The Impact of Summaries: What Makes a User Click?

Khairun Nisa Fachry[1] Jaap Kamps[1,2] Junte Zhang[1]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Faculty of Science, University of Amsterdam

ABSTRACT

Modern retrieval systems are in fact two-tier systems in which a user first views summaries of the results in a hit-list, and only when she decides to "click," the full result document is consulted. Standard information retrieval evaluation ignores the crucial summary step, and directly evaluates in terms of the relevance of the resulting document. In this paper, we investigate the impact of the result summaries on the user's decision to click or not to click. Specifically, we want to find out both what information in the summary triggers a positive selection decision to view a result, and what information triggers a negative selection decision. We use a special document genre, archival finding aids, where results have a complex document structure and currently available systems experiment with structured summaries having both static elements (like the title and a manually compiled abstract by an archivist) and query-biased snippets (showing the matching keywords in context). We conducted an experiment in which we asked test persons to explicitly mark the parts of summaries that trigger a selection decision, and asked them to explain further (i.e. why and how). The results from this user study indicate the importance of sufficient context in the summary. Selection decisions were primarily based on the static elements: the title and abstract of the document. This may be a result of the completeness and coherence of the information in these elements, although also the length played a clear role. A whole paragraph (as in the abstract) triggered a decision more frequently than a short sentence (as in the title) or an incomplete sentence (as in the query-biased snippets).

1. INTRODUCTION

Modern information retrieval systems are in fact two-tier systems. Imagine a scenario about a user with a particular information need. In the first stage, she will inspect summaries of the results in a hit-list and tries to assess which results potentially satisfy her information need. Based on a promising summary, she may decide to "click" and enter a second stage in which she consults the full result document looking for useful information given her information seeking need. In these two-tier systems, the summaries on the hit-list play a crucial role and act as a filter: only when the summary is deemed adequate, the result is inspected.

Standard information retrieval evaluation ignores this crucial summary step and directly evaluates in terms of the relevance of the resulting document. Turpin et al. [20], in their study of including summaries in system evaluation, revealed that summaries need to be evaluated in addition to the document when constructing a test collection. In their experiment, in which users were asked to provide relevance assessments of both summaries and documents, 14% of the highly relevant and 31% of relevant documents were never examined by the users because the summary was judged irrelevant. This shows that the document summary presented by a retrieval system does not always accurately reflect the document content. Since summaries evaluation is the first selection moment for the users, this could results in users missing out some relevant documents.

In this paper, our main aim is to investigate the impact of the summaries of documents on a user's decision to either click or not. Specifically, we investigate the following two research questions:

1. What information in the summaries triggers a positive selection decision to view a result and what information triggers a negative selection decision?

2. Why and how does this influence the decision to click or not to click?

We research these questions for a special document genre, archival finding aids. Archival finding aids are descriptions of archival collections. Since archival collections can be huge, their descriptions may cover 100s of pages. Archival descriptions are structured in a hierarchical way, from general (an overview description of the whole collection) to the specific (a description at the lowest level, most commonly file or item level). Archival finding aids are increasingly encoded in an Extended Markup Language [XML, 21] format called Encoded Archival Description [EAD, 7], which is the *de facto* standard. Archival descriptions are an interesting special case for result summarization, since the documents themselves are long in content and complexly structured. In particular, the descriptions contain various fields such as the title and a human-generated abstract (summary of the whole collection). In addition, short teasers or snippets showing keywords in context can be derived from the textual content of the document. Tombros and Sanderson [19] demonstrated that these can significantly improve both the accuracy and speed of user relevance judgments. Hence, the archival descriptions allow us to experiment with both static elements and query-biased snippets in the result summaries.

We conducted an experiment in which we asked test persons to explicitly mark the parts of summaries that trigger

DIR-2010 January 25, 2010, Nijmegen, the Netherlands.

selection decisions, and asked them to explain why this information triggered their decision. To answer our first research question, we look at two outcomes of a selection decision, i.e. a positive and negative selection decision. For each decision, we count the part of the summaries marked by the test persons and this results in quantitative data. To answer our second research question, we look at the qualitative data on why and how the information triggered the decision.

The remainder of the paper is organized as follows. In Section 2, we describe related work on selection decisions. In Section 3, we describe the methodology of the user study. In Section 4, we describe the result of the user study. Finally, in Section 5 we discuss the results and draw our conclusions.

2. RELATED WORK

In this section, we will discuss related work on selection decisions in literature search, in XML retrieval, and in archival access.

2.1 Selection Decision in Literature Search

A selection decision is based on the (assumed) relevance of a result. The concept of relevance is fundamental in information retrieval, and has attracted continual interest. See Saracevic [17] for the classic framework and overview of early work. More recent contributions include the concept of external (situational) relevance Schamber et al. [18]. Research on selection decision in literature search focused on the ability of users to predict the relevance of documents based on the documents' summaries. For example, Park [16] studied the criteria employed by 10 academic users who were asked to make a selection decision when presented with lists of bibliographic citations. Park categorized user-based characteristics of citation selections as internal, external and problem context. Internal context category describes users perception that are linked at the citation level, for example users perception of author or journals. The external context presents the context stem from individual's search and current research. And lastly, the problem context illustrates why and how the user employs information to construct and solve the information problem. Barry [1] studied the criteria employed by academics to evaluate the representation and the (printed) full text document that has been retrieved specifically for each user's information need. Barry focused on the categorization of user-defined relevance criteria beyond topicality. Her study indicated that the criteria employed by users included tangible characteristics of documents, the provision of references to other sources of information, subjective qualities, and situational factors.

2.2 Selection Decision in XML Retrieval

Selection decision relies heavily on which elements are presented in the summary. XML allows the retrieval and presentation of any individual element in the summary. The presentation of structural text retrieval results is still an open question [10]. Previous user studies have shown the benefits of using XML markup in the retrieval and subsequently information access. Larsen et al. [11] studied whether making elements retrievable is worth the added effort. They found that users find elements useful for their searching tasks, and that they find a lot of the relevant information in specific elements rather than full documents. Betsi et al. [2] found that users liked the idea of being able to gain access directly to the document parts that they were interested in, however, expected the retrieved components to be accompanied by the documents that contain them. Users in this study felt rather uncertain if elements with no contextual information were retrieved. Malik et al. [14] investigated users' behaviors while interacting with XML documents. A result from this study showed that users also appreciated the presentation of XML document structure which is providing context. In terms of elements presented in the summary, only title and authors of documents were displayed as elements summaries in this experiment. As a result, 30 out of 88 test persons in their study commented on the insufficient clues for making a selection decision.

2.3 Selection Decision in Archival Access

Research in users interacting with online finding aids is still in its infancy. Duff and Stoyanova [6] studied elements that were important for users who were looking for archival materials for their research. The following elements were considered to be important: title, information about the creator of the records, call number, scope and content, summary information about content of finding aids, notes of a finding aid, the availability of the finding aids, extent of the material/related records, and types of material/physical description. Since the study mainly focused on the archival display features, they did not elaborate further on the relevance criteria such as user's previous experience and knowledge, sources of information within the environment, and so on.

Duff and Johnson [4] interviewed ten historians focusing on their information seeking behavior. They reported that the historians closely examined finding aids in order to better acquire the sense of the whole collection. They also found that many historians appreciated the addition of summary information about the content of finding aids. This information helped them in their relevance judgments of possible search results. Duff and Johnson [5] studied how genealogists search for information in the archives. Genealogists seek records that contain information about names of people, which might be located in different records. Both studies emphasized the importance of showing the relationships between records (context) and having an overview of the records.

Presenting a list of finding aid elements as a summary has remained a popular method of presenting search results. However, there was no agreement on which elements can be used as summary of the finding aids systems [9]. This can cause a problem since elements used as summary in the search results can vary significantly from one finding aid system to another. For the users, the inconsistency can be very confusing once they interact with several finding aid systems. Presenting why a hit is relevant is strongly related to what to retrieve for the summary. Lee [12, 13] conducted usability studies at the Online Archive of California [OAC, 15], comparing two types of summaries (to which she referred in her study as citations). Long format citation presented title, contributing institution, description (from abstract) and search terms found in information. Short format citation was a Google-like format which presents title and search term. Many of her users preferred long format citation over short format citation, since the long format present more context of the whole collection.

Context of the whole collection was also an issue reported by Fachry et al. [8] where they conducted a study focused

on the effects of presenting context of the whole collection in the hit-list. They conducted a user study were they compare three system: a system that would return the whole fonds[1] (collection level), a system that only returns the individual archival materials (item level) and a system that returns archival material in context (individual items grouped within the same collection). In the first and second systems, the context was omitted, and using this comparison they examined the effects of presenting context in the hit-list. Although the user study showed that the archival material in context system was not optimal, the users had a preference for the third system. The users liked the concept of retrieving archival material in their original context, with users indicating that the system assisted them in assessing relevancy, navigation and direct access to relevant parts of the finding aids.

3. METHODOLOGY

In this section, we discuss the methodology of the user study, specifically we reason our choices of test persons, tasks, summaries, and protocol of the experiment.

3.1 Test persons

The target population of the study included test persons who were novice and expert in searching for archival materials. Although we could elicit more detailed feedback from expert test persons, in this study we also recruited novice test persons, who had no or little experience with archives. They represent a large potential user population for online historical search. In terms of individual differences of test persons, we carefully registered the domain knowledge and archival experiences.

3.2 Tasks

Another important consideration in our study was the tasks. We focus on locating the archival collections of relevance to a given task. In our case, a very specific task as looking for a specific folder number which has a certain topic would be less appropriate. Our interest is in the step before choosing a specific item, where users are presented with a list of results. We prepared four different simulated tasks. The tasks were designed based on the following considerations:

- Tasks were open-ended, requiring test persons to read more than a single summary in order to complete a task.
- The complexity of the tasks was controlled in a way that they were highly similar.
- Each task included background motivation for the search and sufficient information to decide upon the relevance of the viewed summary.

Simulated tasks are presented in the Appendix A. An example of a simulated task is the following:

> *You are interested in the history of slavery in the 18th and 19th centuries. For your history assignment, you are planning to write an essay about anti-slavery movement of that period. To get data for your essay, you are doing research about people who were involved in the anti-slavery movement, who they were and in what way they promoted the anti-slavery movement. Using the digital inventory of the OAC, you would like to check out which archives contain interesting pieces for your research. Depending upon these findings, you should assess whether to visit the archives for your research is worth your time and effort.*

Each test person was assigned two tasks by the experimenters. The order of presentation of the tasks was rotated across test persons. For each task, the test person had to inspect a list of result summaries and decide whether they would view or not to view the result.

3.3 Summaries

In order to operationalize our research questions, we needed "ideal" summaries that contain all potentially useful information. We adopt the summaries used in Online Archive of California [OAC, 15] because they combine both static and query-biased elements in their hit-list. Figure 1 shows a response of OAC finding aid system in relation to the query "Golden Gate Bridge."

We selected ten summaries for each search task. All summaries were prepared by the experimenters. The selection of summaries were based on the following category:

- The selected summaries had a variety of relevance degree to the search task.
- There was a variety of creators. Creators included persons or corporations.
- All summaries were of fonds-level collections that may or may not include series.

Each summary consisted of the following elements:

1. Collection Title, containing the Creator and Title elements
2. Contributing Institution, containing the Repository element
3. Collection Dates, containing the Dates element
4. Items Online, containing the availability and the amount of items online
5. Summary, containing the Abstract element. To avoid confusion between summary element and summary as a whole, the OAC's summary element is referred to as abstract element in this paper.
6. Search term in context. Query-biased summaries/snippets where test persons could see the sentences in which the query terms appeared in the finding aids.

Another methodological consideration was whether to use paper (printed) or digital summary. Summaries printed on paper were chosen rather than digital summaries because:

- A paper summary was an appropriate and sufficient medium to answer our research questions. We were interested to know the contribution of each elements

[1]An archival fonds is all material produced and/or accumulated and used by a person, family or organization over time.

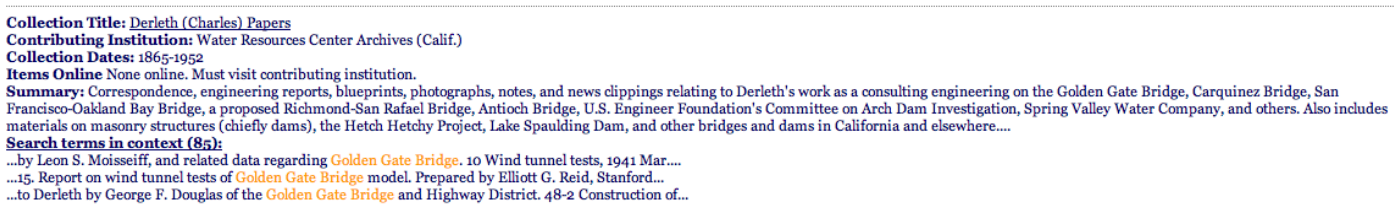

Figure 1: An archival finding aids summary from OAC site (image captured in May 2009)

in test persons' selection decision. Paper summary allowed test persons to easily mark which elements helped them in the selection decision and make notes on how the information in each element helped them in selection decision.

- Paper summary provided ready-transcribed data, the text from test persons' notes can directly be analyzed.

3.4 Protocol

The experiment was designed as follows:

1. Introduction to the experiment and training session.
2. Pre-experimental session in order to collect demographic data of the test persons.
3. Search session I: Judging Summaries. Test person performs the first simulated task, and reviews ten summaries. For each summary, test persons were instructed to:
   (a) examine the summary;
   (b) highlight any portion of the summary that prompted a reaction to pursue the full finding aid;
   (c) for each highlighted portion, comment on the reason to highlight the portion;
   (d) underline any portion of the summary that prompted a reaction not to pursue the full finding aid;
   (e) for each underlined portion, comment on the reason to underline the portion;
   (f) judge the summary as a whole, decide whether to view the full finding aid or not; and
   (g) comment on the reason of the selection decision.
4. Search session II: same as step 2 with a different simulated task.

4. EXPERIMENTAL RESULTS

In this section, we discuss the results of our experiment: the demographics, the elements of the summaries that prompted a positive or negative selection decision, and the motivation behind the choices.

4.1 Demographics data

The total number of 18 test persons (11 male, and 7 female) participated in this study, aged 28–57. All but 5 test persons hold degrees beyond the college (university) level. All test persons were computer-literate with computer experience between 5–15 years. This minimized the possibilities for test persons to find difficulties due to unfamiliarity with common aspects of online navigation.

It is important to emphasize that test persons for this study were carefully registered in terms of their experience with archives. Twelve test persons were recruited from the archive and they all had substantial experience with archives. Ten of the 12 test persons received archival education or training. Out of these 12 test persons, 6 were archivists, 4 were reading room assistants, 1 was a senior adviser and 1 was an ICT manager in an archive. In addition, a thirteenth test person was an amateur-genealogist.

In terms of test persons' experience with archives, 14 test persons had previously conducted historical research (this includes all test persons who were recruited from the archives). When asked about test persons familiarity with archival terminologies, 15 were familiar with archival terminologies in English (this includes all test persons who had conducted historical research). Accordingly, all test persons who were familiar with archival terminologies had visited an archival institution and consulted archival finding aids. However, only 14 of them have visited an archive's site and consulted online finding aids. Since we were using summaries from the OAC, we asked the test persons if they had ever visited the website of OAC. We found out that only 2 test persons had previously visited the OAC website.

4.2 Selection Decision

Table 1 presents test persons' decision in terms of the number of finding aids that they would like to view or not. First, we look at test persons' selection decisions over all tasks. Nine test persons performed simulated task 1, 9 persons did simulated task 2, 10 persons did simulated task 3 and 8 person did simulated task 4. Thus, in total, 36 search sessions were conducted. A total of 360 summaries were examined, since each test person conducted 2 tasks and each task consisted of 10 summaries.

First, we count the number of positive or negative selection decisions based on the test persons' decision to view or not to view a finding aids (see Section 3.4, protocol item 3f). Of the total summaries, test persons decided to view 196 finding aids and not to view 164 finding aids. Thus for each task, a test person decided to view and thus select on average 5.44 finding aids and not to view 4.56 finding aids. The relatively balanced number of view and not view decisions gave us enough feedback to investigate further on processes involved in arriving at a "view" and "not view" selection decision.

Next, we broke down the tasks, and we look at test persons selection decision per task. For each task, did the test persons make the same selection decision? In other words, was there consistency between view or not view decision for each summary within the tasks? For each task, there were 10 cases representing 10 summaries presented to the test persons, 2 categories either 1 (for positive decision) or 0 (for

Table 1: Users' selection decision per task

	Number of Search Sessions	Number of Summaries	Views #	Views %	No views #	No views %	Agreement
Task 1	9	90	43	48	47	52	0.69
Task 2	9	90	49	54	41	46	0.71
Task 3	10	100	56	56	44	44	0.55
Task 4	8	80	48	60	32	40	0.61
Total	36	360	196	54	164	46	

negative decision), and a variety number of raters depending on the number of test person that performed the search task (Task 1=9, Task 2=9, Task 3=10 and Task 4=8). Looking at the agreement for each task using the Kappa statistic [3], the consistency between test persons was substantial for task 1, task 2 and task 4 with K=0.69, K=0.71 and K=0.61, respectively. A moderate consistency was shown for task 3 with K=0.55.

4.2.1 Elements contributing to selection decision

We now focus on processes involved in arriving at a "view" and "not view" selection decision. When presented with summaries of results in a hit-list, what information in the summary trigger a positive selection decision to view a result? Table 2 presents elements of a summary contributed to test persons' decision to view a finding aid which we gathered from elements that were highlighted by test persons when they decided to view a finding aid (see Section 3.4, protocol item 3b). In total, the test persons highlighted 443 elements. On average, for each summary, test persons highlighted 2.26 elements (443 highlighted elements/196 view decisions). The elements abstract, title and snippets came first, second and third, followed by elements dates, item online and contributing institution. Out of all finding aids that the test persons viewed (n=196), abstract element contributed the most to a view decision. Test persons highlighted 147 abstract elements (or 75% of what was viewed). Following the abstract element were title and snippets elements with 103 titles (or 53% of what was viewed) and 101 snippets (or 52% of what was viewed). Furthermore, test persons highlighted 54 date elements (or 28% of what was viewed), 35 item online elements (or 18% of what was viewed) and 3 contributing institution (or 2% of what was viewed).

When presented with summaries of results in a hit-list, what information in the summary trigger a negative selection decision not to view a result? Table 3 presents elements of a summary contributed to test persons' decision not to view a finding aid which we gathered from elements that were underlined by test persons when they not viewed a finding aid (see Section 3.4, protocol item 3d). In total, the test persons underlined 241 elements. On average, for each summary, test persons underlined 1.47 elements (241 underlined elements/164 not view decisions). The elements abstract, title and date elements came first, second and third, respectively, followed by item online element, snippets and contributing institution. As with the view decision, out of all finding aids that were regarded as irrelevant by the test persons (n=164), the abstract element contributed the most to a not view decision. Test persons underlined 96 abstracts (or 59% of what was not viewed). Following the abstract element were the dates and item online elements with user underlined 58 title elements (or 35% of what was not viewed) and 36 date elements (or 22% of what was not viewed). Furthermore, test persons underlined 34 item online elements (or 21% of what was not viewed), 16 snippets (or 10% of what was not viewed) and 1 contributing institution element (or 1% of what was not viewed).

Comparing each elements marked (either highlighted or underlined) by the test persons in Tables 2 and 3, there were two interesting findings. First, we can see that the number of elements marked were higher when test persons decided to view a finding aid (n=2.26 elements per summary) compare to when test persons decided not to view a finding aid (n=1.47 elements per summary).

Another interesting finding was the number of snippets marked. We can see that when test persons decided to view a finding aid, 52% snippets were highlighted. While when test persons decided not to view a finding aid, only 10% snippets were underlined. A plausible reason why this happens is that test persons were first reading the general overview of the finding aid (title and/or abstract element). Once they thought the finding aid was relevant, test persons then went further to the snippets and highlighted the terms they found there. In some cases, if test persons did not find the title or abstract elements relevant to their search task, they did not go further to see the snippets part of the summary. It is also worth noting that many of the snippets were incomplete and too short to judge the relevancy of the document. The snippets were useful to see that the query terms appeared in the finding aids, but the information in the snippets was too little to interpret. This could explain why snippets did not contribute much to the test persons' negative selection decisions.

4.3 Motivation for the selection decisions

We go further on how the individual elements contributed to selection decisions. To answer this question, we focus on how test persons interpreted and used each elements presented as summaries.

4.3.1 Users' assessment of the elements

During the summary judgment phase in the experiment, we asked test persons to comment on the reason that makes them highlight/underline the elements of the summary (see Section 3.4, protocol items 3c and 3e). The result presented in the following is categorized per elements presented in the summary. Our interpretation of factors contributing to selection decision is presented *in italic* and test persons comments are presented "within brackets."

Title In many cases, the title provided *topical relevance*: "Collection title indicates relevance, even without reading the summary I know that there will be a LOT!" On the other hand, the title can also be a reason to reject a summary due

Table 2: Elements trigger a "view" selection decision

	Title		Institution		Dates		Item Online		Abstract		Snippet	
	#	%	#	%	#	%	#	%	#	%	#	%
Task 1	24	56	0	0	6	14	9	21	34	79	22	51
Task 2	36	73	2	4	18	37	22	45	33	67	19	39
Task 3	25	45	0	0	12	21	4	7	46	82	27	48
Task 4	18	38	1	2	18	38	0	0	34	71	33	69
Total	103	53	3	2	54	28	35	18	147	75	101	52

Table 3: Elements trigger a "not view" selection decision

	Title		Institution		Dates		Item Online		Abstract		Snippet	
	#	%	#	%	#	%	#	%	#	%	#	%
Task 1	11	23	0	0	3	6	10	21	38	81	3	6
Task 2	18	44	1	2	13	32	7	17	12	29	4	10
Task 3	10	23	0	0	15	34	11	25	25	57	2	5
Task 4	19	59	0	0	5	16	6	19	21	66	7	22
Total	58	35	1	1	36	22	34	21	96	59	16	10

to its irrelevancy to the information need: "Title implies that the pictures are about the camp and not about the buildings." The *readability* was an important reason for an element not to trigger a selection decision: "The title does not tell me anything." In this case, the title could be too short or mentioned the creator of the collection who was unfamiliar to the test persons. When this was the case, test persons read the other elements or immediately rejected the summary. The title also provided information about the *type of item* available in the collection: "Scrapbook with only pictures of earthquake." For several test persons, a scrapbook was not relevant to them. As mentioned by one test person: "I need written material for my essay, because I do not want to write about the interpretation of the images."

We could also see that the title gave information about the *author* who collected the documents: "This archive was initiated by a state commission, who should do a very thorough work." In this case, the title gave a positive indication since the author seemed to collect reliable materials. While in another case, the author indication in the title could also be an indication to reject a summary: "A very small collection of snapshot made by an unknown individual: just do not know what the photographs are about and are likely not very specific" Another criteria interpreted from the title was the *specificity or broadness* of the collection: "This collection is too broad in subject matter." In this case, the collection was rejected due to its broadness of the topic area.

Dates The dates were important to show the *time period*: "This is excellent for visiting ... over 16 years on the California proposition." Another interesting finding was the dates gave interpretation of *what the collection contains*: "Long period, probably a lot of material about other topics." In this case, since the date period was too long, he thought that the collection would be too broad and contained many other topics (including not relevant topics). The dates also gave indication of the *recency* of the collection: "The dates are too recent."

Items online The availability of online item gave indication of *effort* that test persons needed to spend: "This archive contains online items, which means I can quickly look at the material first before I decide if I need to visit the institute to view the entire collection." The online information also gave indication of *time* that was needed to see the whole collection: "Too much! (referring to 7,000 pages of text)." In this case, the user decided to reject the summary because the amount of text available gave him a clue that he needed to spend a lot of time to read the collection.

Abstract Abstract provided background information such as the time period covered and a brief history of the organization or person who created the records. Abstract elements were most frequently by the test persons. The main reason why abstract was important because it provided the *overview of the whole collection* : "Though this is a very broad collection because of its scope on the African American, it does hold valuable information. First of all on the movement in general and secondly about some of the people involved." Another example showed how a user interpreted the overview of the collection through the abstract. In this case, the test persons rejected the summary because his information need only appeared in some part of the collection: "The summary mentioned that archive focuses mainly on legislation which are not the focus of my research. Though it states includes "some" material on education project, it is not enough for me to view it."

The *type of document* was also shown in abstract: "The summary does not say what these "letters" are about? Although they pertain to the Gold Rush, it is unclear to me whether these are personal letters containing interesting fact about gold seeker's life style or about something else. Too vague." Not only in title element, abstract element also showed indication of *specificity/broadness* of a collection: "The summary was very specific and detailed. It tells me exactly what I can expect from this letter."

From the summary, test persons also predicted the *time and effort* they needed to spend in reading the records: "It is interesting, 100 relevant pages, it is not online, but I know it is one item (a book with 100 pages), it would depend on time." An example where a test person rejected a summary due to *time/effort*: "Description does not indicate that research would be profitable compared to time consumption."

Another selection decision factor is *novelty/new information*: "Personal archive and different type of media, not only governmental archives." In this case, test persons decided to view the finding aid because it could potentially give a new

information to his research. Another important point was the *originality* of the records: "Letters are primary source." In many cases, test persons would like to see the original source of document, not the result that other people have produced: "Scientific info, I would prefer to read original document. Books I can read in the library, I do not need to go to an archive."

Authorship was another factor why abstract was important. Information in the abstract explained the authorship of the records: "The letters might contains personal experience since he wrote to his mother. I expect that the son is writing a long letter with a lot of information." When a record was authored not by the source, the record could be rejected by the test persons: "It is the son's interpretation of his father's life. Probably biased." Especially in one of the tasks when test persons' task was to explore the life of the gold seekers, the originality of the document and the authorship were important selection decision factors. Another selection decision factor was *the types of item*: "Correspondence is interesting. It may give his (Atkin's) personal points of view." This factor was also related to the authorship of the records.

Snippets Snippets were mainly used to indicate *relevance*: "The terms education and tobacco trigger me to have a look." Test persons also looked at the *specificity* of the item in the snippets: "The search term indicate that this professor in history did research himself in this topic..." In this case, the specific information of the item presented in the snippets, was helpful because it provided detailed information that was not shown in other elements. Another important selection decision factor was the test persons' *ability to understand* the snippets: "This snippet does not tell me anything." Often the snippets were too short or repeated in previous elements in the summary. Unavoidably, the length of the snippets influenced our result in terms of the importance of the snippets in supporting test persons to make a selection decision.

5. DISCUSSION AND CONCLUSIONS

In this paper, we investigated the impact of the result summaries on user's decision to either click or not. We researched this question for a special document genre, archival finding aids, where results have a complex document structure and currently available systems experiment with structured summaries having both static and query-biased elements. Static summary elements contains contextual information about the entire collection. Query-biased summary snippets are selectively extracted on the basis of its relation to the searcher's query. The summaries used for our study consist of five static elements (collection title, contributing institution, collection date, items online, and abstract) and multiple query-biased elements (showing keywords in context) per result.

Our first research question was: What information in the summaries triggers a positive selection decision to view a result and what information triggers a negative selection decision? In general, test persons made a selection decision in two steps. First step of selection decision was assessing the general overview of the finding aid to understand what the collection was about. They assessed this by assessing the static elements: title and abstract of the document. Both in the case of a positive decision to view a document, as well as for a negative decision to skip a document, the title and the abstract elements triggered the selection decision the most. Second step of selection decision was assessing the item description which describes the individual document. Test persons assessed this by looking at query-biased summary/snippets. Looking at the elements that contributed to view selection decision, the elements abstract, title and snippets came first, second and third, followed by dates, item online and contributing institution. Looking at the elements contributed to not view decision, the elements abstract, title and dates came first, second and third, followed by item online, snippets and contributing institution.

Our second research question was: Why and how does this information influence the decision to click or not to click? Each element contributed in the selection decision in different ways. Title element indicated relevance, type of item, author information and specificity or broadness of the collection. Dates element showed information of time period, what the collection contained and the recency of the collection. Online element gave indication of effort that test persons need to spend, and time that was needed to read the collection. Abstract element was marked the highest by the test persons which means the summary was the most useful element in selection decision. Abstract element presented the overview of the whole collection, the type of document, specificity/broadness of a collection, time/effort the test persons need to spend, novelty of the collection/new information, originality of the records, authorship, and the types of collection. Snippets provided indication of relevance and specificity of the item. For the title and the snippets elements, we also found that test persons' ability to understand the information played an important role in test persons' selection decision.

Finally, we go back to the overall aim of this paper: What is the impacts of the result summaries on the users' decision to click or not to click? We concluded that contextual information about the document undoubtedly played an important role in supporting test persons in making a selection decision. For both view and not view decisions, test persons needed sufficient contextual information. Often this information was found in the title and abstract elements. This may be a result of the completeness and coherence of the information in these elements. A title element, although it is short, is a complete sentence and that affects the readability of the element. An abstract element, as compared to the other elements, is by far more complete and coherent and presents what the document is about. Only when the test persons could fully comprehend the information in the query-biased snippets, snippets were used to assess relevancy of the material and to see the detailed description of the document. Length of the information presented in the element also played a clear role. A whole paragraph (as in abstract) triggered a decision more frequently than a short sentence (as in title) or an incomplete sentence (as in query-biased snippet). Further research should answer how much information is needed for contextualizing the results, by studying the length of elements and the importance of the presence of the shown, but not marked elements, in this and other document genres.

Acknowledgments We thank the test persons for donating their time. This research is supported by the Netherlands Organization for Scientific Research (NWO) under grant # 639.072.601.

REFERENCES

[1] C. L. Barry. User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science*, 45(3):149–159, 1994.

[2] S. Betsi, M. Lalmas, A. Tombros, and T.Tsikrika. User expectations from XML element retrieval. In *In Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 611–612, 2006.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measures*, 20:37–46, 1960.

[4] W. M. Duff and C. A. Johnson. Accidently found on purpose: Information seeking behavior of historians in archives. *Library Quarterly*, 72:472–496, 2002.

[5] W. M. Duff and C. A. Johnson. Where is the list with all the names? Information-seeking behavior of genealogists. *The American archivist*, 66:79–95, 2003.

[6] W. M. Duff and P. Stoyanova. Transforming the crazy quilt: Archival displays from a user's point of view. *Archivaria*, 45:44–79, 1998.

[7] EAD. Encoded archival description version 2002, 2002. `http://www.loc.gov/ead/`.

[8] K. N. Fachry, J. Kamps, and J. Zhang. Access to archival material in context. In *Proceedings of the 2nd Symposium on Information Interaction in Context (IIiX 2008)*, pages 102–109. ACM Press, New York NY, USA, 2008.

[9] N. G. Huffman. Search features and other characteristics of XML retrieval systems for EAD finding aids: A content analysis. Master's thesis, School of Information and Library Science, University of North Carolina, April 2008.

[10] J. Kamps. Presenting structured text retrieval results. In *Encyclopedia of Database Systems (EDS)*. Springer-Verlag, Heidelberg, 2009.

[11] B. Larsen, A. Tombros, and S. Malik. Is xml retrieval meaningful to users? searcher preferences for full documents vs. elements. In *Proceedings of the 29th ACM SIGIR Conference*, pages 663–664, 2006.

[12] J. Lee. OAC first round usability test findings. *OAC redesign project*, September 2008. `http://www.cdlib.org/inside/projects/oac/oacredesign.html`.

[13] J. Lee. OAC second round usability test findings. *OAC redesign project*, June 2009. `http://www.cdlib.org/inside/projects/oac/oacredesign.html`.

[14] S. Malik, C.-P. Klas, N. Fuhr, B. Larsen, and A. Tombros. Designing a user interface for interactive retrieval of structured documents: Lessons learned from the INEX interactive track. In *10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, pages 291–302, 2006.

[15] OAC. Online Archives of California, 2009. `http://www.oac.cdlib.org/`.

[16] T. Park. The nature of relevance in information retrieval: an empirical study. *Library Quarterly*, 63:318–351, 1993.

[17] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

[18] L. Schamber, M. Eisenberg, and M. Nilan. A reexamination of relevance: toward a dynamic, situational definition. *Information Processing and Management*, 26(6):755–776, 1990.

[19] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–10, 1998.

[20] A. Turpin, F. Scholer, K. Jarvelin, M. Wu, and J. S. Culpepper. Including summaries in system evaluation. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 508–515. ACM Press, New York NY, USA, 2009.

[21] XML. Extensible markup language (XML) 1.0 (fourth edition), 2006. `http://www.w3.org/TR/xml/`.

APPENDIX

A. SIMULATED TASKS

Task 1 You have been asked to organize an activity as part of tobacco education program for high schools students. To get inspiration, you are doing research on previous activities that attempted to give tobacco education for school-age children. For example, you want to know what organizations were actively promoting tobacco education for school-age children, what purposes they had, and what anti-tobacco education projects and activities they implemented.

Task 2 You are writing an article about the damage of the 1906 San Francisco earthquake on buildings of San Francisco. As you know, the earthquake and resulting fire is remembered as one of the worst natural disasters in the history of the United States. To get data for your article, you want to know which buildings the earthquake damaged and to find photographs of the damaged buildings.

Task 3 You are interested in gold rush topic in the California, which happened in the 19th century. For your history assignment, you are planning to write an essay about gold rush at that time. To get some data for your essay, you are doing research about people who came to California as gold seeker, who they were and how their life was as gold seekers during the gold rush period.

Task 4 You are interested in the history of slavery in the 18th and 19th centuries. For your history assignment, you are planning to write an essay about anti-slavery movement of that period. To get data for your essay, you are doing research about people who were involved in the anti-slavery movement, who they were and in what way they promoted the anti-slavery movement.

Search Request (for all tasks) Using the digital inventory of the OAC, you would like to check out which archives contain interesting pieces for your research. Depending upon these findings, you should assess whether to visit the archives for your research is worth your time and effort.