Evaluating the Content and Usability of an Experimental Text Summarization System and Three Web-Based Search Engines

1. Introduction

The World Wide Web offers unprecedented opportunities to provide the general population with access to information about their health care questions. The usability of Web-based systems emerges as a growing concern (Nielsen, 2000), as improvements in retrieving larger amounts of information exacerbate the problem of matching the information needs of patients and the general population. Each day more and more patients and their families try to find information about their health problems on the WWW to aid in their decision making and management of disease. The ability to customize and summarize information relevant to end users' needs is an area where considerable research is currently being directed (Rice *et al*., 2001). In this paper, we describe the evaluation of Centrifuser, a system developed at Columbia University that provides context-sensitive text summarization to support information retrieval. Centrifuser provides three different types of output that are designed to assist users in understanding the documents returned by a search engine (Kan *et al*., 2001): (1) it provides navigation links for users to focus in on specific subtopics present in the documents or to broaden the user's search query; (2) it composes a multi-document summary by identifying sections of repeated information across documents and extracting representative sentences from these sections (with the assumption that repeated information is important); and 3) it identifies the salient, indicative differences between documents and highlights this information for the user. For example, a search on "diabetes" turns up many documents containing information on symptoms, diagnosis and treatment of the disease. In this case, Centrifuser constructs an overview consisting of sentences from each of these common sections, and highlights documents that have unique information (e.g., "Document A has the most information on diabetes treatment.").

In order to evaluate Centrifuser and its comparability to currently available Web-based search engines, we employed an approach based on usability testing and cognitive analysis. In our previous work we have employed video recording of user interactions and audio recording of either subjects' thinking aloud or actual dialogue while using health care information systems (Kushniruk *et al.*, 1997). We have also coded audio and video data to identify potential issues in the design of Web-based information systems (Kushniruk *et al.*, 2001). In this paper, we extend this type of analysis to the evaluation of both the content and user interface of Centrifuser and three commonly used Web-search engines as a method of comparison. Objectives of our work included assessing Centrifuser's usefulness in answering users' real information needs, as well as determining how well the system compares with common search engines.

2. Method

Queries: Medical professionals were consulted to select three widely applicable medical conditions that we used in evaluating the interfaces: diabetes, hypertension (high blood pressure) and angina (chest pain).

Subjects: Thirteen subjects participated in this study. All subjects were recruited from the waiting room at the intensive care unit of a large hospital. All subjects were either friends or relatives of patients undergoing treatment at the hospital for one of the three conditions. Thus, the subjects were ideal to evaluate these interfaces because they had information needs that matched with the interfaces' output.

Procedure: Subjects were asked to select one of the three conditions that they wanted further information about – "tell me about angina", "tell me about diabetes", or "tell me about hypertension". Subjects were then sequentially presented with their selected query results as displayed by the four systems (Centrifuser, Yahoo, Google and About.com) in random order. Subjects were asked to verbalize their thoughts or "think aloud" as they examined each of the interfaces. Additionally, subjects were probed about their thoughts regarding certain aspects of the interface, i.e. its ability to fulfill their information needs, its ability to allow for navigation, and its presentation. After viewing each of the four interfaces, subjects were then asked to complete seven-point Likert scales addressing the following areas: a) usefulness of content, b) types of information available, c) ease of deciding next step, d) ease of locating information, e) layout and f) overall satisfaction

Data analysis: All numerical ratings of the interfaces by subjects were tabulated. The audio portion of the subjects' "thinking aloud" and response to probes were first transcribed verbatim. A coding scheme was adapted from previous work on health information systems (Kushniruk *et al.*, 1996) to tag comments on various aspects of the usability of the interfaces. The scheme included categories for subject comments on: understanding of information, usefulness of information, content of information, linkages to other sites, organization of information, interface consistency, and understanding labels and instructions. Both the audio and video transcripts were enriched with these tags by applying an analysis of video data of human-computer interaction previously developed by the authors (Kushniruk *et al.*, 1996). This involved annotating the verbal transcripts with the codes, i.e. "time-stamping" the coded sections of the transcripts to the corresponding video sequences of the user's interactions with the system.

3. Results

Each hour of video data took about two to three hours for one experimenter to code and analyze. An excerpt from the coded transcript of a subject "thinking aloud" while interacting with one of the three search engines (Yahoo) is given below in Figure 1.

44:58 - SUBJECT SCROLLS DOWN PAGE WITH LINKS TO OTHER SITES
"Well I like how that by the links it has all the information. But if it is most popular sites, it should only have ten links and if you wanna have more you should go to, you should have a next button so they could see more links if they can't find what they are looking

for on the first page"
COMMENT – ORGANIZATION OF LINKAGES

The following is the same subjects initial reaction to the next system he viewed (Centrifuser):

45:56 – SUBECT VIEWS SUMMARY ON SCREEN
"This one is better than the previous system because right away it tells you about the subject, because it only has and tells you about what you are going to be looking at"
COMMENT – CONTENT OF SUMMARY

Figure 1. Excerpt of coded transcript. The time on video counter is indicated, as are the annotations and coding of this section of the transcript

3.1 General Usability of Centrifuser

In general, subjects were positive about their interaction with Centrifuser and found the information it provided in its synopsis to be useful. Table 1 provides the frequency of coded comments, made by each of the 13 subjects where suggestions for *improving* the usability of the Centrifuser interface and/or content of information were offered.

Usability Issue	Subjects												
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
Content – Synopsis	2		1			1					1		
Content – Overall level									1				1
Content – Organization		1								1			
Labeling of Sections					1			1		1	2		2
Labeling of links		1									1		
Too much text					1								
Too little information		1										1	
Format of information					1								
Need for search facility							1						

Table 1 – Frequency of Categories of Subjects' Suggestions regarding Centrifuser's Usability and Content

In order to fine-tune the system, in-depth analyses of the transcripts were conducted to pinpoint the nature of the users' comments for each of the coded and time-stamped issues that they raised. For example, for the category in Table 1 "Labeling of sections", by examining the "think aloud" protocols of subjects 5, 8, 10, 11 and 13, it seemed that these subjects were unclear of what information would be contained in the section of Centrifuser interface labeled "Differences between documents". Comments related to content of the information provided were likewise considered in light of the verbatim transcripts, one subject (S1) indicated that the synopsis generated by the system should begin with a definition of the medical condition the text dealt with, while analysis of interaction with another subject indicated the reading level of the synopsis in terms of medical content might need to be adjusted to take into account users with less education. Based on these analyses, we are currently modifying Centrifuser and plan to conduct a subsequent round of iterative usability testing with a new set of users.

We also extended the analyses of users' interactions with the three search engines (Yahoo, Google, and About.com) which Centrifuser was compared with (described in the

next section). Our preliminary qualitative analyses has pinpointed a number of distinct advantages and disadvantages of each the four interfaces. For example, while subjects liked About.com for its organization, clarity of labeling and its linkages, they were critical of the relevance of links that Google provided, in the context of their specific health care question.

3.2 Comparative Ratings of all Interfaces

As mentioned, after examining the interfaces sequentially, subjects assigned numerical ratings to compare and rank the four interfaces. Subjects were asked to consider each system interface's quality in terms of a) content, b) types of information available, c) ease of deciding next step, d) ease of locating information, e) layout and f) overall satisfaction. Ratings were performed on a seven-point scale where 1 was lowest possible and 7 was the best possible score. As the focus of the evaluation as a whole emphasized capturing qualitative feedback that was quite time intensive, statistical significance was not reached and results in this quantitative section are preliminary. Figure 2 shows the average score across the 13 subjects for each question and system combination.

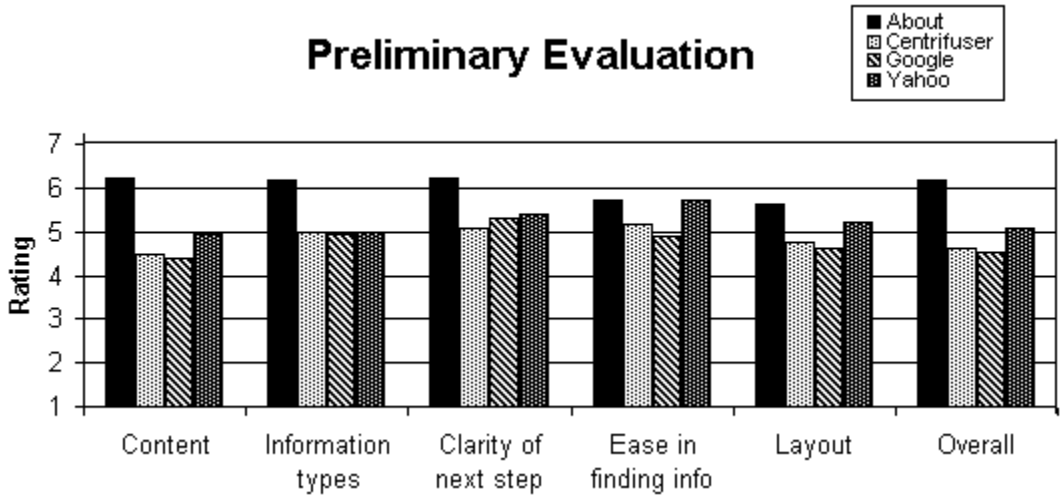

Figure 2. Quantitative Evaluation

About.com's human expert site consistently outperformed all of the other system interfaces, with an emphasis on high quality content and range of different ways to access that information (questions 1 and 2). Yahoo's human created hierarchy performed next best, consistently outscoring or equaling the remaining two systems. Yahoo performed least well comparatively on providing different information access mechanisms (again, question 2). Centrifuser and Google form the lower tier. They both used the same underlying documents (as Centrifuser post-processes Google output), but had their strengths and weaknesses.

According to the subjects, Centrifuser's layout provided an easy way to locate relevant information, whereas Google's consistent placement of links may have been the reason

that subjects found it easy to decide what action to take next. Centrifuser and Google are the least distinct in the evaluation; larger-scale evaluation is necessary to properly assess their differences, and is planned in future work.

4. Discussion

In this paper, we have employed a usability engineering approach to the analysis of a new text summarization system. By collecting and analyzing both video and audio data on users interactions with the system, we have been able to characterize those aspects of WWW interfaces that are useful to patients and their families who are seeking health information in response to questions. Additionally, by coding for categories of user comments we have located areas where the system can be improved. We are currently applying the analysis results to modify Centrifuser for a second round of data collection with new subjects. In general, this approach to data collection, analysis and reprogramming has lead to systems that are more acceptable in areas such as healthcare (Coble *et al*., 1997; Kushniruk *et al*., 1996). With the widespread use of Web-based information resources by patients and their families, this type of user-centered evaluation is increasingly important.

By having subjects compare Centrifuser with three conventional search engines, we found that no one system contained features or capabilities that completely met the needs of all subjects. Although we found general trends where one system was rated slightly higher than another on particular criteria, analysis of the "thinking aloud" data indicated that there was greater consistency of user reactions when the results were categorized by user interface feature than by entire system. For example, the capability of a system to provide relevant linkages (such as Google) was well received by subjects, as was the capability of providing users with a focused summary of multiple sources of information (as Centrifuser does). Our current work aims at teasing apart these factors to provide a rational basis for the engineering of information systems that more closely match the information needs of real users.

References

Coble, JM, Karat, J, Orland, MJ, Kahn, MG. (1997). Iterative usability testing: Ensuring a usable clinical workstation. In Proceedings of the 1997 AMIA Annual Fall Symposium, 744-748.

Kan, M, McKeown KR, Klavans, JL (2001). Domain-specific informative and indicative summarization for information retrieval. In Proceedings of the Document Understanding Workshop (DUC 2001), New Orleans: September 2001.

Kushniruk, AW, Patel, VL, Cimino, JJ, Barrows, R (1996). Cognitive evaluation of the user interface and vocabulary of an outpatient information system. In Proceedings of the 1996 AMIA Annual Fall Symposium, 22-26.

Kushniruk, AW, Patel, VL, Cimino, JJ (1997). Usability testing in medical informatics: Cognitive approaches to the evaluation of information systems and user interfaces. In Proceedings of the 1997 AMIA Annual Fall Symposium, 218-222.

Kushniruk, AW, Patel, C, Patel, VL, Cimino, JJ (2001). "Televalution" of information systems: An integrative approach to the design and evaluation of Web-based systems. International Journal of Medical Informatics, 61(1), 45-70.

Nielsen, J (2000). Designing web usability. Indianapolis: New Riders Press.

Rice, RE, McCreadie, M, Chang, SL. (2001). Accessing and browsing: Information and communication. Cambridge, MA: MIT Press.