Angle Seeking as a Scenario for Task-Based Evaluation of Information Access Technologies

E. Barker, J. Polifroni, M. Walker and R. Gaizauskas
Department of Computer Science
University of Sheffield
initial.surname@dcs.shef.ac.uk

ABSTRACT

In this paper we propose *angle seeking* as an appropriate task for the evaluation of information access technologies. We first describe angle seeking in the context of writing background to breaking news, analysing the types of information seeking activity it typically engenders, and then present a case study in which angle seeking forms the basis for a task-based evaluation in which a novel associative summary technology is compared with a conventional document retrieval engine. While neither technology is conclusively proved superior, this study both provides insights into these technologies and shows how a novel task-based evaluation can provide new information access technologies with a forum in which to establish themselves.

1. INTRODUCTION

Information seeking is typically not an end in itself, but rather occurs in some wider task setting. For example, information may be sought by someone writing a news report to deadline or carrying out a scientific research investigation or deciding what car to buy. The wider task may (1) require different sorts of information seeking activity (e.g. finding all relevant information, finding just one trustworthy source, developing a hypothesis, answering a factoid question) (2) impose production constraints on the information seeking (e.g. deadlines, form of output) and (3) typically be carried out by users with characteristic knowledge states (e.g. scientific investigations are carried out by those already expert in their field; news reports may be written on topics a reporter may know little about before beginning). Such diversity in tasks and in associated information seeking means "one-size-fits-all" information access tools, such as document retrieval engines, are unlikely to be optimal for every task with an information seeking component. It follows that designers of information access technologies should attend to the differing requirements that different task settings throw up for information seekers (as has long been recognised – [17, 7, 10]). One way to drive this process is to design evaluations of information access technologies that assess how well a tool assists a user in carrying out the wider task in whose service information seeking is undertaken. Doing so may help to liberate information access evaluation from the domination of a few standard evaluation measures, such as precision at rank 10, where relevance of retrieved documents to a query is all that is assessed, rather than the utility of a system for carrying out a task.

Various researchers have carried out evaluation of information access technologies in task settings, both simulated and real, e.g. [16, 8, 21] – see Section 3 below. In our view there is room for much more such work, until the implications of different task settings for information seeking are better understood.

One little explored task setting with significant requirements for information seeking is that of writing background to breaking news events, for example for a natural disaster, a political resignation, or company takeover. This task setting is one whose potential to inform the design of novel evaluations we have already explored [5, 12]. In brief the proposal in this earlier work was to assess the utility of different information access technologies by assessing the quality, as determined by task experts (professional journalists), of the written outputs of those using the technologies. That is, how good are the background pieces produced using information access technology A versus those produced using technology B? Experiments showed high intersubjective agreement between judges when they were asked to rank backgrounders written by different users on the same topic - i.e the task appears well-founded. However, there are various logistical difficulties in mounting an evaluation based on this task. In particular one needs a large pool of journalists prepared to write backgrounders on a range of topics, so that one can control for user and topic; one also needs sufficient qualified judges to assess the resulting background pieces. Since producing and assessing each background piece is a significant amount of work, mustering resources to carry out such an evaluation is not easy. Furthermore, since the resulting information artefacts are so rich (full texts) and the steps taken to produce one are so numerous (including, e.g. all the information seeking that may have contributed to the writer's understanding but did not yield any content that found its way into the final product), this task setting makes it difficult to gain understanding into which aspects of a system's behaviour may have contributed positively or negatively to the overall result.

To address these difficulties with the background writing task while retaining its advantages as an evaluation scenario

– a real task setting with a strong information seeking component – we have focussed on one central but limited aspect of the task: *angle seeking*. Angles, discussed in further detail in the next section, are unifying ideas or overarching propositions which frame or position the information reported in the rest of a text. In news articles they are typically conveyed in the headline or lead sentence. Angle seeking is a key, early step in writing a news article, one which can require extensive information seeking but results in a concise output – usually a proposition expressed in a single sentence. As such, angle seeking is an appealing task for task-based evaluation of information access technologies.

To explore the utility and feasibility of angle seeking as a scenario for task-based evaluation of information access technologies we have made two contributions, which we report in this paper: (1) an analysis of the task, what the task is, the information seeking strategies that may be involved, and why is it an interesting challenge task for information access systems (section2); and (2), the design and execution of an evaluation using angle seeking as the task in order to assess two information access systems – a novel association-based approach and a conventional document retrieval engine. We report this work in the rest of the paper. In section 3 we discuss related work on task based evaluations for IA technologies. Section 4 describes the experiment we have carried out based on an angle seeking scenario, including details of the experimental design, the systems compared and the results of the evaluation. In the final section we draw conclusions about the utility of angle seeking as a scenario for evaluation of IA technologies.

2. ANGLES AND ANGLE SEEKING IN NEWS WRITING

The term "angle" may be used to describe both an information artefact and the activity or process that people carry out in producing such an artefact. The OED, reflecting these two uses, describes an angle as a noun: "a position from which something is viewed or along which it travels or acts", and as a verb: "to present information to reflect a particular view or have a particular focus". The term has currency in a number of domains, such as writing and politics, but it has particular significance for journalists researching and writing background for breaking news stories.

A news wire "backgrounder" is an extended prose piece, of around 500 words, sometimes referred to as a sidebar, which is produced when a news editor deems a particular story worthy of dedicated background material. The function of a backgrounder is not to continue to report details of new events, but rather to provide text that supports and contextualises these events. Speed is essential in the production of news wire content. Yet a backgrounder may appear some time after the early instalments of a story have been published on the wire, since the news room requires details of the breaking news to determine whether the story merits a background piece. Furthermore, research must be carried out, typically against a news archive, so that the journalist has a topic of interest to write about.

Developing a newsworthy "angle" is a key goal in the background research and writing scenario. While a precise definition is not something which is easily articulated, journalists have an intuitive understanding of what an angle is. Interviews with journalists and an analysis of a collection of background news wire texts suggest that we can see an angle as a unifying idea, an organizing construct, which links together information such that it might be used to frame the current event in a narrative text that is both coherent and compelling to an audience. We can find intuitive examples of angles expressed in the headline and the opening statements of a background piece, which journalists refer to as the "lead". Together, these lines provide a summary of what the backgrounder is about. Figure 1 shows two backgrounders for the same news event – Clare Short's resignation from the British Cabinet in 2003 – and illustrates how the angle taken in a backgrounder can profoundly affect the interpretation of a foreground event. In the first piece the angle taken is that the resignation is a consequence of Clare Short's character and the piece goes on to supply details of Short's colourful career. In the second, the angle is that Short's resignation is the continuation of a trend of resignations and sackings that have characterised Blair's government.

12/05/03: Clare Short resigns from Tony Blair's cabinet.

Background 1:

'SERIAL RESIGNER' WHO LED A CHARMED LIFE
The surprising thing about firebrand Clare Short's resignation is that her departure from the Cabinet did not happen much earlier.
Ms Short seems to have lived a charmed life as Secretary for International Development, first by describing the Prime Minister as "reckless" and then by missing a key vote last week on the contentious issue of foundation hospitals.
It looked as though she was almost begging to be sacked.
Those who have watched her progress are still astonished that such a volatile person ... has lasted for so long in the top echelons of Government.
Her reputation as what someone once described as "a serial resigner" was made when she served under Neil Kinnock as Leader of the Opposition ...

Background 2:

BLAIR'S CABINET CASUALTIES
Since sweeping to power in 1997, Tony Blair has had to deal with a string of high-profile resignations from his cabinet - and has felt obliged to remove several other senior ministers himself.
The first to quit following Labour's 1997 landslide triumph was Welsh Secretary Ron Davies, who stepped down after a "moment of madness" ...
Social Security Secretary Harriet Harman and her second-in-command Frank Field were both victims of Tony Blair's first major reshuffle - after apparently falling out ...
Peter Mandelson made history when he became the first Secretary of State to resign twice ...

Source: PA News Archive

Figure 1: Two Backgrounders for the same Event

Attfield and Dowell [3] present a model of journalistic information seeking in the context of the task of writing a news story. While not specifically concerned with the scenario of background news writing, their model provides some insights into how angles are sought and developed and the role that they play in the broader context of a news writing task. Given a news topic assignment by a news editor and a set of product and resource constraints, the three stages in the Attwood and Dowell model are:

1. *Initiation* A provisional angle is established and a deadline and word count constraints are determined. (This usually takes place during the initial assignment brief).

2. *Preparation* The angle is tested and either confirmed or refuted. Potential content is gathered, personal understanding is developed and a plan for the report is evolved. During this stage an assignment-specific collection of materials, paper or electronic, is assembled for later use.

3. *Production* The story is written, consulting the assignment collection, based on the understanding and plan developed so far. The writing process may provoke further information seeking and alteration of the plan.

The notion of an angle is central to their model. It is described by them elsewhere [2] as a "proposition, or central factual claim that is to be made by the report. Where the claim involves some speculation the angle takes the form of a working hypothesis or conjecture" and again as the "clearly focused perspective or guiding idea which determines both a solution's space and the writer's information requirements".

This is a compelling account. However, Attfield and Dowell stop short of pursuing in depth the process by which journalists iteratively gather potential content and refine their understanding of a topic. Based on observations of and interviews with journalists engaged in background seeking and a preliminary analysis of a corpus of information seeking dialogues between journalists where one was seeking background and the other providing it [6], we can elaborate on the processes described in the Attfield and Dowell model:

1. *Initiation* When journalists are seeking background information for a breaking news story, they may not always be provided with an angle. Often their job is to discover and establish angles for the story. They often begin the research process by formulating an idea of a topic, or perspective which they want to explore. This is typically derived from the details of the news story and their background knowledge. It may be as simple as a general topic area, e.g. "hurricanes", or more elaborate, e.g. "despite years of worsening weather, this is the worst storm since 1987".

2. *Preparation* The journalist tests and/or refines the provisional angle. Here the journalist is looking for patterns in the data, such as trends or interesting associations, which in his judgement will be sufficient to form the basis for a compelling background to the news story. Our research suggests that journalists have an expert understanding of the kind of information that needs to be examined in order to develop and support an angle and that they may engage in a number of strategies for finding patterns. We note that these are similar to the strategies Collins and Gentner [11] propose for developing and maniuplating ideas in their prescriptive model of the writing process:

   (a) *Collecting similar events* For example, finding other people who have left a Cabinet Office.

   (b) *Comparison* Comparing the current event with (1) a similar event or (2) a group of similar events (e.g. where does this fit on the scale of things?) – i.e. establishing differences or similarities.

   (c) *Viewing and sorting similar events by different attributes* E.g. arranging examples of protests at pay increases in chronological order; grouping earthquakes by their location; ordering hurricanes by windspeed, in the 5 categories of hurricane.

   (d) *Aggregating over similar events* E.g. numbers of caving accidents in a location; how many of these resulted in serious injuries or deaths.

   (e) *Aggregating over attributes* E.g. total numbers of fatalities in earthquakes in Asia in the past fifty years.

   (f) *Finding extreme similar instances* Based on different attributes, e.g. the earthquake to have killed the largest number of people; the most grisly kind of death etc.

   (g) *Newsworthy similar instances* Similar to (f), finding similar events with a newsworthy characteristic, for example "any funded science projects which have been associated with animal rights activity".

When the journalist is satisfied with the angle, he typically selects content from the materials he has examined in order to support and elaborate on the angle in the written background piece (stage 3 in the Attfield and Dowell model).

3. RELATED WORK: TASK-BASED EVALUATIONS FOR INFORMATION ACCESS

For more than a decade there has been growing interest in task-based user evaluations of information access systems.

One line of such work has concentrated on studying the effect that priming a subject with a task context has on the retrieval of relevant documents from a document collection, e.g. [8, 15]. Hansen and Karlgren [15], for example, consider the effect that a work-task scenario description may have on a reader's assessment of the relevance of documents retrieved in a non-native language they know well. While these sorts of study can yield insights into document retrieval technologies, they cannot, given their focus on document retrieval, give insights into the utility of other information access technologies for tasks that could potentially benefit from them

In contrast to this work, and perhaps less well explored, is work on evaluations in which the assessment has focused on measuring the outcomes of system use. Here the emphasis has been on evaluating information access systems indirectly, assessing how well systems have enabled the user to carry out some wider task, such as: answering a clinical question [16], writing a report [21], revealing the topic structure of an archive [22], etc. Apart from providing valuable insights into the benefits systems may bring to tasks, this approach is notable in that it allows for a comparison of systems which have different outputs, e.g. a list of document headlines vs. summaries of document clusters.

We note the task scenario used in McKeown's work [21] is in the same domain as the angle seeking task we describe in this paper. The authors asked subjects to help write reports for an issue in the news e.g. Hurricane Ivan's effects. Key differences are that they described this as a "fact gathering" scenario, where users answer three related questions about an issue in the news. So, a pre-specified topic guides information seeking and as such there is less emphasis on discovery and analysis for the written result, which is in contrast to what we have observed for the angle seeking task. Other task-based evaluations where the user task

shares some characteristics with those of angle seeking for background news task include Baldonado and Winograd [4] who used the wider task of writing a term paper for a graduate seminar (on either cryptography or neural networks) to focus a comparative evaluation for two variations of the Sensemaker information-exploration interface. They asked users to determine the specific topics and then to write down the titles of one or two promising references. However, the evaluation did not include a measure of the task outcome, focussing instead on the character of the interactions in the different conditions and on user satisfaction. There has been a notable line of work on developing IR applications to support the task of generating and testing hypotheses founded in literature collections, e.g. [24]. But to date, and to the best of our knowledge, evaluations have been restricted to demonstrating by critical example as opposed to more systematic evaluations involving multiple users carrying out multiple tasks in different system conditions.

4. AN ANGLE SEEKING EVALUATION

The information seeking activities typical of angle seeking, identified above in Section 2, suggest that a large range of possible information access systems could be applied to the angle seeking task. Document retrieval, similar event searching, topic tracking technology, overview technologies (e.g. scattergather), association mining techniques could all potentially be of help. Furthermore in current practice journalists are limited to document retrieval systems, but express considerable dissatisfaction with this technology for the task. Therefore there is a strong motivation to investigate the benefits which alternate approaches might bring to the task and for an evaluation which allows potential benefits to be assessed.

Since different information access technologies may differ in their objectives and outputs, in the role of the system in application setup, in the type user interactions, and so on, directly comparing the outputs of such technologies may not be feasible. This is one of the strong arguments mentioned above for devising an extrinsic evaluation.

To do this we proceed as follows: (1) identify a task output; (2) gather task outputs as produced by users who employ different information access technologies; (3) get experts to evaluate the "goodness" of the task outputs. This approach is based on the assumption that if two setups A and B, in which humans work with an information system to complete some task, differ only in their embedded information systems S_A and S_B, and A outperforms B according to some evaluation criteria, then S_A is more positively evaluated that S_B.

For the angle seeking task, we propose a setup consisting of a journalist together with an information access system and a text information source, or digital archive. Input to the setup is a breaking news story. The subject is asked to read this story and use the information resources to find as many angles for a background piece to the new event as possible within 15 minutes. The output is a list of angles and for each a list of documents which support the angle.

For this task we can identify various possible evaluation criteria: user satisfaction, effort, quality of output from the setup (the angle plus supporting content), and time to complete. To carry out an evaluation we must operationalise these criteria as measures. For example, user satisfaction could be measured by a post task questionnaire; effort by the number and type (productive or non-productive) of user interactions with the system, quality by experts' judgements on the angles plus supporting documents found by users. In the case study reported below we used two evaluation criteria only: (1) subjects' perception of the utility of each interface as a mechanism for searching for background information; and (2) the quality of the information provided by each interface.

In the rest of this section we describe a case study in using angle seeking as a scenario for evaluation two information access technologies. We first provide some details of the technologies, describe the design of the experimental setup in more detail and then present results.

4.1 Technologies Compared

A new technology that might be suitable for the task of seeking angles for breaking news events is what we refer to as "associative summaries", an approach that takes semantically annotated documents that are topically related to the breaking news event, looks for strong associations in the annotations, and then presents these associations as indexes to document clusters. The intuition here is that these summaries will give the user an idea of what content is available in the archive and of patterns in the data. Our hypothesis is that, given that angle seeking is a task that frequently requires a new event to be seen as the continuation of a pattern or trend, then a technology that actively discovers patterns in the data in areas topically related to the new event will be of more benefit than one which leaves the user, who may know little about either the topic or the archive content, to drive the information seeking process himself. In the evaluation below we compared associative summaries with a conventional document retrieval system, as a baseline, using the angle seeking task as an evaluation scenario.

4.1.1 Associative Summaries for Information Access

The associative summary technique may be summarised as follows (for full details see [23]). First it is assumed that an archive has been semantically annotated for entity types such as *person*, *location*, *date*, *organization* and so on and for *keyphrases* where the latter are single or multiwords terms that are indicative of document content (a variety of techniques exist for identifying these, such as [25]). For the experiment reported here a subset of these entity types was selected, consisting of just *person*, *location* and *keyphrase*.

The technique is applied to a topically coherent subset of documents from the archive. This subset, called the *topic set*, is assembled using a query to a search engine running over the archive (e.g. "China AND polution" – in the experiment one query was selected for each breaking news story for which subjects had to find angles). From the lead segment of each document in this topic set a fixed number of most frequently occurring instances of each of the nominated entity types is identified – in the current case the ten most frequent persons, locations and keyphrases. For each document in the topic set a binary vector representation of length 30 is then created, one position for each of the 30 frequently occurring entities, a 1 in any position in the vector indicating that there is a mention of this entity in this document.

The vector representations of the topic set are input to a clustering algorithm, in this case a modified version of Predictive Apriori with bottom-up agglomerative clustering [1]. The resulting clusters, representing potentially signif-

icant associations, are presented to users using one of two interfaces. The first interface (called the *Full Associations* interface below) shows associations grouped according to the entity types found in the associations. So, for example, all associations involving say keyphrases and locations – for instance "river Russia spill" in the "China AND pollution" topic set – are shown together, as are all associations involving just persons, and so on. Selecting any association takes the user to a page listing the titles of all documents in the archive (not just in the topic set) containing occurrences of the terms in the association (in our example, all documents containing occurrences of "river", "Russia" "spill" "China" and "pollution"). The second interface (*Combined Associations*) simply shows all associations, without grouping them by the types of entities found within them. Again selecting any association leads the user to page listing titles of all documents in which the association is instantiated and links to the full documents.

4.1.2 Baseline Document Retrieval System

The baseline system was the document retrieval system within Ontotext's KIM semantic annotation platform [19], itself built on the Lucence open source search engine library [1], an implemention of the vector space model. For the baseline interface, users constructed search terms themselves for the breaking news story and typed these directly into an interface to Ontotext's search facility. Rather than use the interface provided by Ontotext, a separate page was designed that preserves the look-and-feel of the other two interfaces.

4.1.3 Data Resources

Ontotext Corporation provides an interface to roughly 500,000 news articles from sources such as Reuters, the PA, ABC News, the BBC, and CNN. Each document has been automatically annotated for keyphrases and named entities using the KIM platform. For the experiments described here, Ontotext provided a Java applet that enabled us to query the archive by key term and receive a set of semantically annotated documents in XML format in return.

4.2 Experimental Design

We recruited a total of 18 subjects on the basis of their experience in news writing. Participants included sixteen MSc graduate students in the Department of Journalism Studies, University of Sheffield, and two professional journalists working for the Sheffield *Star*. We asked each participant to read a breaking news story and then, using one of the three interfaces to the Ontotext news archive described above, to find angles that might help in the preparation of the best possible background to the story. We set a 15 minute time limit for the task and asked subjects to find as many good angles as possible within the allotted time. When satisfied with an angle participants were to write down the angle (e.g., "Previous chemical spill in river in China") and to save any documents which supported the angle.

To help them carry out this task, we provided a short scenario which asked a participant to imagine him/herself as a reporter working for an international newswire agency and that the news editor had called for a 500 word background report for the wire to support a breaking story. Each subject carried out three tasks, each on one of three topics, real news stories chosen from AP newswire via Google, from within two weeks of the date of the start of the experiment. Of the three breaking news stories, one was about riots in France following the election of Nicolas Sarkozy as President, one was about a threatened lawsuit by the European Union against Microsoft, and one about new Chinese government measures to address pollution. These topics contain a range of event types/entities: one focussed on a person (e.g. Nicolas Sarkozy), one focussed on an organization and a political entity (e.g. Microsoft and the EU), and one focussed on a country and a keyword (e.g. China and pollution).

Each subject completed three tasks by interacting with each of three interfaces in turn, in a within-subject design.

We varied the interface order across subjects in order to assess the effects of the interface on user behavior and experimental judgment. Across the 18 subjects, each interface was used six times as the first, second, or third interface, respectively. To mitigate the confounding effects of story type on subjects' perception of the interface, we did not also vary story type. Each subject completed the Nicolas Sarkozy task first, the EU/Microsoft task second, and the China task third.

Subjects were given a sample breaking news story as a "warm-up". The three interfaces used in the experiment had been configured for the warm-up story, and subjects were given as much time as they wanted to work through the warm-up task while familiarizing themselves with the interfaces. Experimenters were present to answer questions at this point.

After completing the warmup, subjects returned to the main experimental page, where they were asked to indicate, in general, how familiar they were with each of the topics used in the experiment, rating their familiarity on a Likert scale from 1 to 5, with 5 being "Very familiar".

Subjects then carried out the experiment with a fifteen minute time constraint per task. After finishing each task, subjects were asked to answer two questions about each interface, using a 5-point Likert scale:

- How confident are you that you were able to fully explore the contents of the corpus? (with '1' indicating *Not confident* and '5' indicating *Very confident*)
- Would you use such a system again? (with '1' indicating *Not likely* and '5' indicating *Very likely*)

User input on the first of these is analyzed as the *confidence* metric in Section 4.3; the second as the *reuse* metric. After completing all three tasks and seeing all three interfaces, subjects ranked each interface by its usefulness, again on a Likert scale from 1 to 5 ('1' being *Not useful* and '5' being *Very useful*). This is called the *rank* metric in Section 4.3.

Finally, users were asked to tell us what they liked best and least about each interface, using a free-form text box. This last set of questions was optional, but all subjects except one provided feedback here.

4.3 Results and Analysis

4.3.1 User judgments/input

Overall, the two cluster-based interfaces were ranked as top-choice by our subjects 56% of the time and as either top or equivalent to the *Baseline* 67% of the time.

The average rank users assigned to each of the interfaces is shown in the second column of Table 1. Overall, the

[1] lucene.apache.org

Interface	Rank (average)	Confidence (average)	Use again (average)
Full	3.11	3.11	3.06
Combined	2.94	3.17	3.0
Baseline	3.28	3.33	3.7

Table 1: The scores users assigned to each interface, for overall rank, confidence, and reuse.

Topic/ order	Rank (avg)	Confidence (avg)	Reuse (avg)	Fam. (avg)
Sarkozy	3.06	2.94	3.22	2.722
Microsoft	2.83	3.06	2.83	2.22
China	3.44	3.61	3.67	2.33

Table 2: The scores subjects assigned each story for overall rank, confidence, reuse, and familiarity. The order in the table reflects the order in which the subjects saw each story topic.

highest ranking interface was the *Baseline* system. Preference for the *Baseline* was not significant, however, compared with the *Full Associations* interface, based on paired *t*-tests and a multivariate analysis of variance (MANOVA; Wilks' Lambda, $F(2,16) = .423$, $p = .662$). This lack of significant difference indicates that subjects had no strong preferences among the three interfaces.

Users' confidence in the usefulness of each interface for exploring the archive was also not significantly different in paired *t*-tests and a MANOVA (Wilks' Lambda, $F(2,16) = .242$, $p = .788$). The third column of Table 1 shows these scores. For the *reuse* metric, reflecting users' response to the question about using each particular interface again, averages are shown in the fourth column. As with rank and confidence, paired *t*-tests and a MANOVA showed no significant differences (Wilks' Lambda, $F(2,16)=1.8$, $p = .198$).

We next performed a series of ANOVAs using each of the subjective measures elicited from users as the dependent variable, and type (i.e., *Full Associations*, *Combined Associations*, or *Baseline*) and story topic (i.e., Sarkozy, Microsoft, or China) as independent variables. Table 2 shows these subjective measures as they correspond to story topic. We did not find significant effects or interactions with the independent variable *rank*. However, we found a marginal effect of story topic on confidence ($p = .098$, $F = 2.24$, $df = 2$), although no interaction effects. We also found a slightly stronger, though still marginal effect of story topic on reuse ($p = .066$, $F = 3.4$, $df = 2$), again with no interaction effects.

Recall that, because stories were always presented to users in the same order, story topic is a proxy for order in our analysis. Although there was a marginal effect of story topic on confidence, there was a significant correlation between users' confidence in the systems and story topic (i.e., order). Users' confidence increased monotonically over the course of the experiment, regardless of the order of the interface (see Table 2). The effect of order on user judgment has been seen elsewhere in search-based tasks [9], although with a much smaller subject population. For our subjects, confidence grew as they progressed through the experiment. This suggests that a longitudinal study using these interfaces might yield interesting results.

One possible explanation for the effects shown by story topic is the users familiarity with the story itself. The final column in Table 2 shows familiarity scores by topic, which were not significantly different. Familiarity had a marginal effect on confidence (ANOVA, $p = .06$, $F = 2.9$, $df = 3$), but no effect on rank or reuse. The topic users expressed the greatest familiarity with *a priori*, Nicolas Sarkozy, was also the one that had the lowest confidence scores. The interface to stories about China, about which users had expressed a lower degree of familiarity, had the highest confidence scores.

The lack of a significant difference in any of our subjective measures matches what has been found elsewhere in the literature [14, 18] when comparing interfaces that process data to a Google-like baseline. Simple keyword search interfaces are well-known and frequently used tools, and it is not easy in an hour-long experiment to show superior benefits from a new interface. The fact that two associative summary interfaces were preferred more than half the time is a positive indicator of the utility of associative summaries. Users' confidence grew as they progressed through the experiment, even when they were using the associative summaries in later stages. This indicates that all interfaces met subjects' information-seeking needs to some degree.

4.3.2 *Expert judgment*

In addition to the ratings we elicited from subjects, we also asked a Senior Lecturer in the Department of Journalism Studies at the University of Sheffield, who teaches on the the topic of angles in news stories, to serve as an expert judge on subject output. This expert judge was presented with 54 separate "packages" of documents, one for each of the stories (3) used by each of the subjects (18) to complete their tasks. Each package consisted of a set of angles, followed by the stories the subject found to support each angle [2].

The expert judge read the breaking news story for each topic/interface and answered three questions about each package of angles and background stories. Answers to the questions, listed below, were on a Likert scale of 1 to 5, with 1 indicating a negative opinion:

- How would you rank this package for its usefulness in building a background for the breaking news story?
- How would you rank this package for richness/comprehensiveness of background?
- How would you rank this package for originality/unexpectedness (i.e., does it contain something that is both novel and helps contextualize the event)?

Each package was examined blindly, i.e., the expert had no idea who created the package or what interface was used.

[2]A concern in this part of the experimental protocol was that it was not possible to elicit judgments from more than one expert, given the level and specificity of expertise needed to rate background material, and the time required to examine 54 sets of background angles and supporting documents. Because this is the first time, to our knowledge, that this technology has been both used and evaluated by experts in the same field, we felt that one set of judgments here would contribute to an understanding of the usefulness of the technology, while helping refine an evaluation protocol for use in follow-up experiments.

Interface	Usefulness	Richness	Originality
Full	2.67	2.22	2.06
Combined	2.72	2.61	2.28
Baseline	3.22	3.28	2.83

Table 3: Rankings from the expert judge for each interface, on usefulness, richness, and originality.

Topic	Usefulness	Richness	Originality
Sarkozy	2.83	2.22	2.11
Microsoft	2.78	2.78	2.5
China	3.0	3.11	2.56

Table 4: Rankings from the expert judge for each topic, on usefulness, richness, and originality.

High-achieving subjects			
Interface	Avg. rank	Conf.	Reuse
Full	3.50	3.50	3.50
Combined	3.00	2.75	3.00
Baseline	2.00	3.00	2.50
Low-achieving subjects			
Interface	Avg. rank	Conf.	Reuse
Full	3.25	3.00	3.25
Combined	2.75	3.50	3.50
Baseline	4.00	3.75	4.00

Table 5: Rankings from users, shown by groups, reflecting those whose angles were ranked highly by the expert judge and those that were ranked poorly.

The results of the expert judgments were used to associate measures of *usefulness*, *richness*, and *originality* (corresponding, respectively, to the questions above) with the other experimental variables. Table 3 shows the expert measures as a function of the interface used for the package. Table 4 shows the expert ratings for usefulness, richness, and originality, by topic.

Examined by interface, the baseline performs best along all dimensions. The differences are not significant, however, for usefulness or originality, although a MANOVA indicates significant differences in richness (Wilks' Lambda, $F(2,16) = 4.6$, $p < .05$). Paired t-tests showed a significant difference between the richness scores for *Full Associations* vs. *Baseline* interfaces ($p = .006$, $df = 17$) and for *Combined Associations* vs. *Baseline* interfaces ($p = .048$, $df = 17$).

Scores associated with topic also show significant differences only for richness measured by a MANOVA (Wilks' Lambda, $F(2,16) = 3.8$, $p < .05$). Paired t-tests showed significant differences in richness between Sarkozy angles and China angles ($p < .05$, $df = 17$).

The expert judge used in this experiment was able to provide insight into the interaction between topic type and richness, in an interview conducted after the judgments were elicited. He hypothesized that the topic with the highest scores along all dimensions, "China and pollution", lent itself naturally to the type of background information that he would score highly for richness. He further hypothesized that angles found for the other two topics would be, by their nature, not as interesting from his perspective.

After rating each of the packets, we asked the expert to go back through the angles found for each of the three topics and flag the angle+story combinations that he thought were most interesting. Not surprisingly, he found none that he felt were outstanding along this dimension for either the Sarkozy or the Microsoft story. However, he did find two for the China story, both from the same subject. These two stories were both found using the *Full Associations* interface, and, furthermore, were found by a subject that rated that interface the highest for usefulness. The fact that the only angles felt to be truly outstanding by the expert were found by the same subject, suggested that individual subject performance might be an interesting dimension to investigate.

4.3.3 Examining subjects by performance

Previous research has examined the effects of issues such as personality [13], and experience [20] on users' acceptance of a variety of technologies related to information presentation. Here, we investigate correlations between measures from the expert judge, which evaluate subjects' abilities to identify background for stories, and the preferences expressed by users.

Out of 18 subjects, four fell into a group that scored cumulatively highest on the measures of usefulness, richness, and originality. Those four subjects are classified as the *high achieving* set. Four subjects fell into a group that scored cumulatively lowest on these same three measures and those four subjects are classified as the *low achieving* set. Table 5 shows user-elicited scores for each interface, divided by the high-achieving and low-achieving subjects.

High-achieving subjects ranked the two associative clustering interfaces the highest. The differences here are significant between the rank of the *Full Associations* and the *Baseline* interfaces ($p < .05$, $df = 3$), with the baseline scoring significantly lower. Low-scoring subjects tended to prefer the baseline (although not significantly). Although the dataset is small, and we have only the opinion of one expert judge, this result seems to indicate that the associative summaries technology was able to be used effectively by high-achievers at the task.

5. CONCLUSION

We have introduced the task of angle seeking in the background news domain and shown it to be a rich task with much potential for focussing new applications and evaluations of information access technologies. We also presented an angle seeking evaluation which incorporates an expert's assessment of task outcome. Moreover, we have demonstrated this evaluation in an experiment which compared users' performance on the task in three different information access system setups, two using variants of a novel "associative summary" technology based on finding associations in semantically annotated text and a third using a conventional IR search engine. While the results were inconclusive, in so far as they were not able to establish whether the new technology was more effective in this task setting, they provide important insights into the relative strengths and weaknesses of the new and the conventional information access technologies. In particular by showing that the new technology was preferred by users who were good at the task, the evaluation has helped to establish the potential utility of a new technology, validating the observation we made at the outset that appropriate design of evaluations can help

advance technologies for information access.

6. ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the UK EPSRC grant GR/R91465/01. They would also like to thank sincerely members of the University of Sheffield Department of Journalism for their enthusiastic participation in many aspects of this work, most especially Jonathan Foster, Bob Bennett and David Holmes. Finally we would like to acknowledge the UK Press Association for access to their archive and for expert advice and comment and Ontotext for access to their system and news archive.

7. REFERENCES

[1] J. Alipio. Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In *Proc., 2004 SIAM Int'l. Conference on Data Mining*, 2004.

[2] S. Attfield, A. Blandford, and J. Dowell. Information seeking in the context of writing: a design psychology interpretation of the 'problematic situation'. *Journal of Documentation*, 59(4):430–453, 2003.

[3] S. Attfield and J. Dowell. Information seeking and use by newspaper journalists. *Journal of Documentation*, 59(2):187–204, 2003.

[4] M. Q. W. Baldonado and T. Winograd. SenseMaker: An information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI'97*, pages 11–18, Atlanta, Ga., 1997. ACM Press, New York.

[5] E. Barker and R. Gaizauskas. Evaluating Cub Reporter: proposals for extrinsic evaluation of journalists using language technologies to access a news archive in research. In A. Bailey, I. Ruthven, and L. Azzopardi, editors, *Proceedings of the Workshop on Evaluating User Studies in Information Access, 5th International Conference on Conceptions of Library and Information SCience,(COLIS 2005)*, 2005.

[6] E. Barker, R. Higashinaka, F. Mairesse, R. Gaizauskas, M. Walker, and J. Foster. Simulating Cub Reporter dialogues: The collection of naturalistic human-human dialogues for information access to text archives. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, May 2006.

[7] N. J. Belkin, P. G. Marchetti, and C. Cool. Braque: design of an interface to support user interaction in information retrieval. *Inf. Process. Manage.*, 29(3):325–344, 1993.

[8] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.

[9] P. Borlund and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3):225–250, 1997.

[10] K. Bystrom and K. Jarvelin. Task complexity affects information seeking and use. *Information Processing & Management*, 31(2):191–213, March-April 1995.

[11] A. Collins and D. Gentner. A framework for a cognitive theory of writing. In L. W. Gregg and E. Steinberg, editors, *Cognitive processes in writing: An interdisciplinary approach*, pages 51–72. Lawrence Erlbaum Associates, 1980.

[12] R. Gaizauskas and E. J. Barker. Mice from a mountain: Reflections on current issues in evaluation of written language technology. In J.Tait, ed., *Charting a New Course: Natural Language Processing and Information Retrieval*, pages 195–238. Springer, 2005.

[13] D. Goren-Bar, I. Graziola, F. Pianesi, and M. Zancanaro. The influence of personality factors on visitor attitudes towards adaptivity dimensions for mobile museum guides. *User Modeling and User-Adapted Interaction*, 16(1):31–62, 2006.

[14] C. Gutwin, G. Paynter, I. Witten, C. NevillManning, and E. Frank. Improving browsing in digital libraries with keyphrase indexes. Technical report, Dept. of Computer Science, University of Saskatchewan, 1998.

[15] P. Hansen and J. Karlgren. Effects of foreign and task scenario on relevance assessment. *Journal of Documentation*, 61(3):623–639, 2005.

[16] W. Hersh, J. Pentecost, and D. Hickam. A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science*, 47(1):50–56, 1996.

[17] P. Ingwersen. *Information Retrieval Interaction*. Taylor Graham, 1992.

[18] H. Joho, M. Sanderson, and M. Beaulieu. A study of user interaction with a concept-based interactive query expansion support tool. In *Proc., 26th European Conf. on Information Retrieval*, pages 42–56, 2004.

[19] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, 2004.

[20] S. K. Lippert and H. Forman. Utilization of information technology: Examining cognitive and experiential factors of post-adoption behavior. *IEEE Trans. on Engineering Management*, 52(3), 2005.

[21] K. McKeown, R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg. Do summaries help? In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA, 2005. ACM.

[22] P. Pirolli, P. Schank, M. Hearst, and C. Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 213–220, New York, NY, USA, 1996. ACM.

[23] J. Polifroni. *Enabling Browsing in Interactive Systems*. PhD thesis, University of Sheffield, Department of Computer Science, January 2008.

[24] M. Weeber, H. Klein, L. T. W. de Jong-van den Berg, and R. Vos. Using concepts in literature-based discovery: simulating swanson's raynaud-fish oil and migraine-magnesium discoveries. *J. Am. Soc. Inf. Sci. Technol.*, 52(7):548–557, 2001.

[25] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. KEA: Practical automatic keyphrase extraction. In *ACM DL*, pages 254–255, 1999.