previous work had only focused upon task completion times and accuracy of the first result while not considering their experiences (e.g. [11, 19]). Furthermore, these past works were performed in the context of Web search where the goal was to find one document. However, we acknowledge that our work also has limitations. Here, we examined out research questions – with respect to topic search within a news collection – to explore how behaviour and performance changes when searching for multiple relevant documents. It would be interesting to examine this in other search contexts, such as product search, for example. News article titles also can be crafted differently from documents in other domains. Summaries in this domain may perhaps be more important than in other domains, and so the effects and influences are likely to be larger. Furthermore, we only considered how behaviours changed on the desktop, rather than on other devices where users are more likely to be sensitive to such changes (e.g. [25, 27]). For example, during casual leisure search, multiple relevant documents on tablet devices are often found, and so it would be interesting to perform a follow up study in this area.

To conclude, our findings show that longer result summaries, while containing a greater amount of information content, are not necessarily better in terms of decision making – although subjects perceived this to be the case. We also show a positive relationship between the length and informativeness of result summaries and their attractiveness (clickthrough rates). These results show that the experience and perceptions of users – and the actual performance of those users – is different, and when designing interfaces, this needs be taken into account.

Acknowledgments Our thanks to Alastair Maxwell and Stuart Mackie for their comments, the 53 participants of this study, and the anonymous reviewers for their feedback. The lead author is financially supported by the UK Government through the EPSRC, grant №. 1367507.

REFERENCES

[1] M. Ageev, D. Lagun, and E. Agichtein. Improving search result summaries by using searcher behavior data. In *Proc. 35th ACM SIGIR*, pages 13–22, 2013.
[2] H. Ali, F. Scholer, J. A. Thom, and M. Wu. User interaction with novel web search interfaces. In *Proc. 21st OZCHI*, pages 301–304.
[3] L. Azzopardi, D. Kelly, and K. Brennan. How query cost affects search behavior. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in IR*, pages 23–32, 2013.
[4] H. Bast and M. Celikik. Efficient index-based snippet generation. *ACM Trans. Inf. Syst.*, 32(2):6:1–6:24, Apr. 2014.
[5] H. Bota, K. Zhou, and J. M. Jose. Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proc. 1st ACM CHIIR*, pages 131–140, 2016.
[6] H. Chen and S. Dumais. Bringing order to the web: Automatically categorizing search results. In *Proc. 18th ACM CHI*, pages 145–152, 2000.
[7] M. C. Chen, J. R. Anderson, and M. H. Sohn. What can a mouse cursor tell us more?: Correlation of eye/mouse movements on web browsing. In *Proc. 19th ACM CHI Extended Abstracts*, pages 281–282, 2001.
[8] F. Chierichetti, R. Kumar, and P. Raghavan. Optimizing two-dimensional search results presentation. In *Proc. 4th ACM WSDM*, pages 257–266, 2011.
[9] C. L. A. Clarke, E. Agichtein, S. Dumais, and R. W. White. The influence of caption features on clickthrough patterns in web search. In *Proc. 30th ACM SIGIR*, pages 135–142, 2007.
[10] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proc. 1st ACM WSDM*, pages 87–94, 2008.
[11] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proc. 25th ACM CHI*, pages 407–416, 2007.
[12] S. Dumais, E. Cutrell, and H. Chen. Optimizing search by showing results in context. In *Proc. 19th ACM CHI*, pages 277–284, 2001.
[13] H. Feild, R. Jones, R. Miller, R. Nayak, E. Churchill, and E. Velipasaoglu. Logging the search self-efficacy of amazon mechanical turkers. In *Proc. CSE SIGIR Workshop*, pages 27–30, 2010.
[14] J. He, P. Duboue, and J.-Y. Nie. Bridging the gap between intrinsic and perceived relevance in snippet generation. In *Proc. of COLING 2012*, pages 1129–1146, 2012.
[15] M. Hearst. *Search user interfaces.* Cambridge University Press, 2009.
[16] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context.* 2005.
[17] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. 28th ACM SIGIR*, pages 154–161, 2005.
[18] H. Joho and J. M. Jose. A comparative study of the effectiveness of search result presentation on the web. In *Proc. 28th ECIR*, pages 302–313, 2006.
[19] M. Kaisser, M. A. Hearst, and J. B. Lowe. Improving search results quality by customizing summary lengths. In *Proc. 46th ACL*, pages 701–709, 2008.
[20] Y. Kammerer and P. Gerjets. How the interface design influences users' spontaneous trustworthiness evaluations of web search results: comparing a list and a grid interface. In *Proc. of the Symp. on Eye-Tracking Research & Applications*, pages 299–306, 2010.
[21] T. Kanungo and D. Orr. Predicting the readability of short web summaries. In *Proc. 2nd ACM WSDM*, pages 202–211, 2009.
[22] D. Kelly and L. Azzopardi. How many results per page?: A study of serp size, search behavior and user experience. In *Proc. 38th ACM SIGIR*, pages 183–192, 2015.
[23] D. Kelly, K. Gyllstrom, and E. W. Bailey. A comparison of query and term suggestion features for interactive searching. In *Proc. 32nd ACM SIGIR*, pages 371–378, 2009.
[24] J. Kim, P. Thomas, R. Sankaranarayana, and T. Gedeon. Comparing scanning behaviour in web search on small and large screens. In *Proc. 17th ADCS*, pages 25–30, 2012.
[25] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. Eye-tracking analysis of user behavior and performance in web search on large and small screens. *J. of the Assoc. for Information Science and Technology*, 2014.
[26] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. Pagination versus scrolling in mobile web search. In *Proc. 25th ACM CIKM*, pages 751–760, 2016.
[27] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. What snippet size is needed in mobile web search? In *Proc. 2nd ACM CHIIR*, pages 97–106, 2017.
[28] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.
[29] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proc. 18th ACM SIGIR*, pages 68–73, 1995.
[30] T. Landauer, D. Egan, J. Remde, M. Lesk, C. Lochbaum, and D. Ketchum. Enhancing the usability of text through computer delivery and formative evaluation: the superbook project. *Hypertext: A psychological perspective*, pages 71–136, 1993.
[31] L. Leal-Bando, F. Scholer, and A. Turpin. Query-biased summary generation assisted by query expansion. *J. Assoc. for Info. Sci. and Tech.*, 66(5):961–979, 2015.
[32] Q. Li and Y. P. Chen. Personalized text snippet extraction using statistical language models. *Pattern Recogn.*, 43(1):378–386, Jan. 2010.
[33] G. Linden. *Marissa mayer at web 2.0*, November 2006. *http://glinden.blogspot.com/2006/11/marissa-mayer-at-web-20.html*.
[34] Marcos, M-C. and Gavin, F. and Arapakis, I. Effect of snippets on user experience in web search. In *Proc. 16th Intl. Conf. on HCI*, pages 47:1–47:8, 2015.
[35] A. Muralidharan, Z. Gyongyi, and E. Chi. Social annotations in web search. In *Proc. 30th ACM CHI*, pages 1085–1094, 2012.
[36] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proc. 22nd WWW*, pages 953–964, 2013.
[37] A. Oulasvirta, J. Hukkinen, and B. Schwartz. When more is less: The paradox of choice in search engine use. In *Proc. 32nd ACM SIGIR*, pages 516–523, 2009.
[38] T. Paek, S. Dumais, and R. Logan. Wavelens: A new view onto internet search results. In *Proc. 22nd ACM CHI*, pages 727–734, 2004.
[39] J. Pedersen, D. Cutting, J. Tukey, et al. Snippet search: A single phrase approach to text access. In *Proc. 1991 Joint Statistical Meetings*, 1991.
[40] M. L. Resnick, C. Maldonado, J. M. Santos, and R. Lergier. Modeling on-line search behavior using alternative output structures. In *Proc. Human Factors and Ergonomics Soc. Annual Meeting*, volume 45, pages 1166–1170, 2001.
[41] D. E. Rose, D. Orr, and R. G. P. Kantamneni. Summary attributes and perceived search quality. In *Proc. 16th WWW*, pages 1201–1202, 2007.
[42] D. Savenkov, P. Braslavski, and M. Lebedev. Search snippet evaluation at yandex: lessons learned and future directions. *Multilingual & Multimodal Information Access Evaluation*, pages 14–25, 2011.
[43] B. Schwartz. *The Paradox of Choice: Why More Is Less.* Harper Perennial, 2005.
[44] M. Smucker, X. Guo, and A. Toulis. Mouse movement during relevance judging: Implications for determining user attention. In *Proc. 37th ACM SIGIR*, pages 979–982, 2014.
[45] N. V. Spirin, A. S. Kotov, K. G. Karahalios, V. Mladenov, and P. A. Izhutov. A comparative study of query-biased and non-redundant snippets for structured search on mobile devices. In *Proc. 25th ACM CIKM*, pages 2389–2394, 2016.
[46] K. M. Svore, J. Teevan, S. T. Dumais, and A. Kulkarni. Creating temporally dynamic web search snippets. In *Proc. 35th ACM SIGIR*, pages 1045–1046, 2012.
[47] J. Teevan, E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. André, and C. Hu. Visual snippets: Summarizing web pages for search and revisitation. In *Proc. 27th ACM CHI*, pages 2023–2032, 2009.
[48] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proc. 21st ACM SIGIR*, pages 2–10, 1998.
[49] A. Turpin, Y. Tsegay, D. Hawking, and H. E. Williams. Fast generation of result snippets in web search. In *Proc. 30th ACM SIGIR*, pages 127–134, 2007.
[50] E. M. Voorhees. Overview of the trec 2005 robust track. In *Proc. TREC-14*, 2006.
[51] R. W. White, J. M. Jose, and I. Ruthven. A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Info. Processing & Management*, 39(5):707–733, 2003.
[52] A. Woodruff, R. Rosenholtz, J. B. Morrison, A. Faulring, and P. Pirolli. A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for web search tasks. *J. Am. Soc. Inf. Sci. Technol.*, 53(2):172–185, 2002.
[53] Wu, W-C. and Kelly, D. and Sud, A. Using information scent and need for cognition to understand online search behavior. In *Proceedings of the 37th International ACM SIGIR Conference*, SIGIR '14, pages 557–566, 2014.
[54] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J. Jose, and L. Azzopardi. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information Retrieval*, 16(2):267–305, 2013.