

An evaluation of adaptive filtering in the context of realistic task-based information exploration

Daqing He ^{a,*}, Peter Brusilovsky ^a, Jaewook Ahn ^a, Jonathan Grady ^a,
Rosta Farzan ^b, Yefei Peng ^a, Yiming Yang ^c, Monica Rogati ^d

^a School of Information Sciences, University of Pittsburgh, 135 N. Bellefield Avenue, Pittsburgh, PA 15256, USA

^b Intelligence Systems Program, University of Pittsburgh, 5113 Sennott Square, Pittsburgh PA 15260, USA

^c Language Technologies Institute, Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213, USA

^d Computer Science Department, School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213, USA

Received 3 January 2007; received in revised form 12 June 2007; accepted 9 July 2007

Available online 10 September 2007

Abstract

Exploratory search increasingly becomes an important research topic. Our interests focus on task-based information exploration, a specific type of exploratory search performed by a range of professional users, such as intelligence analysts. In this paper, we present an evaluation framework designed specifically for assessing and comparing performance of innovative information access tools created to support the work of intelligence analysts in the context of task-based information exploration. The motivation for the development of this framework came from our needs for testing systems in task-based information exploration, which cannot be satisfied by existing frameworks. The new framework is closely tied with the kind of tasks that intelligence analysts perform: complex, dynamic, and multiple facets and multiple stages. It views the user rather than the information system as the center of the evaluation, and examines how well users are served by the systems in their tasks. The evaluation framework examines the support of the systems at users' major information access stages, such as information foraging and sense-making. The framework is accompanied by a reference test collection that has 18 tasks scenarios and corresponding passage-level ground truth annotations. To demonstrate the usage of the framework and the reference test collection, we present a specific evaluation study on CAFÉ, an adaptive filtering engine designed for supporting task-based information exploration. This study is a successful use case of the framework, and the study indeed revealed various aspects of the information systems and their roles in supporting task-based information exploration.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Task-based information exploration; Exploratory search; Evaluation framework; Adaptive filtering; CAFE; User study; Intelligence analysts

* Corresponding author. Tel.: +1 412 6242477; fax: +1 412 64870001.

E-mail addresses: dah44@pitt.edu (D. He), peterb@pitt.edu (P. Brusilovsky), jaa38@pitt.edu (J. Ahn), jpg14@pitt.edu (J. Grady), rostaf@pitt.edu (R. Farzan), yep3@pitt.edu (Y. Peng), yiming@cs.cmu.edu (Y. Yang), mrogati@cs.cmu.edu (M. Rogati).

1. Introduction

Studies show that Web searchers have been driven by various information needs, and their corresponding search activities can be classified into three types with labels “lookup”, “learn” and “investigate” respectively (Marchionini, 2006). In this classification, lookup aims at finding specific and possibly existing information in the collection, and is often called “known item” search. Both searching to learn and searching to investigate are more challenging types of search, which require iterative efforts with interpretation, synthesis, and evaluation of the information returned at each iteration. In the literature these types are called exploratory searches (White, Kules, Drucker, & Schraefel, 2006). Although current Web search engines (e.g., Google) do a reasonably good job in lookup searches, their support of the user’s needs during exploratory search is still far from adequate.

A specific type of exploratory search that is considered in this paper is known as *task-based information exploration*. Here the information needs and the corresponding search processes are heavily influenced by the task assigned to the user. This kind of exploratory search is typical for a range of professional users, such as intelligence analysts. To understand the problems and the needs of task-based information exploration, let’s consider the anatomy of an analyst’s work on one task. The work starts with a given Request For Information (RFI). A RFI typically contains one overall investigation goal and a set of more specific questions that call for more information related to a seminal event. The analyst’s job is to collect relevant and useful information from various sources to answer the RFI questions and to prepare a short (one to two-page) report called a *point paper*. The point paper should summarize what has been found and to make specific recommendations for certain actions. For example, if the seminal event is an escape of seven inmates from a prison in Texas, an RFI could be issued to ask for more information and potential actions that are useful to coordinate the recapture of those inmates. It may list such specific questions as where and when they were last sighted (location, time/date), whether they were armed, which kind of vehicle they may be driving, and what steps (e.g., rewards, posting, etc.) have been taken so far by the police to facilitate the recapture.

As the example shows, task-based information exploration driven by complex realistic tasks is a challenging exercise. Its complexity is further increased since the events associated with the task are frequently evolving during the exploration. A typical task requires multiple searches over several sessions to explore the information space and to obtain accurate and updated information. This puts task-based information exploration in the same row with other kinds of exploratory searches and far from simpler lookup searches. Yet, the presence of a clearly defined task makes tasks-based exploration special in at least two aspects. The first aspect is related to search system engineering. An intelligent system can build a model of the given task and provide a better level of user support in the process of exploration. The second one is related to evaluation. The presence of the task and a special structure of task-based information exploration make it possible to build a dedicated evaluation framework to assess and compare different exploration-support systems.

The work presented in this paper is related to both aspects listed above. Our team is involved in a large-scale project that focuses on developing a new generation of systems to support task-based exploratory work of intelligence analysts. While the majority of projects in this area focus on developing visualization systems that help analysts to examine the information space (Acosta-Diaz et al., 2006; Gotz, Zhou, & Aggarwal, 2006; McColgin, Gregory, Hetzler, & Turner, 2006; Wong, Chin, Foote, Mackey, & Thomas, 2006), we focus on artificial intelligence techniques. Among other techniques, our team explores the application of adaptive filtering (AF) (Hanani, Shapira, & Shoval, 2001) to support exploratory searches, especially task-based information exploration. Adaptive filtering is a popular technology in the field of user-adaptive systems. For every user of an information system, an AF engine builds a user model by inferring users’ tasks, interests, preferences, and knowledge from either explicit feedback from users’ relevance judgments or implicit feedbacks by observing the users’ search and browsing activities. The model is then used by the system to predict and recommend potentially relevant information to the user. Such capabilities make AF very attractive in the context of information exploration. For plain search engines, which rely on users’ querying skills, the exploratory search context is a disadvantage, because user information needs and search scope are poorly defined. In contrast, for an AF engine, this context is

advantageous since users' judgments and search/browsing activities, which are abundant in the process of information exploration, provide a good source of information for user modeling. The task-based information exploration is especially attractive for the use of AF since task modeling is both much easier and more reliable in the presence of a clearly defined task.

To explore the potential of information filtering in an exploratory search context, a part of our team developed an innovative AF engine known as Carnegie Mellon Adaptive Filtering Engine or *CAFÉ*. The mechanism of this engine is presented in (Yang et al., 2007). After the first version of *CAFÉ* was developed, we faced the problem of its evaluation. We wanted to confirm our hypothesis that adaptive filtering provides good value in an exploratory search context (at least in comparison with traditional search) and to measure the effect of using *CAFÉ*. The evaluation part, however, appeared to be at least as hard as the development part. As pointed out in (White, Muresan, & Marchionini, 2006), proper evaluation of various information access technologies in the context of exploratory search is a challenge since known evaluation frameworks are limited to those that support minimal to no human-computer interaction. Researchers in the area of exploratory search argue that existing evaluation metrics and methodologies are not adequate in the exploratory search context because of the exploratory search's high level interactions between human and computers. As we discovered, the evaluation of information filtering in the context of task-based exploration is not an exception – all known evaluation frameworks cannot be applicable directly (see Section 2).

To perform a sound evaluation of *CAFÉ* we had to answer the following research questions:

- What kinds of data resources are required to support the evaluation of adaptive filtering in realistic task-based information exploration? Are there any test collections that could serve as a source to build the target collection, and how do they have to be expanded?
- How can we develop an appropriate experiment design to include users in their task-based information exploration, where the process is close to real search scenarios, and the evaluation measures are straightforward and meaningful to the users involved?
- How can we analyze the results of the evaluation experiment? What kinds of measures are relevant in this context to assess the user and system performance? What are the appropriate metrics to express the desired information in numeric form?

As a result, our work on the evaluation of *CAFÉ* went far beyond evaluating one specific adaptive filtering engine. As a byproduct of this process, we developed a complete evaluation framework for studying various information access techniques in task-based information exploration context. We started by analyzing the specific evaluation needs of task-based information exploration needs. Based on this analysis, we prepared an annotated collection of resources that can support the required level of evaluation. We developed the structure of the evaluation process, which is based on humans performing complex tasks and resembles the real information investigation process as much as possible. We also developed evaluation metrics that leverage the users' involvement, including those that use the utility of collected information as the basis for examining the performance.

The framework developed to evaluate the first version of *CAFÉ* turned out to be very useful and, in the longer term, more important than the study it was developed for. We have already used the framework to evaluate several tools for task-based information exploration, saving us a great deal of effort and helping to obtain interesting results. The goal of this paper is to present this framework to the community along with a meaningful example of its use, our evaluation of *CAFÉ*. Since this evaluation produced interesting results, presenting these results becomes the secondary goal of the paper.

In the remainder of this paper, Section 2 reviews previous research on frameworks developed in the fields of information retrieval, filtering and user modeling for evaluating the performance of the related systems and technologies. In Section 3, we concentrate on the presentation of the evaluation framework we developed for testing adaptive filtering engines in the context of realistic and task-based information exploration. In Sections 4–6, we will use our evaluation effort on the *CAFÉ* engine as an example to demonstrate the characteristics of the framework. We will conclude in Section 7 with some discussion of the further development of the framework.

2. Related work

Evaluation has always been an important aspect in developing information systems such as IR systems. The most influential evaluation framework has been based on Cranfield model, even though the model was published about four decades ago (Cleverdon, Mills, & Keen, 1966). The Cranfield model takes the system-oriented view of evaluation, and concentrates on examining the measurements of system performance, such as the precision and the recall of retrieved results. It also relies on a reference collection that consists of a set of documents, search topics/queries, and the corresponding relevance assessments (ground truth) between the documents and the topics. Because of these features, Cranfield-based evaluation frameworks can systematically compare different retrieval algorithms, and the evaluation resources can be reused multiple times. Current widely-used Cranfield-based evaluation frameworks include Text REtrieval Conference (TREC), which mainly concentrates on issues related to English information retrieval (Voorhees & Harman, 2005), Cross-Language Evaluation Forum (CLEF), which focuses on multilingual information retrieval (MLIR) among European languages (Peters et al., 2006), NII Test Collection for IR Systems (NTCIR) on MLIR among Asian languages (Kando, 2005), and INitiative for the Evaluation of XML Retrieval (INEX) on content-oriented XML retrieval (Fuhr, Govert, Kazai, & Lalmas, 2002).

However, Cranfield-based evaluation frameworks have many limitations (see discussions in (Borlund, 2003; Borlund & Ingwersen, 1997; Saracevic, 1995)). For example, the information needs in those frameworks are imposed, simple, and do not evolve as true information needs often do. The relevance assessments are based on static topical relevance, which do not change with different user characteristics and retrieval contexts. The retrieval systems are often assumed to be in batch mode for handling requests, where interactions between users and retrieval systems do not exist. Therefore, although this system-driven IR evaluation approach has greatly improved the effectiveness of retrieval systems and their matching algorithms, further development of IR evaluation needs alternative evaluation frameworks that can remove all or some of the limitations.

The evaluation frameworks based on user-oriented approach has been presented as such alternatives to the Cranfield model. Robertson and Hancock-Beaulieu (1992) present the improvements of the user-oriented evaluation approach over the system-oriented approach as three revolutions. The *cognitive* revolution means that users' cognitive activities should be considered; *relevance* revolution means that users' needs rather than their requests should be the criteria for judging the relevance; and *interactive* revolution means that IR mechanisms (a better term, they think, over systems) should be examined in light of the whole interactive process, rather than a static step of inputting queries and generating rank lists. Borlund (2003) points out that in the user-oriented evaluation approach, the user is the focus; users' information seeking and retrieval processes should be treated as a whole; users' information needs may change over time; and their relevance judgments are subjective and situational. Therefore, the evaluation should focus on how well the users, the retrieval mechanism, and the data collection interact for extracting useful information. Many interesting research results have been achieved via this approach (readers should consult (Ingwersen & Kalervo, 2005) for the latest development in this area). Interestingly, many studies utilizing the Cranfield-based evaluation frameworks actually work on the interactive side of IR, where users are the center of focus – not just the IR systems. For example, TREC had the interactive track from TRECs 3–11, then HARD track from TRECs 12–13. CLEF also has an interactive CLEF track since its beginning. However, it is widely accepted that interactive IR experiments are difficult to design, expensive to conduct, limited in their small scales, and hard to compare cross-site (He & Demner-Fushman, 2003). As Dumais and Belkin (2005) point out, the reasons for this are because the performance of interactive retrieval is greatly influenced by the searches as well as the topics and the systems, and these influences are often complex. Therefore, they state that the key is to reduce variability and separate the effects of searchers, topics and systems.

Both user-oriented and system-oriented evaluation approaches aim at the same goal: the reliability of the IR test performance results. Therefore, it is possible to combine the elements of the two approaches so that the evaluation framework is as close as possible to the actual information retrieval process, and at the same time has a relatively controlled evaluation environment as provided in the Cranfield model (Borlund, 2003). Aiming towards developing an evaluation framework for interactive IR systems, Borlund proposes an evaluation model that contains (1) a commitment of involving potential users, their dynamic information needs, and their relevance assessments that are multidimensional and dynamic; (2) a set of simulated task situations; and

(3) alternative performance measures that are capable of handling non-binary relevance assessments. The involvement of users and their relevance assessments ensure that the IR systems are evaluated under the conditions that the systems are useful and meet users' needs for their given situation. The simulated tasks make the evaluation close to the real operating environment, but also provide some flexibility on how the evaluation can be conducted, in the event the real environment is too complex to be modeled or controlled. The alternative measures then avoid the rigid and unrealistic assumption that the relevance assessments have to be binary. The ideas proposed in Borlund's work are quickly adopted in many areas, including the interactive track of INEX (Larsen, Malik, & Tombros, 2005), the study of implicit feedback (White, Jose, & Ruthven, 2004), and the polyrepresentation principle of IR (Larsen, Ingwersen, & Kekalainen, 2006).

At the evaluation framework level, our proposed framework is not restricted within the Cranfield model, because – agreeing with many others – we think that the system-oriented view of the evaluation framework has the limitations discussed above. These limitations are serious, since our goal is to develop truly useful IR systems for supporting task-based information exploration. In terms of methodology, our framework has no fundamental difference to Borlund's ideas. We share the same idea that user-oriented and system-oriented evaluation approaches can be combined to gain the advantages of both. The users should be included in the evaluation process, and the effectiveness of supporting the users' work is the focus of the evaluation. The interactions between users and the systems are important, and the tasks used in the evaluation should be realistic and close to the actual tasks performed by the users in their work. Finally, the current system-oriented measures are limited in telling us how truly useful the systems are.

However, our work does have differences to Borlund's. Instead of working on a framework for generic interactive IR processes, ours concentrates on testing specific information processing systems under task-based information exploration, where intelligence analysts are the main potential users. We like the idea of simulated task scenarios proposed by Borlund, but the tasks we assumed in our evaluation framework are tightly connected with the actual work of analysts. Our attention on alternative measures is focused more on utility-oriented rather than non-binary relevance. By using a simulated task outcome, we force users to develop a balanced measure between topical relevance and content novelty so that the overall utility can be expressed and measured. Finally, in our design of the evaluation framework, we also developed a reference test collection, an idea borrowed from the Cranfield model. Based on the test collection, the ideas about simulated tasks scenarios are instantiated into concrete tasks. We will discuss our evaluation framework in detail in Section 3.

There have been two previous major reference test collections for evaluating adaptive filtering and its related techniques. Between 1999 to 2002, Text REtrieval Conference (TREC) organized filtering and routing tracks (Robertson & Soboroff, 2002). Their approach concentrated on batch mode evaluation methods and used a set of filtering topics based on the Reuters news collection. To avoid human involvement while achieving the goal of evaluating adaptive filtering techniques across different sites (which means they are in the Cranfield model), they simulated the users' feedback that is necessary for adaptive filtering by assuming all relevant documents recommended by the system as positive feedback from the users. They concentrated on recommending documents rather than useful passages.

The second reference test collection was developed for supporting Topic Detection and Tracking (TDT) research (Allan, 2002). TDT supports five different tasks, in which a tracking task closely resembles adaptive filtering. Each TDT topic is explicitly linked to a seminal event, and the tracking task is to find documents related to the given topic. Similar to the TREC filtering track framework, there is no human involvement in the tracking process, and all feedback is simulated automatically.

As with other evaluation work based on the Cranfield model, both filtering and tracking tracks in TREC and TDT aim at testing algorithms, using simple topics, and simulating users' interactions rather than involving actual users. Their advantages, however, include simple design, easy execution, and good cross-system comparison.

The state of the art in evaluating AF engines in the field of user modeling is somewhat similar. While it is customary in this field to involve human users in the process evaluation, existing approaches use simple scenarios, small document collections, and formal evaluation metrics that do not take into account a number of factors that are critical in the information exploration context (Díaz & Gervás, 2005; Waern, 2004). Typically, the users are required to rate every document in a set of documents that could be incrementally presented to the users over several sessions. The ratings are used in two ways to evaluate the performance of an AF engine

post-factum. On one hand, the ratings are fed to the AF engine to produce a ranked list of recommended documents. On the other hand, these ratings are considered as ground truth, and are compared with the ratings produced by the engine to evaluate its performance using classic relevance-precision metrics.

The reference test collection in our evaluation framework has several key differences to previous collections. First, since our framework aims at supporting task-based information exploration, the tasks we have developed resemble intelligence analysts' actual work scenarios, which are complex and evolving with the events. Second, the evaluated filtering systems under our framework return useful passages rather than documents. Third, our evaluation framework pays attention to the involvement of human subjects in the process rather than using simulation. Finally, our framework considers the utility of the selected passages rather than simple topical relevance. We want to make sure that the usefulness of the system can be revealed through our evaluation framework. We will discuss the detail of the reference test collection in Section 4.

3. An evaluation framework for task-based information exploration

3.1. Reasons for developing the framework

The goal of our research is to evaluate information systems in the context of task-based information exploration performed by intelligence analysts. Although some ideas and resources can be borrowed from existing evaluation frameworks, we still have to develop a new evaluation framework for our evaluation tasks. The key reasons are mainly from the tasks that the information systems are supposed to support, which have not been the foci of previous evaluation frameworks:

- *The tasks that the information systems participate in are realistic.* In our studies, after given an RFI, an analyst starts to explore large volumes of data from various sources with the help of information systems and generates a short summary of information collected, typically a two-page point report. This is the real information exploration process, and the outcome is real too.
- *The tasks modeled in the framework are complex.* Each task is related to a seminal event. The information requests specified in an RFI are focused on that event itself and its different aspects. The events can have multiple aspects. The assigned tasks usually include specific subtasks, all of which are connected to the overall task.
- *These tasks are dynamic.* Events naturally evolve over time. An intelligence analyst then may have to track the development of the events over several sessions before producing the final point paper.
- *Information collected is at passage level.* Although documents would still play important roles in task-based information exploration – because the outcome is a point paper – the useful information is typically contained in specific text passages (snippets). Because of their high density of useful information and small size, these passages are perfect for investigation or report writing.

Because of these task characteristics, our evaluation framework should have methodology, measures and other means to reveal how systems perform in information exploration process, including multiple aspects of the exploration process (such as topical relevance, novelty, and final utility) and multiple stages of exploration process (such as information foraging and sense-making). We think that current evaluation frameworks are especially weak at measures and reference test collections for our tasks. This is why we decided to develop this framework.

The remainder of this section will present the methodology adopted in the framework for obtaining the information, the general metrics for measuring the systems. In Section 4 we will discuss a specific reference test collection we used for the study presented in this paper, including 18 specific tasks and the ground truth assessments between the tasks and the passages in the collection.

3.2. Methodology

Although (White, Muresan, & Marchionini, 2006) points out that the ideal evaluation approach is longitudinal and in a naturalistic setting, there is a need for a cheaper and quicker evaluation approach for the peo-

ple who develop information systems. Our evaluation framework takes the latter approach, combining controlled lab experiments with questionnaires and interviews.

However, our evaluation approach does involve human users conducting task-based exploration and interacting with the information system. Sharing with (White, Muresan, et al., 2006), we believe that information systems that support task-based exploration are highly interactive, where human subjects' interaction behaviors and interaction processes are important aspects to be observed and evaluated.

In addition, our evaluation assumes that intelligence analysts' seeking behavior in information exploration contains two major stages (or loops): information foraging and sense-making (Pirolli & Card, 2005). During the foraging stage, the analysts use their domain and search knowledge to collect potentially useful information in various media and sources in the information foraging stage. In the sense-making stage, analysts process, organize, and distill the collected information into a coherent form so that it can be integrated into their state of knowledge. We believe that the evaluation of the information systems should consider their roles and effects in both stages.

Our evaluation focuses on the quality of selected passages, which is examined not only by how much the contents of the passages match to the requests in RFI, but also whether and how they are used by the analysts in the final reports. Therefore, topical relevance is viewed as an important factor in relevance assessments. However, the most critical assessment criteria are based on the utility of the system in the task-based exploration (i.e., its ability to support analysts' work.) Therefore, a *utility* measure for passages is needed.

Our evaluation settings also try to simulate certain important aspects of analysts' work scenarios. For example, subjects are required to work on a task in multiple sessions, so that subjects encounter the same task-related issues as analysts, such as event evolution and duplicated information.

3.3. Metrics

Metrics provide a set of examination points for studies based on our framework. We want measures that can support our multiple layers of analyzing information systems. These layers include

- Measures focused on system performance. Inherited from existing frameworks, there could be measures on the accuracy and coverage of identified information (i.e., precision and recall) and new ones on the utility that are related to the point paper for a given task. All these measures will concentrate on passage level examination rather than document.
- Measures focused on user formal performance. Our idea here is similar to that of (Marchionini & Shneiderman, 1988), where we examine the paths and decisions taken by the users during the process of completing the tasks, we attempt to make inference regarding the subjects' cognitive activities, which would share some light on how well they are supported by the systems. The measures are expanded to evaluate separately information foraging and sense-making stages.
- Measures focused on user activity patterns obtained by mining the user log.
- Measures focused on data collected from subjective evaluation.

Among the measures above, the utility measure deserves special consideration. It is true that the utility measure is related to information systems' support in task-based exploration, which can be drawn from aspects related to the task outcome – the point paper. However, this approach is less reliable because the quality of the reports is related not only to the support from the information systems, but also to subjects' varying abilities in producing a coherent report based on the collected data. To remove the influence of the latter effect, we suggest to use not a point paper, but a set of collected (i.e., foraging) and organized (i.e., sense-making) passages as the product to be evaluated. We assume that these passages are the basis for writing the point paper; therefore the quality of passage selection and organization can be used to reflect the system features.

We impose a word limit on the selected passages for the final product. This has two advantages. First, it makes the final product more resemble the point paper, which is about one to two pages long. Secondly, the word limit brings in a cost function that is needed to represent our utility measure. Every word added into the final selection means that some other words could not be included. There is, therefore, a trade-off between including as much relevant information to a specific question as possible and covering all the required

questions within the specified word limit. Naturally, such decisions involve the concept of topical relevance (whether the passage contains on-topic information), novelty (whether the passage contains substantially new information not covered by previously selected passages), and usefulness (how useful the passage to the final selection, i.e., to improve the quality of the point paper). Here, usefulness is our utility, as we think that it is related to the topical relevance and novelty of the passages.

4. A reference test collection

4.1. Document set

It takes a great amount of resources, time and effort to develop a reference test collection from scratch. Therefore, our strategy is to reuse an existing collection provided the existing collection contains a large number of documents, saving much time and effort.

Two existing document collections have the potential to be used as the basis for our test collection. Reuters RCV1 collection, which is used in TREC filtering track, contains 800,000 news stories, covering a time period of 12 month between 1996 and 1997. The TDT4 corpus, developed for TDT evaluation, contains news articles from multiple sources in Arabic, Chinese and English languages, and it covers events happened between October 2000 to January 2001. It contains 28390 English documents.

Both collections have their corresponding topics. For example, over the years, several hundreds of filtering topics have been developed for the RCV1 collection. The same thing is true for TDT4 collection. However, because each TDT topic corresponds to a seminal event (Allan, 2002), which share the same characteristics of the topics in our framework, we decide to choose the TDT4 corpus as the document collection.

4.2. Simulated task scenarios

To simulate the task scenarios used by intelligence analysts, we expanded TDT4 topics according to the task scenario development guidelines discussed in Section 3.1.

During the development of tasks, we required the developers (i.e., the authors) to search the TDT4 collection to become familiar with the scenarios and their relevant documents. We also used a two-point access strategy to help us generate sub-task questions that are related to the evolution of the events and tasks. This strategy requires us to identify two time periods, one at the beginning of the event and the other approximately 2 or 3 days later.

Some brief background information, including a seed story (i.e., a good relevant document to the scenario) is provided for each scenario to make sure that subjects would have roughly the same level of knowledge about our scenarios. In reality, a superior officer might hand out a sample story as part of the RFI. Fig. 1 shows a sample task scenario that we developed.

4.3. Ground truth assessments

Human annotators were recruited to mark up the “ground truth” for each developed task scenario. Our assessment of ground truth has the follow features:

- the annotations were at the passage level;
- the annotations were independently collected for three aspects: topical relevance, novelty and utility;
- to accommodate the fact that there are different degrees of relevance, novelty and utility, we collected annotator judgments at three levels: highly relevant/novel/useful, slightly relevant/novel/useful, and not relevant/novel/useful;
- at least two annotators marked each scenario along each aspect (relevant/novel/useful).

We did not have the manpower to process the entire TDT4 collection for the ground truth annotation. By taking advantage of the fact that both our scenarios and our ground truth were in fact elaborations of the original TDT4 topics, we used the set of relevant documents associated with the original TDT4 topics as

G40047: The Deadly Texas Seven

- **Background Information:**
 - Kenedy, TX is a small town near San Antonio, Texas.
 - San Antonio is a big city in Texas.
 - Guns are more common in Texas than in other US states.
 - Cars/trucks are the most common means of transportation within Texas.
- **Short description of the task:**
 - There were seven inmates escaped from a prison in Texas, Your task is to find information needed to coordinate their recapture.
- **From the documents, find snippets of text that contain answers to each of the following questions:**
 1. Where and when were they last sighted (location, time/date)?
 2. Who are their known contacts on the outside , and what is their relationship?
 3. How many convicts are still at large?
- **From the documents, find sentences that can be used to write short summaries about**
 4. What vehicle might they be driving ?
 5. What steps have been taken so far by the police, such as reward (how much?), posting, etc ?

Fig. 1. A task scenario example.

the document pool for ground truth annotation. Later, the adaptive filtering engine we were testing was able to discover small amount of additional relevant documents.¹ These additional documents were subsequently annotated too.

In total, we developed 18 task scenarios. 1916 documents were examined with respect to their relevance, novelty and utility. The relevance annotation produced, on average, 644.4 highly relevant passages, and 230.5 slightly relevant passages per topic. The novelty annotation produced on average 82.4 highly novel passages, and 118.3 slightly novel passages per topic. Finally, the utility annotation produced on average 130.4 highly useful passages and 122.8 slightly useful passages per topic before the final selection, and 150.8 selected useful passages after the final selection. The annotation files are independent from the source data (TDT4 collection) and can be used by anyone interested in running similar studies.²

5. Evaluating CAFÉ: A sample study

To illustrate how the evaluation methodology associated with our framework, this section describe a sample application of the evaluation framework, where we attempted to assess how well an adaptive filtering engine called CAFÉ can help analysts in their task-driven exploratory search. The methodology adopted in the study utilized a controlled lab experiment involving human subjects, and made comparisons between two information access systems: the CAFÉ and a state of the art information retrieval system. The following subsection discusses the four components of the study: data, experiment systems and procedures, measures and subjects.

5.1. Data selection and preparation

To simulate a realistic context where analysts explore the information space over an extended period of time, we divided the collection of documents for each topic into 3 or 4 subsets using time thresholds. This segmentation simulates the unfolding of a seminal event (and an analyst's tracking of the event) over time. Subjects were asked to perform their tasks in multiple sessions over a period of one to two weeks. Each new session added a new segment of data to the pool of documents accessible by subjects.

¹ Comparing to the original TDT4 relevant document set, the latter add-on discovered by CAFE was relative small: 61 new documents, or approximately 6% of the annotated set. Therefore, it seems that the original TDT4 relevant document set has pretty reasonable coverage of useful information.

² The tasks and their ground truth annotations can be downloaded from crystal.exp.sis.pitt.edu:8080/gale/GALE-resources.html.

The subjects' interaction with the system during each session was logged for future analysis and also passed to CAFÉ as a flow of positive and negative feedback. CAFÉ used this feedback to model the user task and to produce a list of passages ranked by their relevance to this task. The baseline system simply executed the same queries over the shifted document set without any adaptation.

We selected eight of the 18 task scenarios as the test topics. The selected scenarios all have a large number of stories distributed over a reasonable period of time. Each topic is divided into segments along its timeline so that a comparable number of relevant articles (topic information density) is maintained within each segment (see Fig. 2 for an example). It simulates the real life situation where the frequency of re-assessing information about an event is related to the speed of its unfolding. In addition, it ensures that there is reasonable number of relevant articles within each session. Four topics with larger number of articles were divided into 4 segments and the remaining four topics into 3 segments (Table 1).

5.2. Subjects

Recruiting real intelligence analysts as our subjects was infeasible for this study. We instead recruited graduate students in library and information science whose knowledge and experience in information access closely fit the profile of intelligence analysts. Subjects were required to be native English speakers and have completed at least one information retrieval course. Familiarity with the news content was not required, because analysts are often asked to research topics outside their domain expertise.

Eight subjects participated in the study from July 12th to August 1st, 2006. Four subjects were assigned to the 3-session topic group, and four to the 4-session topic group. All four subjects in each group completed the

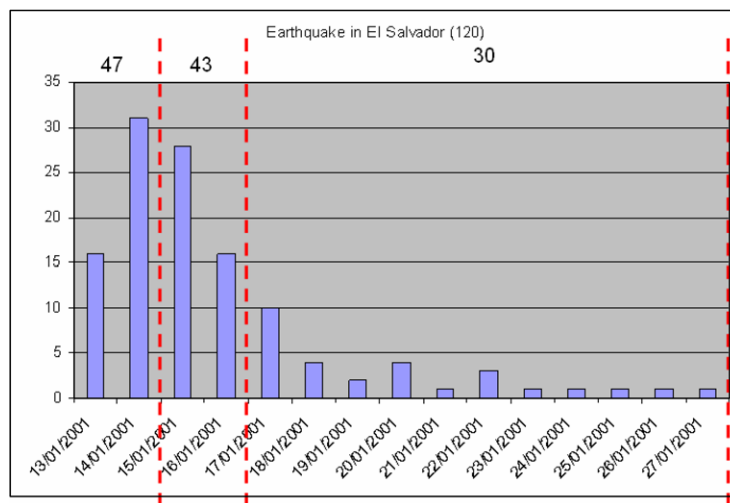


Fig. 2. An example of topic segmentation.

Table 1
The eight selected topics

TDT4 topic	Title	# of Sessions
40004	Russian Submarine Kursk Sank	4
40021	Earthquake in El Salvador	3
40055	Edmond Pope Convicted of Espionage in Russia	4
41005	UN Strengthens Sanctions against Kabul	3
41011	Turkish Prison Riots	3
41012	Trouble in the Ivory Coast	4
41024	Congolese President Laurent Kabila Feared Dead	3
41025	End of the Line for Peruvian President Alberto Fujimori	4

experiment simultaneously with a session interval of 2–4 days. Subjects were required to attend all of their group's sessions.

Six of the subjects were students in the Master of Library and Information Science (MLIS) program at the University of Pittsburgh; one subject was from the Master of Science in Information Science (MSIS) program at the University of Pittsburgh; and one subject was from the Computer Science program at Carnegie Mellon University. Seven of the eight subjects were female and the age range of all subjects was 22–65. On a ten-point scale (10 being the highest), the subjects mean rating of their search abilities was 8.375 with a mode of 8. In terms of time spent reading or viewing news each day, five subjects said they spent less than one hour, and the remaining three spent 1–2 hours.

5.3. Experimental and baseline systems

CAFÉ is an adaptive information filtering system developed at Carnegie Mellon University for utility-based information distillation. It combines the strengths of a state-of-the-art adaptive filtering system (Yang, Yoo, Zhang, & Kisiel, 2005) and a top-performing novelty detection system in benchmark evaluations for Topic Detection and Tracking systems (Fiscus & Wheatley, 2004). Furthermore, to support user interaction with the system in task-based information exploration, CAFÉ provides chronological segmentation of the input stream of documents, passage ranking per query based on both relevance and novelty, and the utilization of task profiles, query logs and recorded user interactions with system-selected passages.

CAFÉ takes the rich information in the task description to construct a profile for the task as the initial setting for adaptive filtering. The task profile is incrementally updated (“adapted”) as soon as new user feedback is received: the feedback indicates the relevance, redundancy or both of the currently processed passages. The user may also add new queries or modify the existing queries as a part of the feedback. The adapted profile is further used to re-rank passages with respect to the current query. The passages already seen by the user are removed from the re-ranked list to avoid repetitive information for user to review.

CAFÉ uses a regularized regression algorithm for the training of task profiles. It estimates the posterior probability of a task given a passage using a sigmoid function

$$P(y = 1|\vec{x}, \vec{w}) = 1/(1 + e^{-\vec{w}\cdot\vec{x}})$$

where \vec{x} is the vector representation of the passage whose elements are term weights, \vec{w} is the vector of regression coefficients, and $y \in \{+1, -1\}$ is the output variable corresponding to “yes” or “no” for the relevance with respect to a particular task. Regularized logistic regression has been found as one of the most successful algorithms in benchmark evaluations of adaptive filtering (Yang et al., 2005; Zhang, 2004). Technical details about the CMU's LR classifier can be found in (Yang et al., 2005).

The baseline system is a non-adaptive passage retrieval engine developed by the first author with the help of a group of researchers at University of Maryland (He & Demner-Fushman, 2003). It uses Indri 2.0³ as the underlying document retrieval engine. The effectiveness of the baseline engine has been demonstrated in TREC HARD 2003 (He & Demner-Fushman, 2003).

5.4. Experimental procedure

To minimize the potential impact of inter-subject difference on the study results, we adopted a within-subject design. During each session, each subject had to work on two tasks with one system, then worked with the other system for another two tasks. We used Latin Square method to rotate the sequence of system and task combinations to remove learning and fatigue effects (Table 2). However, for a given topic, the same system was used to complete the tasks throughout all sessions.

Subjects were given printed instructions on how to use the system and an entry questionnaire to assess their technical ability, search experience, and familiarity with news. A brief (30-min) training session was conducted on the user interface and experiment tasks, including a 10-min practice on a training topic.

³ <http://www.lemurproject.org/indri/>.

Table 2

An example of experimental session structure (Session 1, 3-session topic group) showing points in the experiment when questionnaires were administered and a Latin square rotation of topics and system sequences

Subject				
1	2	3	4	
Entry questionnaire (session 1 only)				
40021-CAFE	40021-Baseline	41024-CAFE		41024-Baseline
Post-search questionnaire				
41005-CAFE	41005-Baseline	41011-CAFE		41011-Baseline
Post-search questionnaire				
Post-system questionnaire				
41011-Baseline	41011-CAFE	41005-Baseline		41005-CAFE
Post-search questionnaire				
41024-Baseline	41024-CAFE	40021-Baseline		40021-CAFE
Post-search questionnaire				
Post-system questionnaire				
Exit interview				

For the purpose of the experiment, we developed a simple interface, which was used for both CAFÉ and the baseline engine. By building this interface, we attempted to achieve two specific goals of this study: to separate the contribution of ranking provided by the adaptive filtering engine or by the baseline system from other factors that may influence user performance and, at the same time, to simulate the exploratory task-driven work of human analysts as close as possible. To satisfy the first goal, we decided to take two simplifications. First, ad-hoc search function was not provided in the interface. The users are known to differ greatly in their query formulation skills, and we wanted to avoid the influence of this factor to the user's performance. Second, in the sense-making part of the interface (Fig. 4), subjects were not asked to write final point paper, instead their tasks were to select or organize the passages that would be used for writing the final report. Through this simplification, we avoid the effect of different report writing skills, which is not the focus of our study. These simplifications may not be necessary when performing a similar study with professional intelligence analysts, but we consider them important for the kind of subjects we used.

The interface simulates the analyst's activities of collecting and organizing potentially useful passages related to a given event, and the final sets of organized passages were used as the surrogate of the point paper that summarizes the collected information.

The interface supports both the foraging and sense-making stages of information access. The foraging interface (see Fig. 3) consists of two frames. The left frame shows the list of the passages ordered by their relevance to the perceived user task. The passages are generated either by CAFÉ or by the baseline engine. The right frame shows a container called the *shoebbox*, which is a traditional information-processing tool used by analysts. The shoebox stores all useful text selected by the user for his/her final report. The user can copy directly any part of the passages to the shoebox, or open a pop up window to view the complete document and select from there. Darker color boxes to the left of the text fragments in the shoebox indicate that the fragments were selected from the passage list directly, and lighter color boxes indicate that the fragments were selected from the full text window. A text fragment can be removed from the shoebox, or can be ordered by the posting date or by the sequence of user's selection.

The selection of a text fragment provides some confidence that its content is relevant to the user task and is considered as a positive feedback by CAFÉ. The user can also provide a negative feedback by removing irrelevant passages from the list. To make this action reversible, the headlines of all removed passages are retained in the passage list. Positive and negative feedback are used by CAFÉ to generate the list of passages for the next session.

The sense-making interface helps the user to organize collected text fragments for the inclusion into the future report. In the work of real analysts, the organized set of passages is used as a source to prepare the final report or some other product. The interface (see Fig. 4) allows the user to remove unwanted text fragments and organize the remainders by associating each with one of the questions of the task scenario. The questions are presented at the right frame for quick reference.

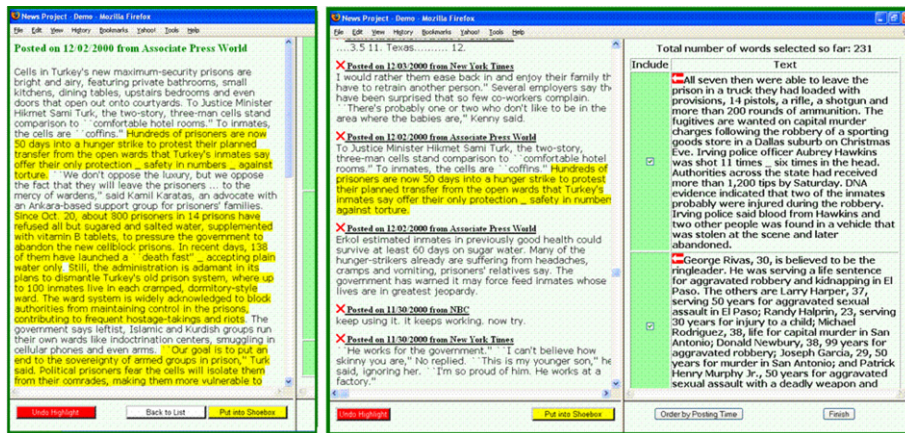


Fig. 3. Information foraging: assembling text fragments in the shoebox.

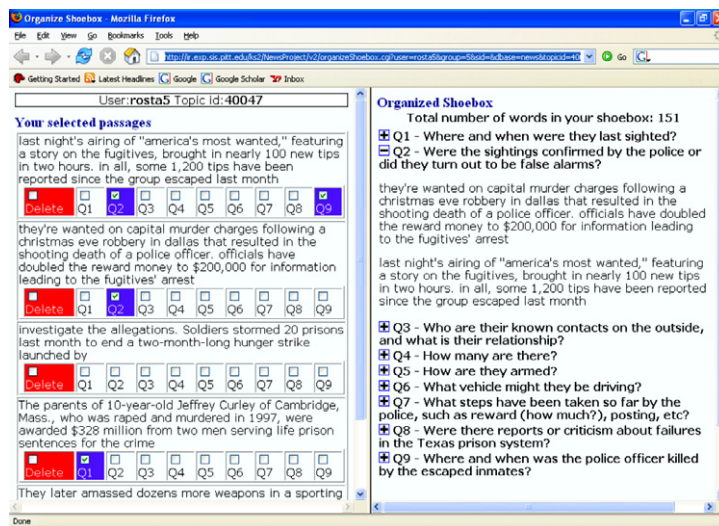


Fig. 4. Sense-making: final selecting and organizing in the shoebox.

For each topic, subjects were given 20 min per session to examine, highlight and add text fragments to the shoebox. Subjects could add a maximum limit of 1000 words to their shoebox during each session. After the 20-min search session, subjects were given 5 min to edit the recently-added contents of their shoebox for usefulness and to meet the 1000-word limit. Each search session lasted approximately 2 hours, including breaks and administering questionnaires.

At the end of the final session for each topic, subjects were asked to compile their final “report” on that topic. In lieu of an actual written report, subjects were presented with all of the contents of their shoeboxes and asked to produce a final shoebox of no more than 2000 words that best summarized the collected information for the topic. Additionally, subjects had to indicate which subtopic question(s) a final selected passage is related to. The final report generation lasted 20 min, followed by a 10-min exit interview.

6. Result analysis

6.1. System comparison based on output ranked lists

Our first result analysis concentrated on examining the two systems intrinsically, i.e., examining the two systems’ output – the rank lists of passages displayed to the subjects. Due to the nature of evaluating adaptive

systems, we paid more attention to precision than recall. The calculation of precision has been modified to handle the fact that the basic unit for recommendation is a passage (e.g., a text snippet) rather than a whole document. We adopted and expanded the precision calculation in HARD03 (Allan, 2003), where all the words in relevant passages in ground truth that are overlapped with the words in at least one retrieved passage will be marked. Each marked word also has a weight calculated based on how many ground truth annotators selected it into the ground truth. The passage precision then is the weighted sum of those marked words over the weighted sum of all the words in the returned passages where each word that is not in the ground truth has weight.

As shown in Fig. 5a, averaged across all topics, the rank lists generated by CAFÉ have better precision scores (0.74 versus 0.45) than that of the baseline when examining the top 20 passages of the ranked lists. This 65% relative improvement is statistically significant (paired t test with $p \leq 0.05$). The superiority of CAFÉ is also evident in the results of top 60 passages. Although the improvement is not as great (0.32 versus 0.27, a 19% relative improvement), the difference is still statistically significant. When we examine further down the ranked lists, the performance of CAFÉ's results becomes almost equal (both 0.19 at top 100 passages) and slightly inferior (0.16 versus 0.17 at top 120 passages) to the baseline. Although the differences are not significant in these two cases, the whole trend does indicate that CAFÉ did a better job of pushing high quality, useful passages to the top of the ranked lists, whereas the baseline performed better at the lower end of the ranked lists. The reason that our analysis stops at the top 120 passages is because CAFÉ only generated around 120 passages for most ranked lists in later sessions, even though we aimed to show the top 200 passages to subjects in our original design. Therefore, calculating precision at 200 made little sense.

6.2. System comparison based on usage profiles

When examining how system's performance extrinsically, there are several aspects to look at. Firstly, there are performance measures on subjects' selection of useful passages. Again we look at precision of the passages.

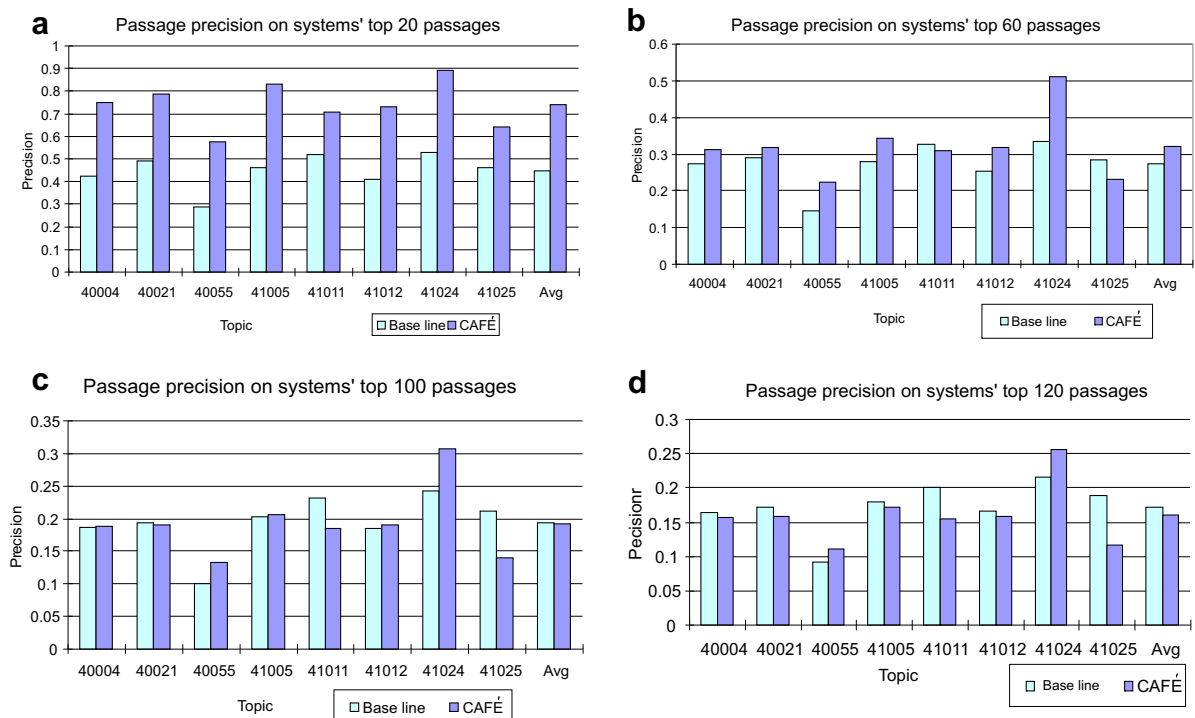


Fig. 5. Passage precision on the rank lists generated by the two systems: (a) based on top 20 passages; (b) based on top 60 passages; (c) based on top 100 passages; (d) based on top 120 passages.

The examination looks into several key points of the exploration process as indicators of the systems utility in supporting subjects' information exploration tasks. These key points include: (1) the three points in the information foraging stage. The three key points are after the first session, which tells us the subjects' ability to use the systems for finding answers when there is no adaptation difference between the two systems; after the final session, where subjects' selection should reflect any difference in the result sets received from the two systems; and the accumulated final selections of passages in the information foraging stage, which tells us the overall system's support in subjects' selection; and (2) The passages kept after the sense-making stage, where the selected passages were viewed as the final product to avoid the unnecessary complication from report writing.

Secondly, to compensate for the fact that differences in topic difficulty might affect the output regardless of system, we designed a *normalized precision* within subjects across all topics that they performed on, which is defined as (1), here $p_{i,j,k}$ means the precision of subject i in session j on topic k . $\bar{p}_{j,k \max}$ means the max average precision for session j in all topics. $\bar{p}_{j,k}$ means the average precision for session j and topic k .

$$np_{i,j,k} = p_{i,j,k} \frac{\bar{p}_{j,k \max}}{\bar{p}_{j,k}} \quad (1)$$

Thirdly, there are usability measures examining the support of the systems in subjects' selection process. This includes, for example, how many passages were selected in each sessions; how quick subjects were able to select useful passages; and how deep in the rank lists did subjects look for useful passages.

Fig. 6 shows the results of the passage precision on subjects' selection of useful passages. On average across topics, CAFÉ obtained better results than the baseline system in helping subjects to forage useful information after the first session (see Fig. 6a), after the final session (b), and at the final information foraging results (c). The differences are not statistically significant. However, when we look at the final results from sense-making stage (see Fig. 6d), the average precision scores of the passages selected using the two systems are almost identical (0.539 versus 0.537, only 0.3% relative improvement).

To remove the effects of topic differences, we utilized the normalized precision. Fig. 7 shows that CAFÉ consistently outperformed the baseline at the three key points of the information foraging stage, as well as the sense-making stage. The differences, however, are not statistically significant. We do see that different subjects performed quite differently despite our efforts to normalize the difficulty of tasks. Within-subject performance also changed dramatically in both the information foraging stage and the sense-making stage (for example, subject 6).

Fig. 8 compares the number of passage selections made on the two systems for each of the eight topics. These data provides some evidence that CAFÉ was better than the baseline in helping subjects discover relevant information. The subjects annotated more passages overall (734 versus 598) with CAFÉ, but this difference was not statistically significant (independent sample t -test, $p = 0.168$). When we considered the six topics where CAFÉ performed better, the difference was statistically significant (independent sample t -test, $p = 0.008$). Looking at the topics where subjects selected more passages with the baseline system, the difference between the two systems was not significant ($p = 0.225$).

How quickly can subjects locate useful passages can be seen as an indicator of the system's support in passage selection. For example, a system that supported more passage selection at the beginning of the session can be viewed as providing more efficient support. Therefore, we counted the number of passage selections at certain benchmarks: at 5, 10, and 15 min after subjects started their tasks (see Fig. 9). These data show that about half of the passage selections were made within the first 5 min in both systems, and there is slightly more passage selections in CAFÉ system at the earlier session stages (≤ 10 min) than that in the baseline.

By examining the ranks of selected passages, we hope to see how deep the subjects had to go in their search for relevant information. Both systems ordered the passages, and the subjects knew this. That is why subjects all started their selection from the top of the rank lists. Therefore, the deeper the subjects have to go, the more effort the subjects had to put into, the less supportive the system is. The descriptive statistics in Table 3 show that passage selections in CAFÉ were generally concentrated on higher ranks in the list. This may show that subjects tended to find useful information closer to the top in CAFÉ than that in the baseline.

One interesting statistic is the mode ranks for each system, which were 26 and 1 for CAFÉ and the baseline, respectively. Mode rank 26 is approximately the second or third page when scrolled down the list, whereas mode rank 1 means the very top of the list.

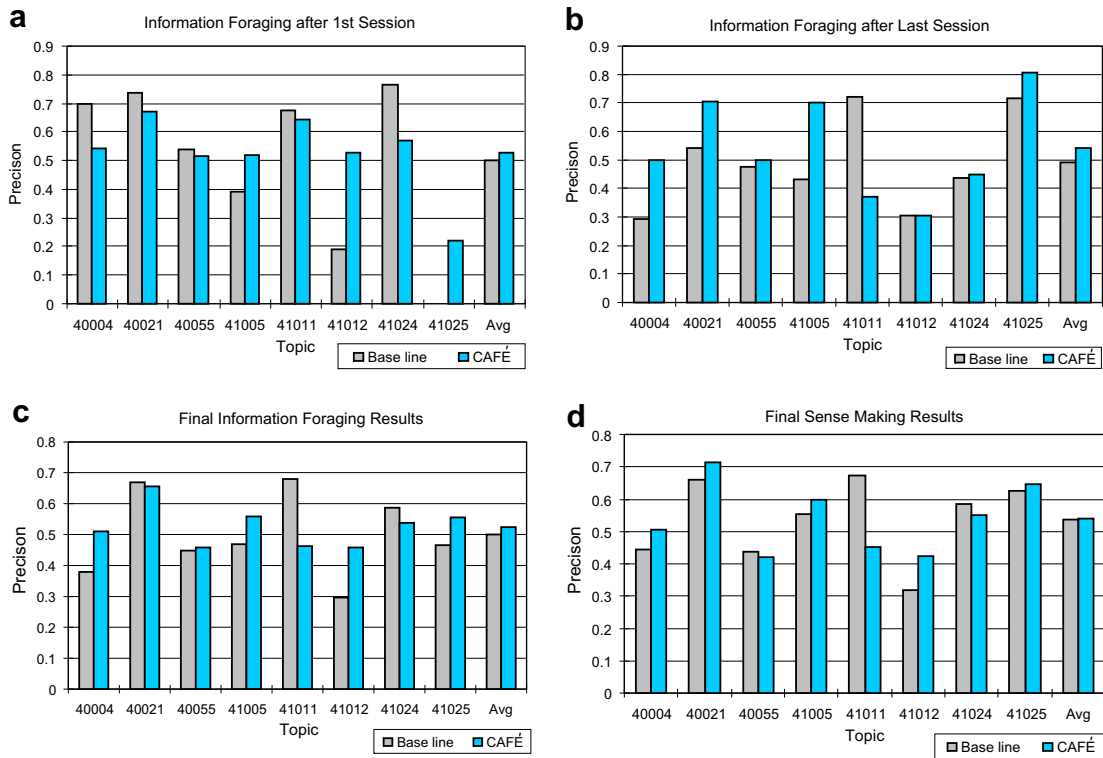


Fig. 6. Passage precision on the selected useful passages at key points of information foraging stage and sense-making stage. (a) Information foraging after the first session; (b) information foraging after the last session; (c) final information foraging results; (d) final sense-making results.

The topic level analysis shown in Fig. 10 is consistent with the overall statistics above. In five out of the eight topics, the average selection ranks in CAFÉ were higher than those of the baseline system. This difference is statistically significant (independent t -test, $p = 0.008$). The difference of the average ranks among the remaining topics (4004, 41,011, and 41,012) is not significant ($p = 0.114$).

The way that subjects selecting passages can also be an indicator of the systems' support. While some passages were selected directly from the ranked list of passages, others were selected after subjects examined the full content of the corresponding document. Fig. 11 shows the number of passages selected directly from the rank lists. Subjects who used CAFÉ made more direct selection of passages from the rank lists. This could indicate that the passages generated by CAFÉ give them more confidence in selecting the passages directly. However, the difference was not significant (paired t -test $p = 0.077$).

An interesting measure we developed was on how strong the information scent is. A snippet selected by the filtering system may not be sufficiently relevant to select for the shoebox, but may bear reasonable information scent (Pirolli & Fu, 2003) to cause the user to open the source document. If the document proves useful enough to select a snippet from it, the additional efforts of opening these documents has paid off. If the document is not sufficiently useful to make a selection, these efforts are wasted. Therefore, the portion of viewed but *not* selected from viewed documents is defined as the indicator of the wasted effort. From this viewpoint, a system that has a larger portion of documents selected from documents viewed by the users might be better.

In the analysis, the opened documents found in the log transaction are regarded as *viewed* documents, and documents from which at least one useful passage was selected and added to the shoebox are called *selected* documents. The proportion of the number of viewed documents that are not also the selected documents divided by the total number of viewed documents is defined as the wasted effort. As shown in Fig. 12a, the

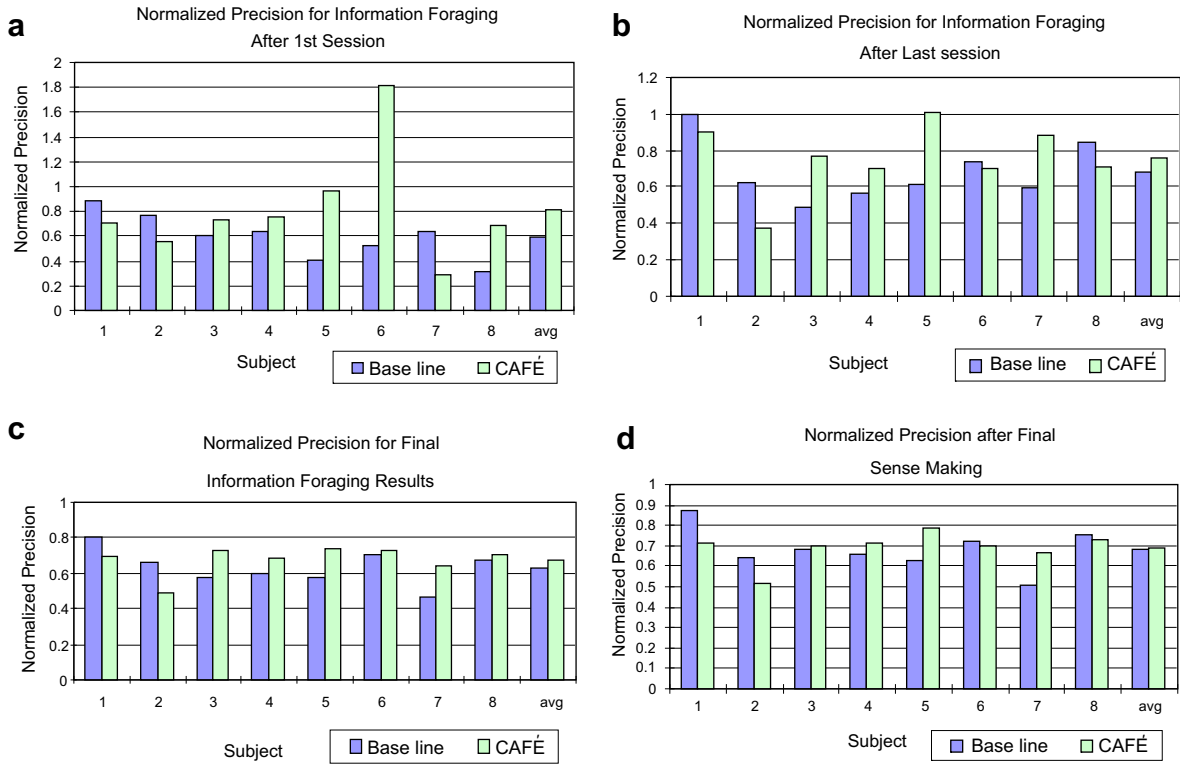


Fig. 7. Normalized precision of subjects' passage selection.

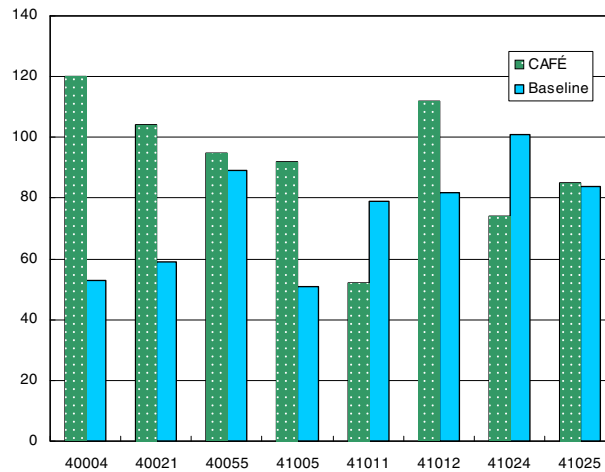


Fig. 8. Number of selected passages in the two systems by topic.

wasted effort observed in the baseline system is higher than that in CAFÉ for almost all subjects. The difference is statistically significant ($p = 0.034$).

The analysis of wasted efforts by session (see Fig. 12b) reveals that the wasted efforts in the baseline system ($M = 0.553$, $SD = 0.088$) was again higher than that in CAFÉ, and the difference is significant ($p = 0.006$). In this analysis, each session, except session 4, had 8 subjects and 8 topics; but session 4 only had 4 subjects and 4 topics.

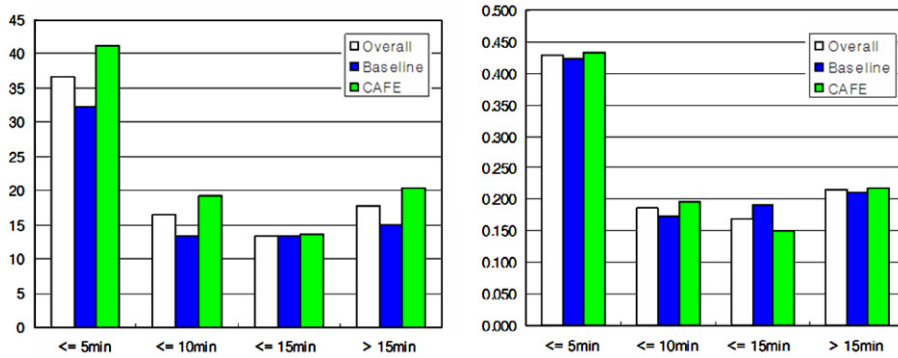


Fig. 9. Number (left) and percentage (right) of selected passages along the duration of a session.

Table 3
Descriptive statistics of the ranks of selected passages

System	Mean	Median	Mode	Standard deviation
Baseline	73.4	48	1	66.6
CAFÉ	50.6	39	26	43.7

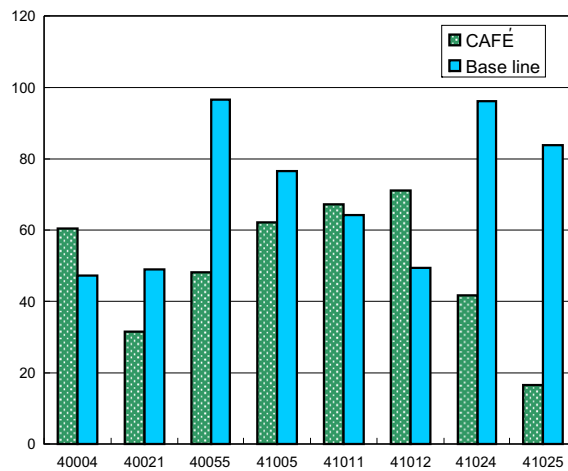


Fig. 10. Average rank of passage selections by topics.

6.3. System comparison based on user feedbacks

Following each search task, subjects were given a post-search questionnaire to assess their satisfaction with the system. Then before switching the systems, they were given a post-system questionnaire. For all questions, subjects were asked to rate their level of agreement from 1 (Extremely) to 5 (Not at all). For both systems, subjects were asked to rank topic familiarity, sufficiency of news, utility of passages, ability to find useful snippets, ease of use, and overall satisfaction. In sessions 2 through 4 for CAFÉ only, subjects were asked to rate their impression of how well the system used their negative feedback to generate subsequent passage lists.

A 2 × 4 within-subjects ANOVA was performed on the post-system questionnaire data to determine significant differences in user answers by system and session. Fig. 13 shows the mean post-system questionnaire responses averaged across all sessions. Though the post-system responses averaged across all sessions for CAFÉ tended to be higher than the baseline's, there were significant differences between the two systems

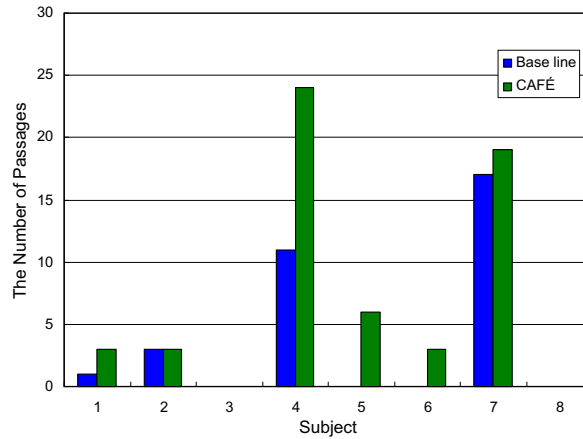


Fig. 11. The number of passages selected directly from the ranked list of snippets.

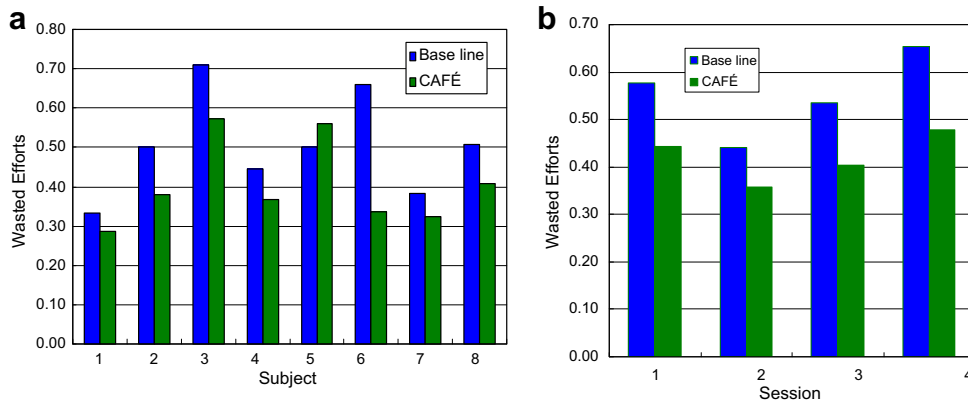


Fig. 12. Wasted efforts by (a) subjects and (b) by sessions.

for any of the questions. These results were reinforced by subjects’ responses in exit interviews: three indicated they preferred CAFÉ, two liked the baseline, and three had no preference. However, there were significant differences in mean post-system responses among sessions (Fig. 14) averaged across systems for perceived sufficiency of news ($F(2, 14) = 7.834, p = 0.005$), ease of use ($F(2, 14) = 6.921, p = 0.008$), and overall satisfaction ($F(2, 14) = 7.440, p = 0.006$).

Simple comparisons revealed significant increases in subjects’ ratings from session 1 to session 3 in both systems for perceived sufficiency of news (BASELINE: $F(1, 7) = 9.000, p = .020$; CAFÉ: $F(1, 7) = 14.538, p = .007$) and overall satisfaction (BASELINE: $F(1, 7) = 8.795, p = .021$; CAFÉ: $F(1, 7) = 7.000, p = .033$). There was also a significant difference in perceived ability to find useful snippets between sessions 1 and 3 for the baseline system only ($F(1, 7) = 5.727, p = .048$), as well as for ease of use for CAFÉ only ($F(1, 7) = 10.309, p = .015$). There were no significant differences in subjects’ ratings of CAFÉ’s use of their negative feedback, nor the utility of passages among sessions.

The data suggest that, over time, subjects felt more satisfied and had greater success completing their tasks with both systems. One factor that may explain this increased satisfaction is that subjects were given identical task descriptions throughout all sessions; thus, they were presented all sub-task questions from the outset of the experiment. Although the temporal nature of the data was explained to subjects in the training, subjects’ inability to find answers to certain sub-task questions in earlier sessions may have led to lower responses. Subjects also had more complaints about slow system response times in sessions 1 and 2 versus 3 and 4, which may have also suppressed earlier ratings.

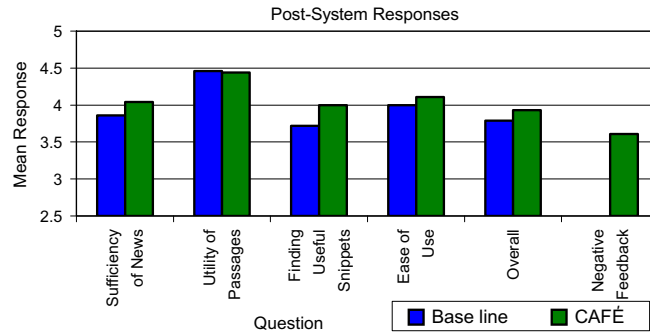


Fig. 13. Post-system questionnaire responses by system, averaged across all sessions.

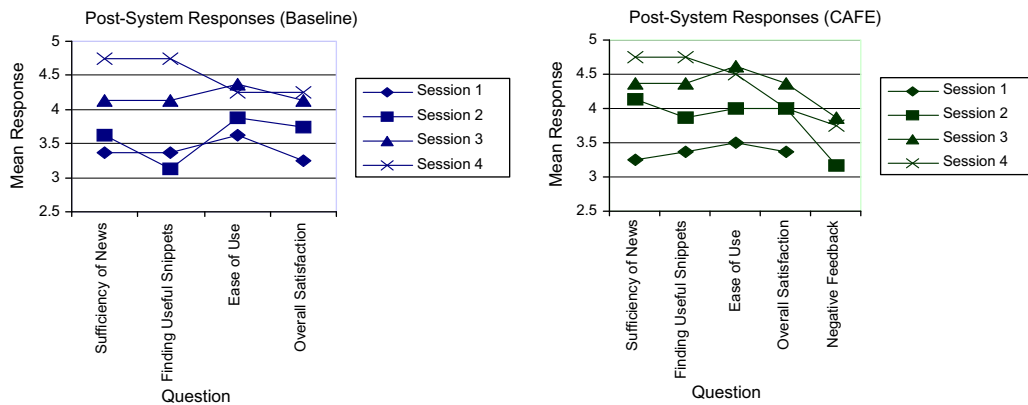


Fig. 14. Mean post-system questionnaire responses for baseline and CAFÉ, by session.

Further analysis of the post-search questionnaires uncovered some interesting patterns in users’ subjective feedback among the topics. In the 4-iteration group (see Fig. 15), topic 40055’s feedback on all questions tended to be higher versus other topics for both systems, while topic 41012’s tended to be in the bottom half among all topics. Topic 41012’s feedback correlates positively with subjects’ performance, which was lowest or

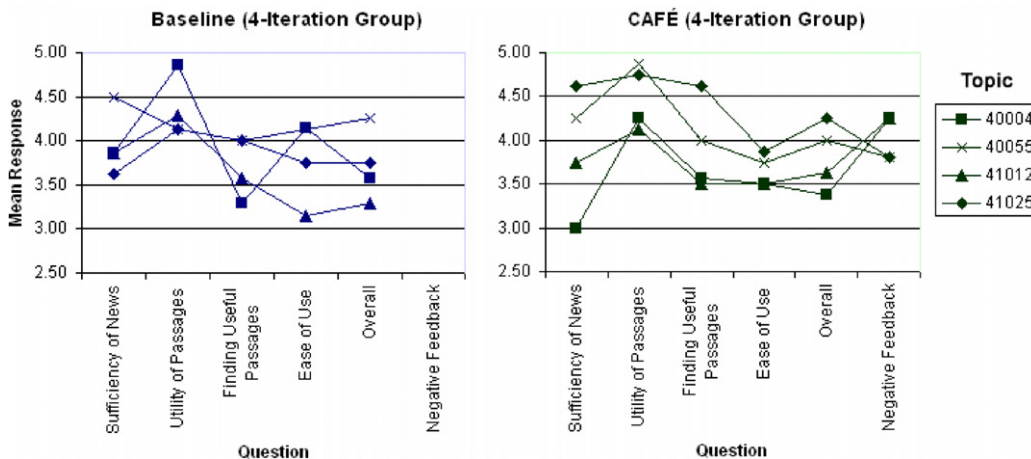


Fig. 15. Mean post-search questionnaire responses from subjects in the 4-iteration group, by topic.

among the lowest for all precision measures and percentage of subqueries answered correctly. Feedback from subjects who used CAFE for topic 40004 was also lower relative to other topics (particularly sufficiency of news), but the feedback from baseline users is less clear-cut. While users found the passages generated by the baseline system highly useful when deciding to open the full article, they ranked the overall utility of passages the lowest for topic 40004. Many of the subjects complained about the readability of the articles within this topic, particularly of those articles that were translated by machine from foreign languages, which may help reconcile the differences in ratings between these two questions. Percentage of subqueries answered correctly was also lower than average for topic 40004, while precision measures were average to below average versus all topics. We found similar results in the 3-iteration group.

7. General discussion and conclusion

In this paper, we presented an evaluation framework designed specifically for assessing and comparing performance of innovative information access tools created to support the work of intelligence analysts in the context of task-based information exploration. The motivation for the development of this evaluation framework came from the evaluation need of the project, which our team is focusing on. As we discovered, none of the existing evaluation frameworks can satisfy the evaluation needs defined by task-based information exploration context. We needed a framework that connects closely with the kind of tasks that intelligence analysts perform: complex, dynamic, multiple facets and multiple stages. We also needed a framework, which views the users rather than the information systems as the center of the evaluation, and examines how well the users can be served by the systems in their tasks. Finally, we needed a framework that could allow us to evaluate separately the user performance during each major stage of their work such as information foraging and sense-making.

In this paper we presented the main components of the developed evaluation framework – the methodology, the ideas of simulated tasks, and the set of metrics. We also presented a reference test collection and an evaluation procedure as the extra part of the framework. All components feature valuable innovations, which we already tested while evaluating several innovative information access systems.

Our reference test collection offers 18 tasks scenarios completed with corresponding passage-level ground truth annotations at three independent aspects: topical relevance, novelty, and utility. The three annotations provide multiple aspects examinations on the performance of the systems. The suggested evaluation approach supports assessing user performance in complex realistic tasks, and allows separate evaluation of the system impact on foraging and sense-making stages. Metrics recommended in the framework provide multiple layers of analyzing the information systems. These layers include measures focused on system performance, measures focused on user formal performance and informal indicators of the process, and those on subjective evaluation.

To demonstrate the usage of our evaluation framework and the reference test collection, we presented the framework itself in parallel with a specific evaluation study of CAFÉ, an adaptive filtering engine designed for supporting task-based information exploration. This study can be considered as a successful use case of the framework. The multiple layers of evaluation measures indeed revealed various aspects of the information systems.

One interesting design of our experiment is the usage of multiple sessions to capture the dynamic development of the event and the tasks. According to the results collected from the experiment, it seems that this feature indeed showed the complexity of the tasks, the changes of systems performance and the varieties of subjects' behaviors in different sessions. The separation of information foraging and sense-making in subjects' tasks, which is another interesting design in our study, also helped to review the different support between the two systems.

We hope that this work will contribute to the establishment of evaluation approaches for exploratory search systems. We also hope that the reusable framework we developed will be further utilized to explore various ideas on study design and evaluation parameters.

Beyond this general contribution, our work brought some interesting observation from the specific study of the CAFÉ engine. It seems that existing IR systems are polished to work well with existing evaluation frameworks – i.e., produce good formal relevance. Both CAFÉ and the Indri baselines are developed to push more relevant passages to the top. However, as our study hints, formal relevant is not equal to better user support.

We can't take for granted that users can perform better with formally better outputs. From the users' task-oriented view, CAFÉ system was able to win over the baseline on many aspects, including better rank lists, better adaptation, less wasted effort, subjects' comments and log analysis, however, it did almost identical when examining how well the passages were selected. Therefore, this gives us a ground to argue that evaluation involving human users is essential to obtain truly useful testing results that are meaningful. To make further progress with developing systems that can provide better support to their users we need frameworks and evaluation approaches that take users and the work context into account. We consider the framework that we developed and the evaluation approach that we suggested as useful steps in this direction.

Another interesting result of our study is the observation that system performance may differ a lot from topic to topic. This phenomenon was observed for every kind of explored performance measures. Even though, our normalized precision measure tries to remove this factor for one set of measures, the difference between topics is still visible. Not only specific performance measures vary from topic to topic, but also the performance balance between systems vary. While in user-oriented measures CAFÉ was better in general and for most of the topics, there were always topics where the baseline system demonstrated better performance. Our study demonstrated that CAFÉ is a great *enhancement* to a search-only system, but not a *replacement* since it still can't beat traditional search for certain topics. A proper approach is to combine both so that one engine's advantages can cover the other's weakness.

As we mentioned above, the presented work was performed in the context of a large-scale collaborative project aimed to develop a new generation of systems for task-based information exploration. The current design does combine a search engine and an adaptive filtering engine. In this context, our next evaluation goal is to evaluate the combined system in task-based information exploration using the same evaluation framework and a similar study design. We are also currently working on expanding what we have learned from this study to develop evaluation frameworks and new evaluation measures for a range of similar information exploration systems.

Acknowledgements

We want to thank Dr. Shulman and his QDAP center for their help in developing the ground truth annotations. Thank Qi Li and Jongdo Park for helping in results analyses. This work is partially supported by DARPA GALE project.

References

- Acosta-Diaz, R., Guillen, H. M., GarciaRuiz, M. A., Gallardo, A. R., Pulido, J. R. G., & Reyes, P. D. (2006). An open source platform for indexing and retrieval of multimedia information from a digital library of graduate thesis. In *Proceedings of world conference on E-learning in corporate, government, healthcare, & higher education E-learning 2006 Honolulu* (pp. 1822–1829). Hawaii: AACE.
- Allan, J. (2002). *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers.
- Allan, J. (2003). HARD track overview in TREC 2003 high accuracy retrieval from documents. In *The twelfth text retrieval conference*.
- Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3).
- Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3), 225–250.
- Cleverdon, C. W., Mills, J., & Keen, M. (1966). *Factors determining the performance of indexing systems*. Cranfield: ASLIB Cranfield Project.
- Díaz, A., & Gervás, P. (2005). Personalisation in news delivery systems: item summarization and multi-tier item selection using relevance feedback. *Web Intelligence and Agent Systems*, 3(3), 135–154.
- Dumais, S. T., & Belkin, N. J. (2005). The TREC interactive tracks: putting the user into search. In E. M. Voorhees & D. K. Harman (Eds.), *TREC: Experiment and evaluation in information retrieval* (pp. 123–152). MIT Press.
- Fiscus, J., & Wheatley, B. (2004). Overview of the TDT2004 evaluation and results. In *Proceedings of TDT-04*.
- Fuhr, N., Govert, N., Kazai, G., & Lalmas, M. (2002). INEX: INitiative for the evaluation of XML retrieval. In *Proceedings of the SIGIR 2002 workshop on XML and information retrieval*.
- Gotz, D., Zhou, M. X., & Aggarwal, V. (2006). Interactive visual synthesis of analytic knowledge. In P. C. Wong & D. Keim (Eds.), *IEEE symposium on visual analytics science and technology, VAST 2006* (pp. 51–58). Baltimore, MD: IEEE.
- Hanani, U., Shapira, B., & Shoval, P. (2001). Information filtering: overview of issues, research and systems. *User Modeling and User Adapted Interaction*, 11(3), 203–259.

- He, D., & Demner-Fushman, D. (2003). HARD experiment at maryland: from need negotiation to automated HARD process. In *Proceedings of text REtrieval conference (TREC) 2003*.
- Ingwersen, P. J., & Kalervo (2005). *The Turn: integration of information seeking and retrieval in context*. Springer.
- Kando, N. (2005). In *Proceedings of the fifth NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access Tokyo, Japan*.
- Larsen, B., Malik, S., & Tombros, A. (2005). The interactive track at INEX 2005. In *The workshop of INEX 2005*.
- Larsen, B., Ingwersen, P., & Kekalainen, J. (2006). The polyrepresentation continuum in IR. In *1st international conference on IR in context*.
- Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4), 41–46.
- Marchionini, G., & Shneiderman, B. (1988). Finding facts vs. browsing knowledge in hypertext systems. *IEEE Computer*, 21(1), 70–79.
- McColgin, D., Gregory, M., Hetzler, E., & Turner, A. (2006). In: White, R.W., Muresan, G., Marchionini, G. (Eds.). *From question answering to visual exploration in workshop on evaluating exploratory search systems at SIGIR 2006*. (pp. 47–50).
- Peters, C., Gey, F., Gonzalo, J., Mueller, H., Jones, G., & Kluck, M. (2006). *Accessing multilingual information repositories: 6th workshop of the cross-language evaluation forum*. Springer.
- Pirolli, P., & Card, S.K. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of 2005 International Conference on Intelligence Analysis*, McLean, VA, 2–4 May 2005.
- Pirolli, P., & Fu, W.-T. (2003). SNIF-ACT: A model of information foraging on the World Wide Web. In P. Brusilovsky, A. Corbett, & F. d. Rosis (Eds.), *9th international user modeling conference* (pp. 45–54). Berlin: Springer-Verlag.
- Robertson, S., & Soboroff, I. (2002). The TREC 2002 filtering track report. In *Proceedings of TREC 2002*.
- Robertson, S. E., & Hancock-Beaulieu, M. M. (1992). On the evaluation of IR systems. *Information Processing and Management*, 28(4), 457–466.
- Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In: *Proceedings of SIGIR '95* (pp. 138–146).
- Voorhees, E. M., Harman, D., K., 2005. TREC: Experiment and evaluation in information retrieval: MIT Press.
- Waern, A. (2004). User involvement in automatic filtering – an experimental study. *User Modeling and User Adapted Interaction*, 14, 201–237.
- White, R. W., Muresan, G., & Marchionini, G. (2006). Evaluating exploratory search systems. In: *Evaluating exploratory search systems, a workshop of ACM SIGIR06*.
- White, R. W., Jose, J. M., & Ruthven, I. (2004). An implicit feedback approach for interactive information retrieval. *Information Processing and Management*, 42(1), 166–190.
- White, R. W., Kules, B., Drucker, S. M., & Schraefel, M. C. (2006). Supporting exploratory search. *Communications of the ACM*, 49(4), 37–39.
- Wong, P. C., Chin, G., Jr., Foote, H., Mackey, P., & Thomas, J. (2006). Have Green – a visual analytics framework for large semantic graphs. In P. C. Wong & D. Keim (Eds.), *IEEE symposium on visual analytics science and technology, VAST 2006* (pp. 67–74). Baltimore, MD: IEEE.
- Yang, Y., Lad, A., Lao, N., Harpale, A., Kisiel, B., & Rogati, et al. (2007). Utility-based information distillation over temporally sequenced documents. In *Proceedings of ACM SIGIR'2007*.
- Yang, Y., Yoo, S., Zhang, J., & Kisiel, B. (2005). Robustness of adaptive filtering methods in a cross-benchmark evaluation. In *28th Annual international ACM SIGIR conference Salvador* (pp. 98–105). Brazil: ACM Press.
- Zhang, Y. (2004). Using bayesian priors to combine classifiers for adaptive filtering. In *27th Annual international ACM SIGIR conference, Sheffield, United Kingdom* (pp. 345–352).