Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 44 (2008) 437–462

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Contextual factors affecting the utility of surrogates within exploratory search

Ian Ruthven *, Mark Baillie, Leif Azzopardi, Ralf Bierig, Emma Nicol, Simon Sweeney, Murat Yaciki

University of Strathclyde, Department of Computer and Information Sciences, 26 Richmond Street, Glasgow G1 1XH, United Kingdom

Received 21 December 2006; received in revised form 9 August 2007; accepted 10 August 2007
Available online 19 September 2007

Abstract

In this paper we investigate how information surrogates might be useful in exploratory search and what information it is useful for a surrogate to contain. By comparing assessments based on artificially created information surrogates, we investigate the effect of the source of information, the quality of an information source and the date of information upon the assessment process. We also investigate how varying levels of topical knowledge, assessor confidence and prior expectation affect the assessment of information surrogates. We show that both types of contextual information affect how the information surrogates are judged and what actions are performed as a result of the surrogates.

Keywords: Surrogates; Exploratory search; Ciqa; Relevance assessment

1. Introduction

Information retrieval interfaces act as communication devices through which people interact with digital resources. Good information retrieval interfaces not only support the technical aspects of searching, such as entering queries, but also the cognitive aspects of searching such as creating good search requests, investigating an information space, and interpreting search results. As noted by Xie (2002) the strategies people develop to interact with a search system are directed by the interactive functionality offered by the search interface, that is the way people perform a search is influenced by the functionality and information presentation of the interface. Consequently the design of any interface, and the choice of what interactive functionality to offer, is an important factor in the success of a search interface.

One important set of interface components are surrogates: representations of information or search objects presented to the searcher at the interface level. Search interfaces can offer many types of surrogate. For text alone these can include summaries, text snippets, titles, abstracts, passages, sections, and top-ranking

[*] Corresponding author. Tel.: +44 1415483098.
E-mail address: ian.ruthven@cis.strath.ac.uk (I. Ruthven).

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.08.002

sentences and these surrogates may be created by the document author, such as titles, or automatically by the system, such as summaries or abstracts (Tombros & Sanderson, 1998; White, Jose, & Ruthven, 2003).

Surrogates of information objects are typically used to aid a searcher in making decisions about a search. Surrogates are especially common in presenting the results of a search – lists of titles being the most common form of retrieval results presentation – and are especially useful in exploratory search. In exploratory search the searcher typically engages in some process of investigation either to uncover more about the content of an information space, the structure of the space or simply to refine an information need. This exploration is usually facilitated by surrogates and in exploratory search, where the searchers are "*selectively seeking and passively obtaining cues about where the next [retrieval] steps lie*" (White, Kules, Drucker, & Schraefel, 2006), surrogates are vital to aid the process of selecting next search moves. A natural question, therefore, is how useful are surrogates and what information is it useful to include in a surrogate?

In this paper we describe a study to investigate two questions: what factors affect the utility of surrogates and what are effective forms of surrogate presentation. In the first question the factors in which we are interested relate to the searcher, including aspects such as the searcher's knowledge of the topics being searched, and the type of search being undertaken. In the second question we examine how the presentation of a surrogate can affect what decisions a searcher makes upon the surrogate.

Our study is based on our participation in the *ciqa* (complex interactive question answering) track of TREC 2006.[1] This task investigates interactive support when searching for answers to complex questions. Exploratory searches are typically complex searches, and one form of exploratory search are searches that involve a process of answering questions about an information space to narrow down the search space to target the most useful area of investigation. In this study we see answers to complex questions and the surrogates that present these answers as a way of iteratively narrowing down the search space within exploratory search. Our study is aimed at investigating what factors affect the use and effectiveness of surrogates within this type of exploratory search.

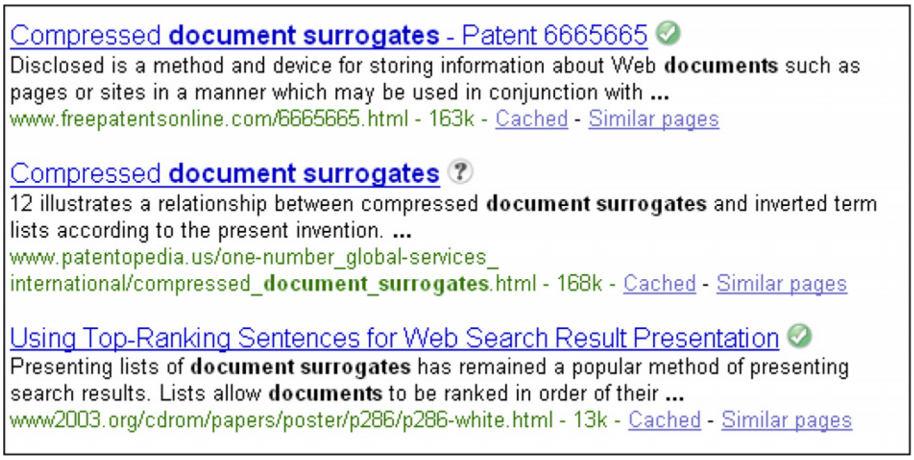
In Section 2 we examine previous research, in Section 3 we outline the *ciqa* track and in Section 4 we discuss the variables we investigated within our study. In Sections 5–7 we present the results of our findings, concluding with a discussion in Section 8.

2. Related work

From the earliest library catalogue systems searchers have relied on surrogates to help them assess information without being forced to examine the whole of an information object. Most search systems will use some form of surrogates and these surrogates can take many forms. They might, for example, be *part* of an information object, e.g. a key-frame from a video segment (Smeaton, Lee, O'Connor, Marlow, & Murphy, 2003) or an abstract of a document (Johnson, 1995), they might be meta-data about an object, such as library catalogue cards, or they can be some specially created representation of the object such as the text snippets commonly presented in web searching. Surrogates may be fixed representations, such as document titles, that are shown to all searchers or they may be dynamically created representations such as query-biased summaries, which are tailored to individual queries or searchers. In Fig. 1, for example, taken from a Google search on '*document surrogates*', searchers are presented with fixed title and URL surrogates and an automatically created text-snippet surrogate consisting of two lines of continuous text containing at least one query term.

Surrogates are particularly popular in presenting the results of a search and there are many possible types of surrogates. Recently the most popular class of surrogates have been summaries, especially the class of summaries known as indicative or suggestive summaries (Chuang & Yang, 2000; Tombros & Sanderson, 1998; White et al., 2003). Although summaries are mostly used for text retrieval they can be used for most media types and mixed media objects. Xu et al., for example, proposed summarisation techniques for music videos (Xu, Shao, Maddage, & Kankanhalli, 2005). Web retrieval interfaces have also utilised summarisation techniques, e.g. (White et al., 2003) and several researchers have examined what form useful summaries might take for web pages: standard text-based summaries or summaries that incorporate more visual aspects commonly

[1] http://www.umiacs.umd.edu/~jimmylin/ciqa/2006.



Compressed **document surrogates** - Patent 6665665 ✓
Disclosed is a method and device for storing information about Web **documents** such as pages or sites in a manner which may be used in conjunction with **...**
www.freepatentsonline.com/6665665.html - 163k - Cached - Similar pages

Compressed **document surrogates** ?
12 illustrates a relationship between compressed **document surrogates** and inverted term lists according to the present invention. **...**
www.patentopedia.us/one-number_global-services_ international/compressed_**document_surrogates**.html - 168k - Cached - Similar pages

Using Top-Ranking Sentences for Web Search Result Presentation ✓
Presenting lists of **document surrogates** has remained a popular method of presenting search results. Lists allow **documents** to be ranked in order of their **...**
www2003.org/cdrom/papers/poster/p286/p286-white.html - 13k - Cached - Similar pages

Fig. 1. Google surrogates.

found in web pages. Woodruff, Faulring, Rosenholtz, Morrison, and Pirolli (2001), for example, found that different types of surrogates work well for different types of search task. However, some form of aggregate surrogate, incorporating web page thumbnails and text, was best for most tasks and appeared to be a safe default. Dziadosz and Chandrasekar (2002) also found that different surrogates work better for different tasks (e.g. thumbnails can be poorer than textual summaries when searching for unknown items) and found that the presence of both thumbnails and text more led to more predictions of relevant material but also more incorrect predictions of relevance.

The notion of prediction – that useful surrogates should help people make reliable decisions about the material being summarised – is important to most work on surrogates. Two studies particularly examined how accurate predictions of relevance based on surrogates were compared to relevance decisions on the complete object. Vechtomova and Karamuftuoglu in this way examined the quality of query-biased sentences (Vechtomova & Karamuftuoglu, 2006). The assessors in their study were generally good at predicting relevance – correctly identifying 73% of sentences as coming from documents later judged as relevant. This level of accuracy naturally, however, is dependent on the quality of the sentence selection – surrogate production – algorithm. Ruthven, Baillie, and Elsweiler (2007) also examined sentences, in this case introductory sentences from newspapers rather than query-biased sentences. This second study, showed that sentences could lead to good prediction of relevance but the predictive value of sentences was dependent on the personal characteristics, such as level of topical knowledge, of the person making the assessment. People with high levels of topical knowledge, for example, can over-estimate the presence of relevance information leading to poor predictions. In this study the ability to predict relevant material (material that the assessor would mark as being relevant after reading the full document) was not related to their ability to predict non-relevant material. That is, assessors could be good at identifying sentences that came from non-relevant documents but weak at identifying sentences from relevant documents. The prediction success, of either predicting relevance or non-relevance, was also based on a searcher's willingness to predict. That is when given the choice some assessors would rather not make a prediction and would rather see the complete information object.

The presentation of surrogates is also important. Sweeney and Crestani (2006), point to an interesting distinction between effectiveness and preference: in their study of optimal summary length for handheld device presentation they found that people prefer longer summaries on larger devices but longer summaries did not make them more accurate at using summaries to predict relevance. In (Bell & Ruthven, 2004) presenting *many* sentence surrogates was shown to be useful in reducing the complexity of searches by allowing searches to easily see an overview of the retrieved material without having to access individual documents serially. In White, Ruthven, and Jose (2005) an interface that layered exposure to different types of surrogate was shown to be useful in supporting different stages within a search.

Through the use of novel elicitation methods such as eye-tracking we can learn more about how people use surrogates in searching. The study by Lorigo et al. (2006) for example indicated that in over half of their investigated web searches, users reformulated queries based on scanning the surrogates alone without examining any pages and that navigational (website-finding) queries are often answered by surrogates alone. That is if searchers are willing to accept occasional false hits when making decisions based solely on surrogates. False hits in this case are pages appearing to be relevant because the surrogate misrepresents a page's content; this can arise from surrogates being created from cached pages (automatically generated surrogates), from deliberate misrepresentation of the page's content (Lynch, 2001) or from the surrogate-creation process resulting in a surrogate that does not represent the real content of an object.

Given the prevalent use of surrogates within search interfaces, particularly for web search interfaces, we examine in this study what contextual factors might influence the effectiveness of surrogates. The contextual factors we explore are those relating to the person using the surrogates (personal contextual factors), and contextual factors relating to the presentation of summaries (the context in which information is displayed). Our study is carried out as part of the 2006 *ciqa* track.

3. Complex interactive question answering (*ciqa*)

The *ciqa* task arose out of the HARD (High Accuracy Retrieval from Documents) track of TREC which ran from 2003–2005 (Allan, 2005). The HARD track was designed to utilise interaction from an assessor in order to improve retrieval effectiveness. Rather than assuming that there was one result list that would be good for all searchers, HARD investigated whether employing information from individual assessors could be used to personalise, and hopefully improve, retrieval performance. To enable this personalisation, later versions of the HARD track facilitated the capture of contextual, or personal, information through the use of so-called clarification forms. These clarification forms were HTML questionnaires given to the TREC assessors who would assess documents retrieved by participating retrieval systems.

The HARD track finished in 2005 but the useful idea of exploiting a dialogue with the document assessors was carried on to the *ciqa* (complex interactive question answering track) task. *ciqa* in 2006 was an optional sub-task of the main Question Answering track in TREC (Kelly & Lin, 2007). In *ciqa* the questions are complex questions where the complexity arises from the relationships between two or more entities. An example is shown in Fig. 2 where the relationship of interest is a financial relationship between the two entities *drug companies* and *universities*. These questions are seen as more complex than the simpler factoid type questions (e.g. "who killed President Kennedy?") previously investigated in question answering, partly because the structure is more complicated – by relating concepts or entities – and also because the underlying information need may be more complex comprising of several sub-questions. This aspect of complexity will be discussed in more detail in Section 4.2.

Associated with each question is a description (or narrative) that explains more about the information need associated with the question. As can be seen in Fig. 2 the description provides additional information regarding the underlying information need not evident from the simple question itself, i.e. the aspects of universities potentially biasing their findings and financial support of universities by drug companies.

The protocol for *ciqa* was as follows:

1. Each participating group submits an initial run consisting of the top answers retrieved by their system ranked in order of perceived answer quality. At the same time each group submits up to two interaction forms (see Section 4) to gain further information from the question creator that might be used to improve the answers provided.

Question: What financial relationships exist between drug companies and universities?

Description: The analyst is concerned about universities which do research on medical subjects slanting their findings, especially concerning drugs, towards drug companies which have provided money to the universities.

Fig. 2. *ciqa* topic number 32.

2. The TREC assessor who created the question completes the interaction forms and the form answers are returned to the participating groups. The assessor was allowed only 3 min to complete each form. If the form was not completed or submitted at the end of 3 min then the form was timed out and automatically submitted.
3. The participating groups use the answers to their own interaction forms to produce a new set of answers.
4. The TREC assessor evaluates both initial and final set of answers.

Each answer is only assessed by one assessor, the assessor who created the question, and both sets of answers are evaluated at the same time. *ciqa* 2006 involved eight assessors who each created between two and five questions of different types.

In our participation to *ciqa* 2006 we did not attempt any novel Question Answering research. Rather we wished to explore the assessment process involved in assessing answers and how personal characteristics, as well as characteristics of the questions and answers themselves, affected the assessment process. As we aim to show in the remainder of the paper, the results we obtained are useful in understanding the role of surrogates in exploratory search.

4. Variables under study

In our study we investigated three groups of variables and relationships between these variables. The three groups of variables focus on variables relating to the assessors[2] themselves (Section 4.1), the questions set (Section 4.2) and the answers presented (Section 4.3). In this section we discuss the variables we selected, how we measured them and the relationships in which we are interested.

The majority of variables were investigated through the use of interaction forms – HTML forms presented to the assessors and completed by the assessors. Each participating group in *ciqa* was allowed to submit two interaction forms per topic and, as noted before, each form must be capable of being completed within 3 min. Beyond this there are few limitations placed on the design of the form.[3]

We designed two interaction forms: Interaction Form 1 gathered information on the assessor and on the question, Interaction Form 2 gathered information on answers to the questions. Both forms are shown in the Appendix, Figs. A.3 and A.4.

4.1. Assessor variables

The first group of variables relate to the assessors themselves. In *ciqa* the same person who created the question and description fields assesses the answers to the question and completes the interaction forms. Knowing more about this person could provide useful information regarding their preferences for answer formats or how their personal characteristics affect the answer assessment process.

To gain information on the assessor we asked for four responses:

- *Topical knowledge*. Topical knowledge has been shown to be one of the major factors in assessing relevance (Hsieh-Yee, 1993; Michel, 1994; Vakkari & Hakala, 2000; Wen, Ruthven, & Borlund, 2006). We asked the assessor to rate their topical knowledge ("*How much do you think you know about the topics in this question*:") of the entities involved in the question on a three-valued scale "*not much/same as most people/know quite a lot*".
- *Confidence in assessing answers*. Although the questions are created by the assessors themselves this does not guarantee that they will find the task of assessing answers easy.[4] In a previous study (Ruthven et al., 2007) we found that asking assessors to rate their confidence in assessing retrieved material was a useful question in

[2] We use the term assessors rather than searchers or users as the assessors do not search or use a search interface during the course of this study. Rather their primary role is to complete the forms and assess answers.

[3] Aside from a few technical restrictions, e.g. the form cannot invoke java applications or cgi scripts.

[4] In addition to returning answers, participating groups must also provide the document identifier of the document supporting the answer in order that assessors can verify the answers if they are unsure of the answer.

identifying searcher preferences regarding presented material. For example, we found a useful relationship between an assessor's willingness to predict the relevance of an object, based on a surrogate, and the final assessments made by the assessor on the object. Consequently, we asked the assessors to rate their confidence in assessing correct answers to their own questions ("*For this question, how confident are you that you could recognise* ***correct*** *answers to this question?*") on a three-valued category ("*very confident/depends on the answers returned/not very confident*"). We asked about recognising correct answers because we were interested in how useful surrogates might be in providing answers without reference to the full document, as will be explained in more detail in Section 5.3.3. We used a categorical reply, rather than a simple scale, based on our previous experience of using this attribute, explained in (Ruthven et al., 2007), in which splitting assessors based on willingness to assert a confidence level gave useful distinctions in the data.

- *Prior expectations*. Next we asked the assessor if they already had an expectation of an answer to the question set. We asked "*Do you have an answer in mind for this question (could you provide an answer without searching)?*" and solicited responses on a four-category scale ("*yes/no/I could provide a partial answer/no answer but have an idea of what an answer might look like*"). The questions set by the assessors are original questions so we expect they have some exposure to the topics in their questions even if they may know little about the topics. The prior expectation question was designed to elicit some information about their knowledge of possible *answers*, as opposed to the topical knowledge question which elicited information about their knowledge of the entities involved in the *question*. The latter two possible responses ("*I could provide a partial answer/no answer but have an idea of what an answer might look like")* were intended to differentiate between situations where the assessor is confident enough to provide a partial answer (but may require more information to provide a full answer) and situations where the assessor does not know the answer but has an expectation of the likely answers or at least the direction of the answer. For example, for topic 33 "*The analyst is especially interested in opinions of scientists as to whether there is a family link between dinosaurs and birds, and what evidence they cite concerning their opinions*", the assessor may suspect that there is a family link between dinosaurs and birds and is looking for confirmatory evidence. We asked about prior expectations because we were interested in how existing knowledge might affect the assessment of answers provided. Different relevance criteria may be applied at different stages in an information problem (e.g. Vakkari & Hakala, 2000) and that exposure to information can alter the searcher's perception of their information problem (Park, 1993). An assessor with clear expectations of the answers to be found could signal a situation where the assessor is simply expecting confirmation of existing knowledge. In such a case we might expect the assessor to be more definite in judging the quality of answers provided and less susceptible to presentation of answers.
- *Variety of opinions*. The final question we asked the assessor was what they would prefer in terms of a *set* of answers. All answers provided by a retrieval system can be assessed individually but a searcher is typically presented with a set of answers or surrogates when searching. Although searchers are forced to read each surrogate individually, the set of answers as a whole is important because all answers are available for inspection. We asked the assessor about what type of answer set they required – "*For this question, would good set of answers contain?*" – and asked them to respond using one of two options "*a variety of similar opinions (or evidence)/as many different opinions as possible*". Although few topics gave clues to which of these two options would be applicable to the topic, we felt that assessors may well have reasons for asking for the information and may require answers of particular direction, e.g. to confirm an existing belief, to search for any evidence that refutes a belief, or to obtain the most likely (most common) answer to a question. This aspect also relates to the novelty of information, novelty being one of the main relevance criteria used in searching (Barry & Schamber, 1998). A set of very similar answers may provide less *new* information after the first answer, whereas a set of very different answers will provide more new information but may raise doubts as to the correct answer.

4.2. Question variables

The second group of variables relate to the questions and question descriptions, Fig. 2. In studying the questions we are interested in five variables:

Transport: What evidence is there for transport of [goods] from [entity] to [entity]?

Relationship: What [relationship][5] exist between [entity] and [entity]?

Influence: What influence/effect do(es) [entity] have on/in [entity]?

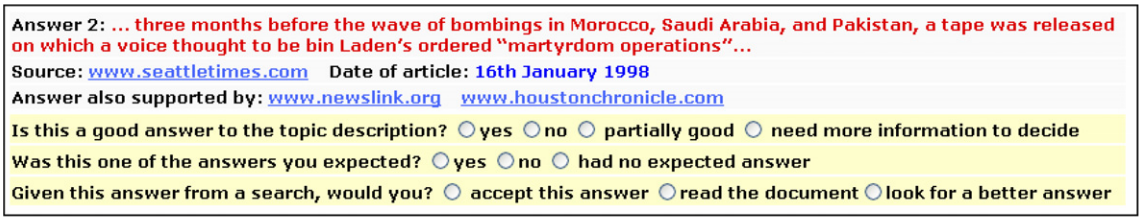
Position: What is the position of [entity] with respect to [issue]?

Evidential: Is there evidence to support the involvement of [entity] in [event/entity]?

Fig. 3. *ciqa* templates.

- *Time of answer*. The questions used in *ciqa* often relate to news events and the time of these events may be important in detecting good answers. The narrative for question 45 for example "*What is John McCain's position toward Jerry Falwell's Moral Majority and Pat Robertson's Christian Coalition? Does he support the organisations or does he oppose them?*" suggest that what is primarily required is recent information not historical information, i.e. does John McCain currently support or oppose these organisations? The narrative for question 49, on the other hand, "*The analyst would like to know how Richard Seed felt about human cloning. Specifically, the analyst would like to know what his feelings were regarding human cloning and what actions he took as a result?*" suggests, through the use of past tense, that the information required may be older information. For other questions there appears to be no bias in time information, e.g. question 42, "*The analyst is interested in the effect of aspirin on coronary heart disease and stroke. Specifically, what does aspirin do and how does it do it?*" For each question the assessor was asked about preferences on the date of good answers "*For this question, would good answers come from?*" The response was modelled as a categorical variable ("*recent articles/older articles/any articles*").
- *Number of entities*. As noted in Section 3 the questions in *ciqa* relate entities. The number of entities, as marked by the question creators, is our second variable. Number of entities is a categorical variable with values of 2 (e.g. "*What effect does [aspirin] have on [coronary heart disease]?*") or 3 (**e.g**. "*What [financial relationships] exist between [the United States] and [supporters of the Irish Republican movement]*").
- *Type of relationship*. *ciqa* questions are modelled on templates as shown in Fig. 3 with six questions from each template. *ciqa* investigated five types of relation in 2006 – transport, relationship (which encapsulated financial relationships, organisational ties, familial ties, and common interests), effect relationships, attitude (or position) and evidential relationships. Each question could therefore be assigned to one of the categories: transport/relationship/influence/positional/evidential.
- *Predicted difficulty*. Our fourth variable was the predicted difficulty of the search. Although the assessors were being asked to judge only retrieved answers to their own questions, rather than perform a search themselves, we felt it would be useful to ask for their opinion on how difficult the information problem (i.e. the description field) would be to answer using a general purpose search engine such as Google. We ask this in case the assessors have any pre-conceived view on the information problem that might affect their judgement of the quality of the answers presented in the second interaction form. This relates in someway to the process known as "anchoring"[5] in which people estimate unknown values (the quality of answers in our case) by starting from an initial value which "*may be suggested by a formulation of the problem*" (Tversky & Kahneman, 1974). That is, if the assessor believes the question is easy to answer they might rate answers more strictly than if they believe the question to be difficult to answer. The predicted difficulty was measured using a 5 point scale ("*very difficult/fairly difficult/cannot predict how difficult/fairly easy/very easy*") to the question "*If you were searching for answers to this question using a web search engine such as Google, how easy do you think it would be to find good answers?*".

[5] Proposed and named by Tversky and Kahneman (1974) but related to information seeking and retrieval by Blair (1980).



Answer 2: ... three months before the wave of bombings in Morocco, Saudi Arabia, and Pakistan, a tape was released on which a voice thought to be bin Laden's ordered "martyrdom operations"...

Source: www.seattletimes.com Date of article: 16th January 1998

Answer also supported by: www.newslink.org www.houstonchronicle.com

Is this a good answer to the topic description? ○ yes ○ no ○ partially good ○ need more information to decide

Was this one of the answers you expected? ○ yes ○ no ○ had no expected answer

Given this answer from a search, would you? ○ accept this answer ○ read the document ○ look for a better answer

Fig. 4. Answer template in Interaction Form 2.

- *Complexity*. The questions are designed to be complex in the sense that they relate several entities or concepts (e.g. *drug companies* and *universities* in Fig. 2). The description field describes why the information is required and other details of the information need promoting the question (e.g. "*the analyst is concerned about universities which do research on medical subjects slanting their findings, especially concerning drugs, towards drug companies which have provided money to the universities*") from question 32 in Fig. 2. The description field, therefore provides, additional information that will be used to assess the answers returned by participating groups. For some questions this additional information simplifies the original question, e.g. "*Specifically, the analyst seeks evidence that smugglers use the island of San Andres for such a purpose*"; for other questions the description field extends the original question to include additional questions. For example for question 33 the question asks "*What familial ties exist between dinosaurs and birds?*" whereas the narrative makes it clear that an answer should contain both opinions on whether dinosaurs and birds are related *and* the evidence for such positions ("*The analyst is especially interested in opinions of scientists as to whether there is a family link between dinosaurs and birds, and what evidence they cite concerning their opinions*"). We deemed a question with several sub-questions that require answer as being more complex to answer. Conflating complexity with number of sub-questions contained within the question description we performed an internal classification of the number of sub-questions in each topic description. Three internal assessors were asked, independently, to count the number of sub-questions in each topic description. For all except six topics the assessors agreed on the number of sub-questions. A short discussion resolved the disagreement on these six topics to give a number of sub-questions for each topic[6] which we treat as a measure of complexity of the question. This variable is distinct from the number of entities, which is expressed in the question itself.

4.3. Answer variables

Interaction Form 2 gathered information on answers by presenting a series of eight answers to the assessor. Each answer had a common layout consisting of three off-white[7] fields comprising the answer and contextual information and three pale yellow fields containing our questions to the assessor regarding the answer. An example is shown in Fig. 4. The first answer line contained the answer to the question presented in red font, the second line contained a source for the answer presented as a URL and the date of the article presented in a dark blue, and the third line contained sources that supported (agreed) with the answer.

Our surrogates are, then, a combination of content and meta-data. The content are selections of the page text, similar in style to text snippets from search engines, and the meta-data provides information on the web page containing the information.

All presented answers were selected from manual searching of the web. Although the answers to be returned to *ciqa* for assessment were to be answers from the AQUAINT news collection we felt that we could

[6] Topic 31 was agreed to have three sub-questions, topics 26, 27, 29, 30, 33, 34, 35, 38, 41, 42, 44, 45, 46, 48, 49, 50, and 51 had two sub-questions and the remaining topics had only one sub-question. We found no correlation between the number of entities in the question and number of sub-questions in the topic description.

[7] The colour of these fields may not appear very strong in this paper version. After initial pilot testing we reached a balance between contrast of information and visual separation of answers to *ciqa* questions (offwhite) and questions to the assessor (yellow).

obtain better answers from the general web and we tried, through manual searching, to simulate the results of a good question answering system. That is, we tried to find answers that related to as many points of the question as possible within a short text span.

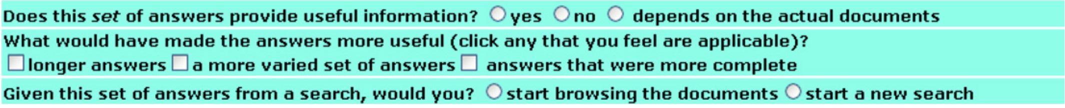
Three authors selected answers for the questions; each question being assigned to only one author. One author checked all answers for clarity and length to avoid any major differences in these attributes. Short answers, rather than whole sentences or paragraphs, were selected through manual examination of the top results from submitting manual queries to the major search engines. Short answers were preferred to simulate the main question answering task. We allowed multiple answers from the same document although in practice this was only necessary in a small number of cases where the answers were distinct (i.e. contained different information). We lacked knowledge of many of the domains covered by a number of questions and so deliberately did not attempt to select 'correct' answers. Rather we tried to form a representative set of answers; a set of answers that seemed to reflect the distribution we found when searching. So, if most documents gave similar answers then most answers presented in the interaction forms were similar although none were identical.

All answers contain a textual answer with ellipsis to denote missing text if the answer is a fragment of a sentence e.g. "*...an appearance on Oprah or Today can shoot book sales through the roof...*".

When presenting the answers we set out to investigate three variables

- *Time of answer*. All answers were presented with information on the date of the source containing the answer. This was to test whether the date of the information was useful in assessing the answer. Answers were randomly assigned dates from one of two lists of dates: recent dates, in this case only from the first six months of 2006,[8] or older dates, in this case prior to 2004.
- *Quality of source.* Each answer was associated with a source which was a website URL. All URLs shown were presented as hypertext links but were not linked to any other page, i.e. clicking on the text did not transfer the assessor to a new page. Although all answers presented were genuine, the sources were manually assigned and did not correspond to the actual sources of the answer. Rather we sought to distinguish between high and low quality sources of information. High quality sources of information were ones that we felt that most assessors would recognise as established sources of reputable information, even if they did not agree with any particular political stance or editorial policy of these sources. We developed a list of sources which were primarily chosen from a list of top 10 US newspapers (assessment was carried out in US so we assumed assessors would recognise these newspapers), several well known television stations such as CNN, CBS, and BBC and quality online resources such as Wikipedia. Low quality sources of information were ones that we felt assessors would be unfamiliar with, primarily sources that had unusual names. These URLs were nearly all fictitious. In the case that these URLs now correspond to real URLs, we do not assert that these are indeed low quality sources of information, rather than at the time of assessment they would not be recognised as known, reliable sources of information. A list of the two sets of URLs is presented in the Appendix, Figs. A.1 and A.2. The answers provided bear no relation to the actual content of these sources; the sources were only used to test whether the source of the information was important in assessing the quality of an answer presented on the interaction form. Lin et al. (2003) also investigated the role of source in answer presentation but distinguished between biased, neutral and unknown sources; biased sources being ones that are known to have specific views on a topic (such as a pressure group or company), neutral sites being ones such as Wikipedia that are intended to present strictly factual, rather than opinionated information, and unknown sources being ones where the authority of the source has not been established. In Lin et al.'s study the interest was in the perceived trust of a source: opinionated sources may provide answers that require more evidence before answers are accepted. Although their subjects felt that the stance of individual sources would be important in deciding whether or not to accept an answer from the source, the experimental results indicated that the nature of the source was not important. In our study we investigate familiarity rather than trust: our high-quality sites are sites which may be opinionated or not.

[8] The interaction forms were completed in August 2006 so these dates are effectively dates from 6 months preceding the answer assessment.



Does this *set* of answers provide useful information? ○ yes ○ no ○ depends on the actual documents

What would have made the answers more useful (click any that you feel are applicable)?

□ longer answers □ a more varied set of answers □ answers that were more complete

Given this set of answers from a search, would you? ○ start browsing the documents ○ start a new search

Fig. 5. Questions on quality of answer set.

- *Supporting evidence.* According to Barry and Schamber (1998) one of the important criteria in assessing relevance is the presence of supporting or confirmatory evidence. That is, evidence that information (in our case an answer) is supported by multiple sources can lead to the information being more likely to be assessed relevant. Accordingly we presented some answers as having multiple sources of information agreeing on the answer. For example in Fig. 4 the answer is provided by www.seattletimes.com and supported by www.newslink.org and www.houstonchronicle.com, i.e. claiming that these two sources also provide the same answer. If an answer had supporting sites these corresponding to the perceived quality of the original source, i.e. high-quality sources were supported by high-quality sources and weak-quality sources were supported by weak-quality sources.[9] It would have been useful to mix these two conditions (quality of original source vs. quality of supporting sources) but the number of combinations required would have been too many to assess within the three minute condition set by the *ciqa* organisers. These supporting sources were also manually assigned and bear no relationship to the actual content of the sources. This allowed us to test whether supported answers were preferred to unsupported answers.

The cross combination of three variables (recent vs. older information, high-quality source vs. low-quality source, supporting vs. no supporting sources) gives eight combinations of answer presentation. Thus, for each topic eight surrogates were provided to assessors in Interaction Form 2: one answer from each combination.

For each answer we asked the assessor to assess:

- *Quality of answer*. The first question, asked on all answers, was on the general quality of the answer "*Is this a good answer to the topic description*" and assessors were asked to respond using the categories "*yes/no/partially good/need more information to decide*". "*Partially good*" was intended to reflect answers that supply some useful information but not necessarily all the required information, and "*need more information to decide*" to reflect the situation where the assessor would need more context from the document to decide on the actual value of the answer.
- *Expectation of answer*. Next we asked about the fit of the answer to the assessors prior expectation of the answers – *"was this one of the answers you expected"* which was to be answered using the categories*"yes/no/had no expected answer"*.
- *Next action*. Finally we asked what the assessor would do given this answer from a search "*Given this answer from a search would you?*". In this case the answers were limited to "*read the document/look for a better answer/accept this answer*". The answers "*accept this answer*" and "*look for a better answer*" both reflect a decision to use the answer as presented without checking the context of the answer in the whole document. Simply accepting the answer as presented could be seen as a high degree of trust in the accuracy of the answer whereas moving to get a better answer would indicate the answer is seen as poor or not relevant.

A final set of questions, displayed after all answers, asked the assessor about the set of answers as a whole, Fig. 5. As we noted in Section 4.1, although answers may be read serially, a searcher will be aware of all the answers and can make assessments on the complete set of results presented. We first asked whether the set of answers provided useful information. The answers themselves might fulfil the assessors' need without reading the full text of the documents ("*yes*"), or might be inappropriate answers ("*no*") or the assessor may require to

[9] We ensured that no two answers on the same form (i.e. answers to the same question) had the same source either as primary source or supporting source.

Table 1
Summary of variables and measurement

Assessor	How measured	Nature
Knowledge of topics in question	Self-assessment in Interaction Form 1	Three point ordinal scale
Confidence in assessment of answers	Self-assessment in Interaction Form 1	Three point ordinal scale
Prior expectation of answers	Self-assessment in Interaction Form 1	Four-valued category
Predicted difficulty	Self-assessment in Interaction Form 1	Five point ordinal scale
Variety of opinions required	Self-assessment in Interaction Form 1	Two-valued category
Question		
Complexity of question	Sub-question analysis	Count of sub-questions
Number of entities	Mark-up of questions from *ciqa* organisers	Count of entities
Type of relationship	Categorisation from *ciqa* organisers	Five-valued category
Time of relevant information	Assessor opinion from Interaction Form 1	Three-valued category
Answer		
Quality of source	Presence of good/weak named source in Interaction Form 2	Binary variable (good/weak source)
Presence of supporting sources	Presence/absence supporting sources in Interaction Form 2	Binary variable (presence/absence supporting sources)
Date of answer	Presence of recent/older dates in Interaction Form 2	Binary variable (recent/older information)
Quality of answer	Assessor opinion from Interaction Form 2	Four-valued category
Expectation of answer	Assessor opinion from Interaction Form 2	Three-valued category
Next action on answer	Assessor opinion from Interaction Form 2	Three-valued category
Answer set		
Utility of answer set	Assessor opinion from Interaction Form 2	Three-valued category
Increased utility of answer set	Assessor opinion from Interaction Form 2	Three-valued category (allows for multiple responses)
Next action on answer set	Assessor opinion from Interaction Form 2	Two-valued category

read the documents to judge how useful the answers were ("*depends on the actual documents*"). Next we asked what, if anything, would have made the answers more useful. Here we had three choices and the assessor could select any combination. Answers could have been more useful if they were longer, more varied (i.e. contained more different types of information or different answers) or more complete (i.e. answered more sub-questions of the description). Finally we asked what the assessor would do given this form (set of answers) from a search, either browse the documents themselves or start a new search. As we mentioned previously we were interested in the use of answers and surrogates in web search. A poor set of answers might lead to a new query whereas a good set of answers could encourage the searcher to explore the documents retrieved. This final question was intended to reflect an overall assessment of the answers.

4.4. Summary of interaction forms

Our two interaction forms, shown in the Appendix, Figs. A.3 and A.4, consisted of a range of questions to gather information on 18 variables. Interaction Form 1 asked about all the assessor variables and the time of relevant information variable (question variables), Interaction Form 2 gathered information on all the answer and answer set variables. The remaining variables were measured using data gathered internally (complexity), or from data obtained from the *ciqa* organisers. Table 1 summarises the variables investigated, how these were gathered and the type of each variable.

5. Individual variables

In this section we first report on our overall findings on individual variables within each group to provide background information which we will use to contextualise the results in later sections when we consider the relationships between these variables.

5.1. Assessor variables

Interaction Form 1 gathered information from the assessor. Below we show the distribution of responses to the questions asked.

- *Knowledge of major topics in question* (know a lot ($n = 5$), same as most ($n = 17$), not much ($n = 8$)).
- *Confidence in assessment of answers* (very confident ($n = 22$), depends on the answers returned ($n = 8$), not very confident ($n = 0$)).
- *Prior expectation of answer* (yes ($n = 4$), no answer but have idea ($n = 7$), could provide a partial answer ($n = 15$), no ($n = 4$)).
- *Predicted difficulty* (fairly easy ($n = 3$), cannot predict how difficult ($n = 9$), fairly difficult ($n = 16$), very difficult ($n = 3$)).
- *Variety of opinions required* (as many different answers as possible ($n = 20$), variety of similar answers ($n = 7$)).
- *Time of required information* (recent ($n = 3$), older ($n = 4$), any ($n = 23$)).

The assessors, on the whole, asserted at least an average level of topical knowledge with a high confidence in their ability to assess the accuracy of answers to the questions even though the questions were perceived to be difficult to answer. Even though the assessors felt that they only had at least average knowledge of the topics for most questions (22 out of 30) they felt that they had sufficient information to guess at least a partial answer to the question suggesting a certain degree of existing topical knowledge and an awareness of possible information available. The most desired common response from the system was a variety of opinions, i.e. different answers, from any date range although for a significant minority of topics the assessors were interested in a range of similar opinions.

We compared the assessor variables against each other, using the non-parametric Spearman Rank test, but found no significant correlations between the assessor variables.

5.2. Question variables

We proposed four question variables: complexity, number of entities, type of relationship and time of relevant information. Complexity was measured using a sub-question analysis approach which showed most topics to have only 1 or 2 sub-questions per question and the number of sub-questions were not related to the number of entities within the question. If we group the topics by relationship type, as in Table 2 where we present the average number of sub-questions and entities per question, we see that some question types are simpler than others. The evidence type for example has on average fewer sub-questions and deals with a smaller number of entities. The transport type of question on the other hand has more sub-questions and more entities. As we will discuss in Section 7.4 the type of question asked appears to have more of an effect on the quality of answer than the complexity or number of entities involved.

The date of answers in most cases was perceived not to be important for these questions with assessors predicting that good answers would come from recent articles in three questions, older articles in three questions and any articles irrespective of date for the remainder of questions. However, as we shall show in Section 6.2 presentation of time information can affect search behaviour.

Table 2
Sub-questions and entity count by relationship type

Relationship type	Sub-questions	Entities
Transport	**2**	**3**
Relationship	1.5	3
Effect	1.5	2
Position	1.83	2
Evidence	1.33	2

Highest value shown in bold.

Table 3
Distribution of answers to answer quality variable

Response to question "*Is this a good answer to the question?*"	Percentage
Yes	**57.89% (*n* = 121)**
No	16.75% ($n = 35$)
Need more information to decide	14.35% ($n = 30$)
Partially good	11.00% ($n = 23$)

Highest value shown in bold.

Table 4
Distribution of answers to answer expectation variable

Response to question "*Was this one of the answers you expected?*"	Percentage
Yes	**50.48% (*n* = 105)**
No	22.12% ($n = 46$)
Had no expected answer	27.40% ($n = 57$)

Highest value shown in bold.

Table 5
Distribution of answers to answer expectation variable

	Yes ($n = 102$)	No ($n = 54$)	Had no expected answer ($n = 65$)
Accept	**75.49% (*n* = 77)**	12.96% ($n = 7$)	41.79% ($n = 26$)
Move	1.96% ($n = 2$)	**51.85% (*n* = 28)**	8.96% ($n = 6$)
Read	22.55% ($n = 23$)	35.19% ($n = 19$)	**49.25% (*n* = 33)**

Highest value shown in bold.

5.3. Answer variables

For individual answers we have six variables: three which reflected direct questions to the assessor (quality of answer, expectation of answer, next action on answer) and three which were related to the presentation of answers (quality of answer source, presence of supporting sources, date of answer).

5.3.1. Quality of answer

The majority of answers (58%)[10] were deemed to be good answers to the questions, Table 3. For a small percentage of answers (on average 1 presented per form) the assessor felt that they could not decide on the quality of the answer without reference to the entire document. That is, the answer on its own did not allow the assessor to make a decision with seeing the answer in the context of the entire article.

5.3.2. Expectation of answer

After each answer we asked the assessor whether the answer matched their prior expectation of the answer. For half the answers the assessor felt that the answer did match their expectations and for just over a fifth of answers, the answers were unexpected, Table 4. In Table 5 we present the predicted next actions according to the assessors' opinions on the expectation of the answer. For answers that the assessor expected to receive, the most common predicted next action was to accept the answer, followed by reading the document. Reading the document may resolve the case where the assessor had a partial answer in mind rather than a complete answer. For unexpected answers the most common action (51%) was to move in search for a different answer, followed

[10] A small number of questions in our forms were not answered. We present percentages for fairer comparison between categories but report on the actual number of responses within the tables ($n = x$).

Table 6
Distribution of answers to answer quality variable

Response to question "*Given this answer from a search, would you?*"	Percentage
Accept this answer	**41.20% (*n* = 85)**
Read this document	40.30% ($n = 83$)
Look for a better answer	18.50% ($n = 19$)

Highest value shown in bold.

Table 7
Distribution of predicted next actions over answer quality

Predicted next action	Answer quality			
	Good ($n = 121$)	Poor ($n = 34$)	Partially good ($n = 21$)	Cannot decide ($n = 25$)
Accept	**69.42% (*n* = 84)**	0.00% ($n = 0$)	0.00% ($n = 0$)	0.00% ($n = 0$)
Move	0.00% ($n = 0$)	94.12% ($n = 32$)	14.29% ($n = 3$)	20.00% ($n = 5$)
Read	30.58% ($n = 37$)	5.88% ($n = 2$)	**85.71% (*n* = 18)**	**80.00% (*n* = 20)**

Highest value within each answer assessment category shown in bold.

by reading the document. For the situations where the assessor had no expected answer the most common action was to read the document, although a sizeable number would simply accept the answer as given.

5.3.3. Next action on answer

For most answers presented the assessors were split between reading the document and simply accepting the answer as presented, Table 6. For the minority of answers (18.5%), a figure comparable to the percentage of poor answers in Table 3, the assessors would reject the answer and carry on searching.

In Table 7 we examine the predicted next actions against the quality of answers (good/poor/cannot decide/partially good), e.g. for 30.58% of good answers the assessor felt their next action would be to read the document. From Table 7, for good answers the most likely action is simply to accept the answer, for poor answers in almost all cases the predicted next action is to look for a better answer and for cases where there is a level of uncertainty (partially good answers or cannot decide) the predicted next action is to read the document.

5.3.4. Summary

In this section we presented some general findings regarding the main classes of variables under study. The assessors can be seen as relatively knowledgeable, confident assessors who have some idea of the answers they expect from what they see as difficult search topics. In the following sections we examine how their judgements differ when we examine the surrogates (Section 6), differences in the assessor characteristics (Section 7) and the questions being tackled (Section 8).

6. Results on contextual information in surrogates

The variables described in Sections 5.3.1,5.3.2,5.3.3 were investigated through direct questions to the assessor. Some of the results are not surprising – the fact that assessors felt they would read documents containing answers on which they could not decide the quality or that they would reject poor answers. What we examine in this section is what factors might affect these decisions to see whether the remaining variables (quality of source, date of answer, presence of supporting sources) led to different types of response from the assessors, e.g. did the date of the answer affect the decision to mark an answer as good, poor, etc or did it lead to different predictions regarding the actions that the assessor would perform on the answer.

In each of the following three sections we examine each variable (supporting evidence, date, and quality of source) in turn. Where appropriate we test for statistical significance using the Wilcoxon test.

6.1. Comparing answers with or without supporting evidence

First we compare whether the presence of supporting evidence leads to answers being assessed differently. In Table 8 we present the percentage of answers within each answer assessment category (*good answer/poor answer/cannot decide/partially good*) that had supporting evidence or no supporting evidence. For example, on average, 60.5% of supported answers were rated as good (row 2) whereas only 54% of non-supported answers were assessed as good (row 6).

Overall Table 8 indicates that supported answers were more likely to be rated as good than poor whereas unsupported answers were more likely to be rated as only partially good. If we collapse these two categories into a single 'good' category then both supported and unsupported answers have approximately 68% of presented answers judged as good. The implication is that supporting sources are slightly more likely to lead an answer to be rated good whereas the lack of supporting sources appears to lead to a degree of doubt as to the quality of the answer. There is also a higher percentage of unsupported answers being marked as poor (13% supported vs. 20% unsupported). The difference between the number of good supported answers and good unsupported answers was significant ($p = 0.046$).

Table 8 shows what percentages of supported or unsupported answers were assessed as good/poor/etc., i.e. how likely (un)supported answers were to be rated good or otherwise. An alternative way of looking at the data is to examine what percentage of good and poor answers were supported, i.e. how likely good/poor/etc. answers were to be supported or not. Analysing the data this ways shows that 56% of the good answers were supported compared whereas only 44% were unsupported. For poor answers 43% were supported compared to 57% which were unsupported. Again there appears to be some preference for supported answers being more likely to be rated as good and unsupported answers as poor.

We also asked, for each question, what action the assessor would take based on the answer presented – to read the document, accept the answer without reference to the document or simply look for a better answer. Table 9 shows the percentages of responses for this question across supported and unsupported answers, e.g. for 39.28% of supported answers the assessor felt they would read the document against 42.1% of unsupported answers. From Table 9 there is little difference in whether the assessors felt they were more likely to read the document based on whether it is supported or not. However, if an answer is supported then the assessors showed a tendency to accept the answer without confirmation and if the answer was unsupported to look for better answers. Our notion of supporting sources in this study is relatively coarse, we simply present that supporting evidence exists from particular sites; we do not present any information from these sites. However the highest percentage in the accept decision – where the assessor simply trusts the answer based on the system presentation – came from supported. The difference between accepted supported answers compared to accepted unsupported answers was significant ($p = 0.049$).

Table 8
Percentage of supported/unsupported answers rated under different assessment categories

Answer quality	Supported answers ($n = 114$)	Unsupported answers ($n = 100$)
Good answer	**60.53% (*n* = 69)**	54.00% ($n = 54$)
Poor answer	13.16% ($n = 15$)	**20.00% (*n* = 20)**
Can't decide	**18.42% (*n* = 21)**	12.00% ($n = 12$)
Partially good	7.89% ($n = 9$)	**14.00% (*n* = 14)**

Highest value within each answer assessment category shown in bold.

Table 9
Predicted next actions for supported/unsupported answers

Predicted next action	Supported answers ($n = 112$)	Unsupported answers ($n = 95$)
Accept	**43.75% (*n* = 49)**	36.84% ($n = 35$)
Move	16.96% ($n = 19$)	**21.05% (*n* = 20)**
Read	39.28% ($n = 42$)	**42.10% (*n* = 40)**

Highest value within each predicted next action category shown in bold.

6.2. Comparing recent against older answers

Our next variable was the age of an answer. This was a binary variable with two values: recent answers (with dates of six months preceding the assessment date) and older answers (with dates of prior to 2004). For most questions the assessors predicted that the date of answers would not be important (Section 5). This prediction was made on the question alone, before seeing any answers. However, after viewing answers in Interaction Form 2 there seemed to be a preference for recent answers to be rated as good i.e. an average 60% of recent answers were good compared to only 54% of older answer were good, Table 10. The difference between recent answers marked as being good and older answers being marked as good was significant ($p = 0.001$).

In the category of partially good answers there was a more tendency for older articles to be rated partially relevant (9.9% of recent answers rated as partially good compared to 11.7% of older answers). Collapsing partially and good into a single category still shows a preference for more recent information (70% of recent answer were rated as good vs. 66% of older answers). This may reflect a conservative approach to older information, also indicated by the slightly higher use of '*cannot decide*' for older answers.

As before we compared what percentage of good or poor answers were recent or old. In this case 53.2% of good answers were recent compared to 46.7% of good answers which were older so there is some preference for assessing more recent answers as good. 51.4% of poor answers were recent compared to 48.6% of poor answers which were older indicating less effect of date in determining poor answers. So recency may be important in combination with other factors in leading to an assessment of good answers but not for poor answers.

If we compare the assessors' reactions to the answers when classified under time information, there was little in difference in willingness to look for new answers, Table 11. However, there were bigger differences in how the assessors considered they would treat the answers with the assessors being more likely to read documents containing older answers (perhaps for verification) and a preference for more recent answers to be accepted without verification. The difference between recent answers that would be accepted and older answers that would be accepted was significant ($p = 0.031$).

6.3. Comparing good against weak sources

Finally we consider the nature of sources and here we differentiated between two sets of resource: good sources and weak sources. There was little difference for good answers: answers presented as being from good sources were not rated as being good more often than those presented as coming weak sources. However, there was some tendency for answers from weak sources to be rated as poor, Table 12. Collapsing good and par-

Table 10
Percentage of recent/older answers rated under different assessment categories

Answer quality	Recent answers ($n = 111$)	Older answers ($n = 103$)
Good answer	**60.30% ($n = 67$)**	54.36% ($n = 56$)
Poor answer	**16.21% ($n = 18$)**	16.50% ($n = 17$)
Can't decide	13.51% ($n = 15$)	**17.48% ($n = 18$)**
Partially good	9.90% ($n = 11$)	**11.65% ($n = 12$)**

Highest value within each answer assessment category shown in bold.

Table 11
Predicted next actions for recent/older answers

Predicted next action	Recent answers ($n = 105$)	Older answers ($n = 102$)
Accept	**44.8% ($n = 47$)**	36.3% ($n = 37$)
Move	**19.0% ($n = 20$)**	18.6% ($n = 19$)
Read	36.2% ($n = 38$)	**45.1% ($n = 46$)**

Highest value within each predicted next action category shown in bold.

Table 12
Percentage of answers presented as from good/weak sources under different assessment categories

Answer quality	Answers from good source ($n = 112$)	Answers from poor source ($n = 102$)
Good answer	**58.03% (*n* = 65)**	56.86% ($n = 58$)
Poor answer	14.29% ($n = 16$)	**18.67% (*n* = 19)**
Can't decide	15.12% ($n = 17$)	**15.68% (*n* = 16)**
Partially good	**12.50% (*n* = 14)**	8.823% ($n = 9$)

Highest value within each answer assessment category shown in bold.

tially good into a single category shows some preference for answers presented as coming from good sources (70% of good/partially good answers for good sources vs. 65% for weak sources).

When comparing what percentage of good or poor answers were from good or weak sources there is a very slight preference for good answers to come from good sources (52.8% of good answers were from a good source compared to 47.2% of good answers from a poor source) but a larger preference for poor answers to come from weak sources (45.7% of poor answers were from a good source compared with 54.3% from a weak source).

Regarding the predicted next actions there was a slight preference to read documents from good sources but stronger tendency to seek a better answer (move) for answers from a weak source, Table 13.

6.4. Summary

In Table 14 we summarise the main findings against the predicted next action, the data in this table arises from a breakdown of the data into the eight categories given by the three variables. The assessors felt that they would be most likely to *read* the full text of articles from good sources that were presented as being older and where the answer had no supporting evidence. Reading the article in this case may be seen as a way of disambiguating the quality of an answer. The assessors, on the other hand, were least likely to read the full text of articles from good sources that were recent and had supporting sources. The assessors were, in fact, most likely to simply *accept* these answers based on the answer alone and least likely to *move* in search of a better answer when presented with answers in this category. The assessors were least likely to accept, without verification, answers from older articles without supporting sources whether or not the answers were from good or weak sources.

The predicted next action is an implicit assessment of the quality of an answer. This relationship is reinforced when we examine the distribution of answers over the predicted next action (Table 15) e.g. 47.61%

Table 13
Predicted next actions for answers presented as coming from good/weak sources

Predicted next action	Answers from good source ($n = 106$)	Answers from poor source ($n = 101$)
Accept	**41.51% (*n* = 44)**	39.60% ($n = 40$)
Move	16.04% ($n = 17$)	**21.78% (*n* = 22)**
Read	**42.45% (*n* = 45)**	38.61% ($n = 39$)

Highest value within each predicted next action category shown in bold.

Table 14
Summary of most likely predicted next actions

Decision	Category
Most likely to read	Older articles that have no supporting sources but come from a good source
Least likely to read Most likely to accept Least likely to move	Recent articles that have supporting sources and come from a good source
Least likely to accept	Older articles that have no supporting sources
Most likely to move	Recent articles that have no supporting sources and come from a weak source

Table 15
Distribution of answers over predicted next action

Answer quality	Accept ($n = 84$)	Move ($n = 40$)	Read ($n = 84$)
Good	**100.00% (*n* = 100)**	0.00% ($n = 0$)	47.61% ($n = 40$)
Poor	0.00% ($n = 0$)	**80.00% (*n* = 32)**	2.38% ($n = 2$)
Cannot decide	0.00% ($n = 0$)	7.50% ($n = 3$)	**23.80% (*n* = 20)**
Partially good	0.00% ($n = 0$)	12.50% ($n = 5$)	**26.19% (*n* = 22)**

Highest value within each answer assessment category shown in bold.

of answers where the assessor felt their next action would be to read the document were ones rated as good answers. As can be seen in Table 15 answers that are simply accepted are all good answers, answers that are rejected (move) are nearly all poor answers and answers where they would like to read document have level of indecision or are good.

The answers most likely be simply accepted and least likely to be rejected are ones that the assessor sees as good answers (Table 15) and ones presented as being recent answers that have supporting sources and come from a good source (Table 14). The presentation of the answer, therefore, does seem to have an effect on how the assessor views the quality of an answer and the actions performed on the answer.

In Sections 5 and 6 we analysed factors on variables relating to different components (assessors, questions, answers). In Sections 7 and 8 we relate variables from across groups to compare the effect of variables from one component on other components, for example the effect of assessors' topical knowledge on their predicted next action or the effect on question complexity on answer quality.

7. Results on assessor and answers

In this section we compare the assessor variables against the answer variables in order to gauge the effect of the assessor's personal context on the assessment process.

7.1. Knowledge, answer quality and next actions

First we compare how topical knowledge affects the judgements on the answers given and the predicted next actions based on the answers. In Table 16, we compare the percentage of answers rated as good/poor/etc under the variables for topical knowledge.

For the topics where the assessor feels they know little (*not much*) the tendency is to be conservative: relatively low use of the definite categories (good/poor answer) and higher use of the partially good and cannot decide categories compared to the other assessor groups. Indeed the majority of answers for low topical knowledge reflect some uncertainty regarding the quality of the answer which requires resolution from the whole document. This is also indicated in Table 17 where the most common next action for assessors with low topical knowledge is to decide they would read the whole document.

Assessors with higher levels of topical knowledge (*same as most*, *know a lot*) can be more decisive about the quality of answers presented with at least 85% of answers being rated as good or poor and few cases where the

Table 16
Knowledge and answer quality

Answer quality	Level of topical knowledge		
	Not much ($n = 56$[a])	Same as most ($n = 117$)	Know a lot ($n = 36$)
Good	32.14% ($n = 18$)	**70.94% (*n* = 83)**	55.56% ($n = 20$)
Poor	7.14% ($n = 4$)	15.39% ($n = 18$)	**36.11% (*n* = 13)**
Cannot decide	**37.50% (*n* = 31)**	6.84% ($n = 8$)	2.78% ($n = 1$)
Partially good	**23.21% (*n* = 13)**	6.84% ($n = 8$)	5.56% ($n = 2$)

Highest value within each answer assessment category shown in bold.

[a] *n* in this table and subsequent tables in this section refer to the number of answers not the number of assessors in each category.

Table 17
Knowledge and next action

Predicted next action	Level of topical knowledge		
	Not much ($n = 56$)	Same as most ($n = 117$)	Know a lot ($n = 36$)
Accept	**21.43% (n = 12)**	45.30% ($n = 53$)	**55.56% (n = 20)**
Move	8.93% ($n = 5$)	17.09% ($n = 20$)	38.89% ($n = 14$)
Read	**69.64% (n = 39)**	37.61% ($n = 44$)	5.56% ($n = 2$)

Highest value within each topical knowledge category shown in bold.

assessor cannot decide on the quality of the answer or rates the answer as being partially good. These assessors also have a higher than average rate of rating answers as poor. Assessors with the highest level of topical knowledge are far more likely to act on the answer itself without recourse to the full text as, for 95% of answers, the predicted next action is to either accept the answer as presented or move to find a better answer. For the middle range of topical knowledge (*same as most*) the most likely action is one based solely on the answer (*accept* or *move*) but for almost 40% of answers the assessor would seek further information from the document (*read*).

We performed a chi-square test to determine whether the percentage of people with varying levels of topical knowledge was equal in frequency. In other words, was there a difference in the answer assessment pattern between assessors with different levels of knowledge? The p-value <0.05 indicated that the distribution of percentages was different across the topical knowledge, meaning that people with high topical knowledge rated answers more often as poor or good, compared to assessors with low topical knowledge, who could not give definite assessments to the answers.

7.2. Confidence, answer quality and next actions

Next we compare the assessors' declared confidence in being able to detect correct answers to the questions presented against their perception of the quality of the answers presented (Table 18). The results indicate that assessors who declared a high level of confidence in being able to assess answers are certainly more confident as they reach definite (*good/poor*) opinions on nearly 75% of answers presented and there are fewer cases where they felt they could not decide on the quality of an answer (*cannot decide*). Assessors who felt that they were less confident, before seeing any answers, of their ability to recognise correct answers only reach definite opinions on 64% of answers and showed some uncertainty over the quality of an answer on over a third of answers (see Table 19).

However, these more conservative assessors (*depends* category) appeared to be more conservative only about the poor category, i.e. they marked a similar proportion of answers as good but far less as being poor. Instead they are more likely to use the partially good/cannot decide categories to reflect some uncertainty in the quality of the answers. This confidence, however, does not completely accord with their predicted next actions where they have similar rates of the actions except that the conservative assessors predict that they would *accept* more answers and reject fewer. Conservative here may then mean that they tend not to reject possible answers, which accords with the higher rate of cannot decide judgement.

Table 18
Pre-assessment confidence and answer quality

Answer quality	Confidence level	
	Depends ($n = 52$)	very confident ($n = 209$)
Good	**59.62% (n = 31)**	57.89% ($n = 121$)
Poor	3.85% ($n = 2$)	**16.75% (n = 35)**
Cannot decide	**21.15% (n = 11)**	14.35% ($n = 30$)
Partially good	**15.38% (n = 8)**	11.00% ($n = 23$)

Highest value within each answer assessment category shown in bold.

Table 19
Pre-assessment confidence and next action

Predicted next action	Confidence level	
	Depends ($n = 53$)	Very confident ($n = 206$)
Accept	**52.83% (*n* = 28)**	40.29% ($n = 83$)
Move	7.55% ($n = 4$)	**18.45% (*n* = 38)**
Read	39.62% ($n = 21$)	**41.26% (*n* = 85)**

Highest value within each predicted next action category shown in bold.

7.3. Prior expectation, answer quality and next actions

Examining the relationship between prior expectation and answers (Table 20) we see that assessors who have no prior expectation of what answers might look like have a very distinct pattern reflecting a conservative approach to assessment: no answers are rated poor, an almost even split between good and partially good and a high rate of cannot decide decisions. This group of assessors also felt they would read the majority of documents, Table 21. On the other hand assessors who felt they had a good idea of what answers to expect accepted most answers and rated most answers as good.

The two middle categories (*no, but idea in mind* and *partially good*) are interesting in that these categories reflect situations where the assessor has some preconception regarding what answers to expect: either a partially formed idea of what answers to expect (*no, but…*) or have a partial answer and require further information (*partially good*). We might expect, then, that there would be more uncertainty in these cases and a higher use of the categories cannot decide/partially good. In fact the assessors in these categories made more definitive assessments of quality (good/poor) and fewer uncertain assessments (cannot decide/partially good). The assessors in this group also felt inclined to reject more of the answers shown (move) and accept less than the more confident *yes* assessors and fewer than the less confident *no* assessors.

7.4. Summary

From the results presented in detail in Section 7 it is clear that some personal characteristics, such as level of topical knowledge, have an affect on the assessors' perceptions of answer quality and predicted next action. Low levels of topical knowledge or low expectations of answers seem to lead to read decisions whereas higher

Table 20
Prior expectation and answer quality

Answer quality	Prior expectation			
	Yes ($n = 28$)	No but idea ($n = 46$)	Partial ($n = 104$)	No ($n = 31$)
Good	**64.29% (*n* = 18)**	**54.35% (*n* = 25)**	**65.38% (*n* = 68)**	32.26% ($n = 10$)
Poor	7.14% ($n = 2$)	32.61% ($n = 15$)	17.31% ($n = 18$)	0.00% ($n = 0$)
Cannot decide	14.29% ($n = 4$)	8.70% ($n = 4$)	10.58% ($n = 11$)	**35.48% (*n* = 11)**
Partially good	14.29% ($n = 4$)	4.35% ($n = 2$)	6.73% ($n = 7$)	32.26% ($n = 10$)

Highest value within each answer assessment category shown in bold.

Table 21
Prior expectation and next action

Predicted next action	Prior expectation			
	Yes ($n = 29$)	No but idea ($n = 45$)	Partial ($n = 101$)	No ($n = 31$)
Accept	**55.17% (*n* = 16)**	42.22% ($n = 19$)	40.59% ($n = 41$)	22.58% ($n = 7$)
Move	10.34% ($n = 3$)	**33.33% (*n* = 15)**	19.80% ($n = 20$)	0.00% ($n = 0$)
Read	34.48% ($n = 10$)	24.44% ($n = 11$)	39.60% ($n = 40$)	**77.42% (*n* = 24)**

Highest value within each predicted next action category shown in bold.

levels of topical knowledge or clearer expectations of answers lead to accept, or confirmatory, decisions. Knowledge of how the searcher relates to the topic being searched (how much they know, what sort of answers they expect) might help an IR system create more appropriate surrogates for individual searchers. An assessor with high topical knowledge or who is at the later stages of a search may require different surrogates – ones that present different information – than a searcher starting out on a new search or a search who knows little about the topic being searched.

We also investigated a number of other variables in this study. Due to small numbers of data points in some categories for these variables we cannot draw strong conclusions but present indicative findings.

For example, we asked the assessors in Interaction Form 1 to predict how difficult their question might be to answer if they were searching for answers using a standard Web search engine. This was to examine whether the assessors were influenced by their preconceptions of how difficult the question might be. Most assessors either could not predict how difficult the question might be or predicted the question as being fairly difficult, Section 4. Although the data points are small in number there does appear to be a relationship between predicted difficulty and predict next action. For questions the assessor felt were more difficult to answer, they were more likely to accept the system answer or reject it (move) whereas for questions that they felt were easy to answer they were more likely to want to read the document. For difficult questions the assessors are perhaps basing their interactive decision on the system response – placing a higher degree of trust in the system – whereas for easier questions they preferred to read the text containing the answer.

There was little difference in predicted behaviour (predicted next action) or answer assessment between assessors who wanted similar answers in response to their question or those who wanted different answers.

However, when examining the questions themselves, that is the types of relation involved in the question, it appeared that some types of question are easier to judge than others. For evidence questions – ‘‘what evidence supports the involvement of X in Y?’’ – over 93% of answers could be judged as good or poor. For relationship questions – ‘‘what the relationship exists between X and Y?’’ – only 52% of answers could be conclusively evaluated and almost 48% of answers were only partially good or assessor could not decide on the quality of the answer. There also appears to be some relationship between the number of entities and ability to provide clear answers; the two types of question with the lowest proportion of definite (good/poor) answers were those with the highest number of entities (transport and relationship from Table 2) whereas the two types of relationship with the highest proportion of definite answers were those with the lowest number of entities and sub-questions (effect and evidence Table 2).

8. Discussion

This study was designed to investigate the role of specific variables within the process of complex question answering. Our variables spanned the assessor, the answers and the question itself. The results come from our participation in the *ciqa* track. As a result there are distinct limitations to our study: we cannot interview the assessors or gain qualitative information after the assessment process and the numbers of assessors in individual categories are sometimes quite small. Due to the small numbers of assessors in some groups[11] we have avoided carrying out formal analyses on some research questions that have arisen from the analysis presented here. We see this study as primarily exploratory in nature, blending together a number of variables to generate preliminary data and to generate new hypotheses for future investigation. However, the results we have presented do highlight some useful findings, particularly with relation to exploratory search.

In Section 1 we characterised the role of surrogates as helping a searcher decide which action to take next, based on a range of information presented by the search interface. Good surrogates should help searchers make good decisions, e.g. reading documents that will be useful or navigating to good sources of information, but this process of selecting which route to explore may not be neutral. That is, the creation of the surrogates

[11] The small number of assessors in some data groups is a fact that cannot be controlled as easily in TREC as in a more controlled study.

may influence the process of navigation in exploratory search by making some routes seem more attractive to explore. A surrogate that looks like it leads somewhere interesting or appears to link to reliable information could influence the searcher's decision to use the surrogate (by accessing the full document) more than a surrogate that looks less enticing. Alternatively, searchers may only base their exploration on the content of a surrogate and may not be influenced at all by the presentation of the surrogate. Knowing more about what factors affect this use of surrogates can help us design better surrogates and understand more about how they are used within exploratory search.

Our investigation is focussed on complex question answering, a process where a searcher is looking for answer to a complex question. As noted before, a complex question may not be answered with one single text string. Rather it may require answering several smaller sub-questions, or related questions, to provide a good answer. In such a case, a searcher may require to run multiple searches or examine several documents to complete their search. The surrogates may help answer the question, or parts of the question, or simply lead to a document that moves the search on. Exploratory searches, in a more general sense, often have to be broken down into sub-questions to interrogate the database and focus the search.

Although our investigation used many variables, we report in detailed on two important groups: variables that reflect contextual information about the answer presented and the assessors' personal context. The contextual information about the surrogates came in the form of meta-data (information source, date, and quality of source) which were experimentally manipulated. The assessor's personal context was information on aspects such as topical knowledge which were gathered through direct questions. In both cases we analysed the results by looking at the assessors' rating of the answers and their predicted next action. From a methodological point of view we found that asking about the next action – what the assessor would do based on an answer – to be a very useful question for gathering information and one that highlighted many useful results.

From Section 6, the presence of supporting evidence, whether from pretended weak or good sources, seems to be important in judging both the quality of answers and acting upon these answers. The presence of supporting evidence (a number of different sources of information agreeing on the answer), led the assessors to mark more answers as good and, more importantly, to accept the answers. Accepting answers does not mean that a search is complete: many different answers may be accepted and a choice of which to immediately follow is required. Nevertheless, accepting an answer based on a surrogate alone is an important vote of trust in the answer. We also found a similar pattern with the pretended age of answers; answers that were presented as being more recent were accepted more often and rate as being good answers more often.

In exploratory search we assume there is a higher level of uncertainty than with other types of search, such as known-item finding for example, and that searchers require more information clues to help progress in their search. Contextual meta-data, such as those we investigated in this paper, can help provide such clues about worthwhile information avenues and help searchers prioritise which information they wish to view and wish route to follow.

The results presented in Section 6 also indicate that searcher effort might be more usefully employed if the retrieval algorithms prioritised certain types of information before presentation to the searcher. When we asked the assessors to predict what action they would take after reading an answer, they predicted different actions for different types of surrogate. Surrogates from good, recent sources where the answers were supported by other sources were most likely to be accepted without further verification whereas other answers, such as older answers or answers without support, were more likely to require the document being read. Within an exploratory search surrogates that are useful without reading an entire document may be more useful for filtering initial results to a narrower set of potentially useful results.

The assessors' personal context, presented in Section 7, affected how they judged answers and it was clear that the assessors in some circumstances felt they would use surrogates differently from other circumstances. The three variables we examined in depth – topical knowledge, confidence and prior expectation – did show some interesting trends. Assessors with low topical knowledge, for example, have more uncertainty regarding the quality of information presented in surrogates and, consequently, may require better and more detailed surrogates to make information decisions. In exploratory search searchers with low topical knowledge may be common and finding good surrogates for these searchers may be more important than those with high topical knowledge. Similarly assessors who felt less confident about their ability to detect correct answers or who had no prior expectations of an answer also had a high rate of uncertain decisions – marking more answers as only partially good or being unable to decide on the quality of the answer. Assessors with high topical knowledge or high confidence would appear to find it easier to judge surrogates without reference to the entire document and make more decisions based on the surrogate alone. Assessors in this situation may benefit from seeing more surrogates at the interface level to help explore the information space more quickly.

As noted before, some of the categories involve small numbers of assessors which mean we cannot directly answer some interesting questions that have arisen. For example, does the presence of contextual meta-data such as date or information source affect all assessor groups equally or are some groups affected more by the content of surrogates than others? Anecdotally, our results suggest not: assessors who are more uncertain about the assessment process are more likely to be influenced by contextual information than those who have clear ideas of what information to expect. From this exploratory study, however, we have started to map out such questions for future investigation.

9. Conclusion

Our study investigated a number of variables relevant to exploratory search and was centred on an important construct in exploratory search, namely information surrogates. Within this study we have pointed to important factors in understanding how surrogates might be created – what information within a surrogate leads to changes in assessment behaviour. We have also investigated personal factors relating to the assessors themselves. Some of these variables, such as topical knowledge, are already established variables for investigation and we hope to have added more to their understanding within the literature on search behaviour. Other factors, such as prior expectation, are less well investigated but their influence on search behaviour is no less of interest. By this study we hope to contribute to an understanding of how surrogates can be designed and used for exploratory search interfaces.

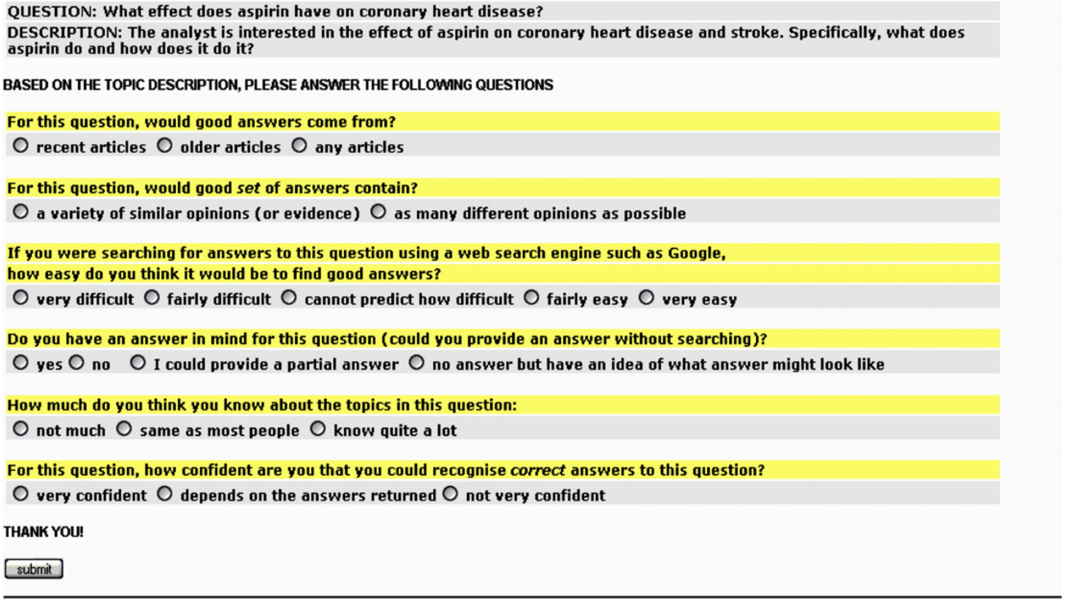
Appendix. Figs. A.1–A.4

www.abc.com
www.baltimoresun.com
www.bbc.co.uk
www.bostonglobe.com
www.cbs.com
www.cnn.com
www.dallasnews.com
www.en.wikipedia.com
www.foxnews.com
www.heraldtribune.com
www.houstonchronicle.com
www.latimes.com
www.miamiherald.com
www.newslink.org
www.nytimes.com
www.reuters.com
www.seattletimes.com
www.theassociatedpress.com
www.the-times.co.uk
www.time.com
www.usatoday.com
www.wallstreetjournal.com
www.washingtonpost.com

Fig. A.1. “High quality” sources.

www.bubblegumfink.com
www.ilab.com
www.politicalgateway.com
www.afnan.com
www.bash.org
www.baz.nl
www.bizzare.com
www.blogger.com
www.buzz.com
www.defamer.com
www.digital-ed.net
www.dopefly.com
www.eopinions.com
www.everything2.com
www.fabiosblog.it
www.fmbbx.ie
www.geocities.com
www.gidforums.net
www.internetmonk.com
www.itnauts.com
www.jackbenimble.co.nl
www.join-me.org
www.jonniesblog.com
www.leif.com
www.marketocracy.com
www.monkeyontheedge.com
www.monkeyrap.com
www.petittoujours.com
www.pointless.co.uk
www.refdesk.sa
www.revie.com
www.starsol.co.uk
www.thejackol.com
www.therealtruth.co.jp
www.thirteen.co.uk
www.typepad.com

Fig. A.2. “Low quality” sources.



QUESTION: What effect does aspirin have on coronary heart disease?

DESCRIPTION: The analyst is interested in the effect of aspirin on coronary heart disease and stroke. Specifically, what does aspirin do and how does it do it?

BASED ON THE TOPIC DESCRIPTION, PLEASE ANSWER THE FOLLOWING QUESTIONS

For this question, would good answers come from?

○ recent articles ○ older articles ○ any articles

For this question, would good *set* of answers contain?

○ a variety of similar opinions (or evidence) ○ as many different opinions as possible

If you were searching for answers to this question using a web search engine such as Google,
how easy do you think it would be to find good answers?

○ very difficult ○ fairly difficult ○ cannot predict how difficult ○ fairly easy ○ very easy

Do you have an answer in mind for this question (could you provide an answer without searching)?

○ yes ○ no ○ I could provide a partial answer ○ no answer but have an idea of what answer might look like

How much do you think you know about the topics in this question:

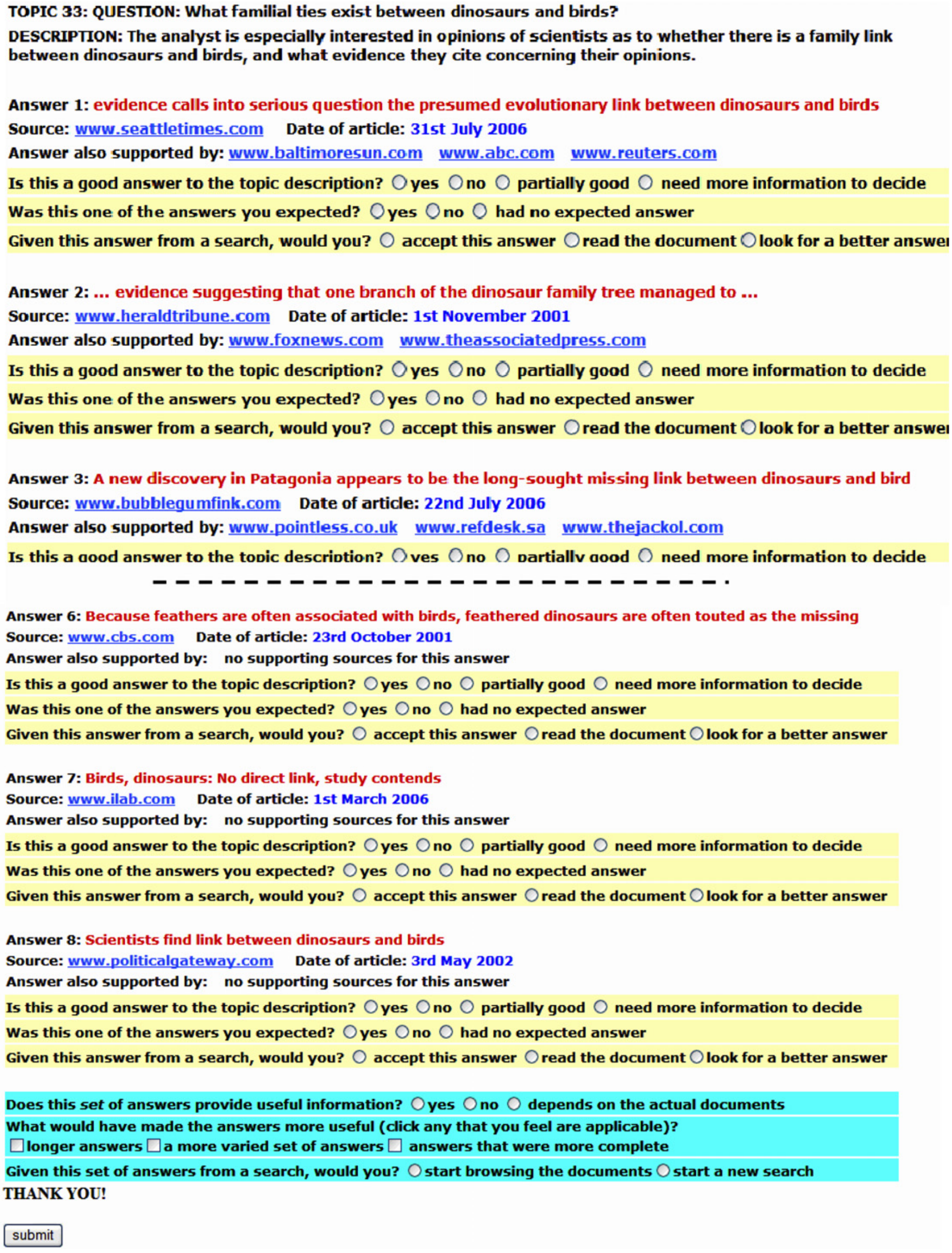
○ not much ○ same as most people ○ know quite a lot

For this question, how confident are you that you could recognise *correct* answers to this question?

○ very confident ○ depends on the answers returned ○ not very confident

THANK YOU!

Fig. A.3. Interaction Form 1.



TOPIC 33: QUESTION: What familial ties exist between dinosaurs and birds?

DESCRIPTION: The analyst is especially interested in opinions of scientists as to whether there is a family link between dinosaurs and birds, and what evidence they cite concerning their opinions.

Answer 1: evidence calls into serious question the presumed evolutionary link between dinosaurs and birds

Source: www.seattletimes.com **Date of article:** 31st July 2006

Answer also supported by: www.baltimoresun.com www.abc.com www.reuters.com

Is this a good answer to the topic description? ○ yes ○ no ○ partially good ○ need more information to decide

Was this one of the answers you expected? ○ yes ○ no ○ had no expected answer

Given this answer from a search, would you? ○ accept this answer ○ read the document ○ look for a better answe

Answer 2: ... evidence suggesting that one branch of the dinosaur family tree managed to ...

Source: www.heraldtribune.com **Date of article:** 1st November 2001

Answer also supported by: www.foxnews.com www.theassociatedpress.com

Is this a good answer to the topic description? ○ yes ○ no ○ partially good ○ need more information to decide

Was this one of the answers you expected? ○ yes ○ no ○ had no expected answer

Given this answer from a search, would you? ○ accept this answer ○ read the document ○ look for a better answe

Answer 3: A new discovery in Patagonia appears to be the long-sought missing link between dinosaurs and bird

Source: www.bubblegumfink.com **Date of article:** 22nd July 2006

Answer also supported by: www.pointless.co.uk www.refdesk.sa www.thejackol.com

Is this a good answer to the topic description? ○ yes ○ no ○ partially good ○ need more information to decide

Answer 6: Because feathers are often associated with birds, feathered dinosaurs are often touted as the missing

Source: www.cbs.com **Date of article:** 23rd October 2001

Answer also supported by: no supporting sources for this answer

Is this a good answer to the topic description? ○ yes ○ no ○ partially good ○ need more information to decide

Was this one of the answers you expected? ○ yes ○ no ○ had no expected answer

Given this answer from a search, would you? ○ accept this answer ○ read the document ○ look for a better answer

Answer 7: Birds, dinosaurs: No direct link, study contends

Source: www.ilab.com **Date of article:** 1st March 2006

Answer also supported by: no supporting sources for this answer

Is this a good answer to the topic description? ○ yes ○ no ○ partially good ○ need more information to decide

Was this one of the answers you expected? ○ yes ○ no ○ had no expected answer

Given this answer from a search, would you? ○ accept this answer ○ read the document ○ look for a better answer

Answer 8: Scientists find link between dinosaurs and birds

Source: www.politicalgateway.com **Date of article:** 3rd May 2002

Answer also supported by: no supporting sources for this answer

Is this a good answer to the topic description? ○ yes ○ no ○ partially good ○ need more information to decide

Was this one of the answers you expected? ○ yes ○ no ○ had no expected answer

Given this answer from a search, would you? ○ accept this answer ○ read the document ○ look for a better answer

Does this *set* of answers provide useful information? ○ yes ○ no ○ depends on the actual documents

What would have made the answers more useful (click any that you feel are applicable)?
□ longer answers □ a more varied set of answers □ answers that were more complete

Given this set of answers from a search, would you? ○ start browsing the documents ○ start a new search

THANK YOU!

Fig. A.4. Interaction Form 2.

References

Allan, J. (2005). HARD track overview in TREC 2004: high accuracy retrieval from documents. In E. M. Voorhees, & L. P. Buckland (Eds.), *Proceedings of 13th text retrieval conference (TREC-13)* (pp. 24–37). NIST Special Publication: SP 500-255.

Barry, C. L., & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information, Processing and Management, 34*(2/3), 219–237.

Bell, D. J., & Ruthven, I. (2004). Searchers' assessments of task complexity for web searching. In S. McDonald, & J. Tait (Eds.), *Proceedings of the 26th European conference in information retrieval (ECIR 04)* (pp. 57–71). Springer.

Blair, D. C. (1980). Searching biases in large interactive document retrieval systems. *Journal of the American Society for Information Science, 31*(4), 217–277.

Chuang, W. T., & Yang, J. (2000). Extracting sentence segments for text summarization: a machine learning approach. In N. J. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 200)* (pp. 152–159). ACM.

Dziadosz, S., & Chandrasekar, R., (2002). Do thumbnail previews help users make better relevance decisions about web search results. In K. Järvelin, M. Beaulieu, R. Baeza-Yates, & S. H. Myaeng (Eds.), *Proceedings of the 25th annual international ACM conference on research and development in information retrieval (SIGIR 2002)* (pp. 365–266). ACM.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science, 44*(3), 161–174.

Johnson, F. (1995). Automatic abstracting research. *Library Review, 44*(8), 28–36.

Kelly, D., & Lin, J. (2007). Overview of the TREC 2006 ciQA task. *ACM SIGIR Forum, 41*(1), 107–116.

Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., et al. (2003). What makes a good answer? The role of context in Question Answering. In M. Rauterberg, M. Menozzi, & J. Wesson (Eds.), *Proceedings of Interact'03 (pp. 25–32)*. International Federation for Information Processing (IFIP).

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., & Gay, G. (2006). The influence of task and gender on search and evaluation behaviour using Google. *Information Processing and Management, 42*(4), 1123–1131.

Lynch, C. (2001). When documents deceive: Trust and provenance as new factors for information retrieval in a tangled web. *Journal of the American Society for Information Science and Technology, 52*(1), 12–17.

Michel, D. (1994). What is used during cognitive processing in information retrieval and library searching? Eleven sources of search information. *Journal of the American Society for Information Science., 45*(7), 498–514.

Park, T. K. (1993). The nature of relevance in information retrieval: An empirical study. *Library Quarterly, 63*(3), 318–351.

Ruthven, I., Baillie, M., & Elsweiler, D. (2007). The relative effects of knowledge, interest and confidence in assessing relevance. *Journal of Documentation, 63*(4), 482–504.

Smeaton, A. F., Lee, H., O'Connor, N., Marlow, S., & Murphy, N., (2003). TV News Story segmentation, personalisation and recommendation. In G. Jones, (Ed.), *Intelligent multimedia knowledge management: Papers from the 2003 spring symposium* (pp. 24–26). Technical Report SS-03-08. American Association for Artificial Intelligence, Menlo Park, California.

Sweeney, S., & Crestani, F. (2006). Effective search results summary size and device screen size: Is there a relationship? *ormation Processing and Management, 42*(4), 1056–1074.

Tombros, A., Sanderson, M., (1998). Advantages of query biased summaries in information retrieval. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, & J. Tait (Eds.), *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2005)* (pp. 2–10). ACM.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation, 56*(5), 540–562.

Vechtomova, O., & Karamuftuoglu, M. (2006). Elicitation and use of relevance feedback information. *Information Processing and Management, 42*(1), 191–206.

Wen, L., Ruthven, I., & Borlund, P. (2006). The effects on topic familiarity on online search behaviour and use of relevance criteria. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), *Proceedings of the 28th European conference in information retrieval (ECIR 2006) (pp. 456–459)*. Springer.

White, R. W., Kules, B., Drucker, S., & Schraefel, M. C. (2006). Supporting exploratory Search: A special section of the communication of the ACM (editorial). *Communications of the ACM, 49*(4), 36–39.

White, R. W., Jose, J. M., & Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing and Management, 39*(5), 707–733.

White, R. W., Ruthven, I., & Jose, J. M. (2005). A study of factors affecting the utility of implicit relevance feedback. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, & J. Tait (Eds.), *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2005)* (pp. 35–42). ACM.

Woodruff, A., Faulring, A., Rosenholtz, R., Morrison, J., & Pirolli, P. (2001). Using thumbnails to search the web. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI 2001)* (pp. 583–590). ACM.

Xie, H. (2002). Patterns between interactive intentions and information-seeking strategies. *Information Processing and Management, 38*(1), 55–77.

Xu, C., Shao, X., Maddage, N. C., & Kankanhalli, M. S. (2005). Automatic music video summarization based on audio-visual-text analysis and alignment. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, & J. Tait (Eds.), *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2005)* (pp. 361–368).