Usability Evaluation of an Experimental Text Summarization System and Three Search Engines: Implications for the Reengineering of Health Care Interfaces

Andre W. Kushniruk, PhD[1], Min-Yen Kan[2], Kathleen McKeown, PhD[2], Judith Klavans[2], PhD, Desmond Jordan[3], MD, Mark LaFlamme[2], MD, Vimla L. Patel, PhD, DSc[3]

[1] Department of Mathematics and Statistics, York University, Toronto, Ontario,
[2] Department of Computer Science, Columbia University, New York, New York,
[3] Department of Medical Informatics, Columbia University, New York, New York

ABSTRACT

This paper describes the comparative evaluation of an experimental automated text summarization system, Centrifuser and three conventional search engines – Google, Yahoo and About.com. Centrifuser provides information to patients and families relevant to their questions about specific health conditions. It then produces a multidocument summary of articles retrieved by a standard search engine, tailored to the user's question. Subjects, consisting of friends or family of hospitalized patients, were asked to "think aloud" as they interacted with the four systems. The evaluation involved audio- and video recording of subject interactions with the interfaces in situ at a hospital. Results of the evaluation show that subjects found Centrifuser's summarization capability useful and easy to understand. In comparing Centrifuser to the three search engines, subjects' ratings varied; however, specific interface features were deemed useful across interfaces. We conclude with a discussion of the implications for engineering Web-based retrieval systems.

INTRODUCTION

With the growing amount of health related literature on the World Wide Web (WWW), the efficient retrieval of health information relevant to the information needs of patients is becoming a major problem. Patients and families now commonly use Web-based resources to aid in decision-making and disease management. The effectiveness and usability of accessing these Web-based resources is emerging as a growing concern[1]. Conventional search engines often generate a large number of "hits", presented as a ranked list of documents. For example, using traditional search engines, a search on the patient's question "what is angina?" turns up a list of documents containing widely dispersed information on symptoms, diagnosis and treatment of the condition. The cognitive load and time needed to process these lists limits the efficiency and usability of these search engines in many real-world healthcare contexts. Thus, the ability to customize and summarize information and present it in a usable manner relevant to users is growing in importance.

Given the above considerations, the Centrifuser system was developed to provide context-sensitive text summarization[2]. Centrifuser post-processes articles retrieved by a conventional search engine to provide a customized summary with respect to a user's query (see Figure 1). The system also highlights the documents that have unique information (e.g. "Document B has the most information on angina"). Centrifuser extracts representative sentences from the different documents based on the principle that information that is repeated in different documents is likely to be important. The interface also presents an automatically generated summary with which the user may then select from a set of navigational links to focus on a specific subtopic or to broaden the search for information.

We conducted an evaluation to assess the potential of a system such as Centrifuser, in comparison to currently available Web-based search engines. We used an approach based on usability testing and cognitive analysis, which employed subjects consisting of relatives and

We found 4 documents relevant to your question: angina

Browse to narrower subtopics: [definition][causes][cause][signs and symptoms][diagnosis][prevention][prognosis][home remedies and alternative therapies][male/female differences][treatment][coronary arteriography][continuous ecg monitoring][essential workup][laboratory][imaging/special tests][exercise tolerance testing][angiography][antiplatelet drugs][medication][initial stabilization][variant angina][aspirin]

Synopsis of the documents: Angina, also called angina pectoris, is temporary chest pain or a sensation of pressure that occurs while heart muscle isn't receiving enough oxygen. Typically, angina is described as a pressing or squeezing pain that starts in the center of the chest and may spread to the shoulders or arms (most often on the left side although either or both sides may be involved), the neck, jaw, or back. If the angina continues in spite of the use of medications or occurs more often or with greater intensity, your physician may consider coronary angioplasty or coronary artery bypass surgery (see pages 665 and 666). Treatment begins with attempts to prevent coronary artery disease, to slow its progression, or to reverse it by dealing with its known causes (risk factors). interfere with the effects of the hormones epinephrine (adrenaline) and norepinephrine (noradrenaline) on the heart and other organs.

Differences between the documents:

- Mayo Clinic family health book contains information on rare topics, contains topics such as "signs, symptoms", "medication", "exercise" and "angina pectoris" and is a lot shorter than others.
- The Columbia University College of Physicians and Surgeons complete home medical guide and The Merck manual of medical information are generally related to your query. All of the documents discuss topics such as "coronary artery bypass surgery " and "angina". The second document is longer than most other documents. The first document contains significantly less material than average.
- 5 minute emergency medicine consult doesn't seem to be related to the main sense of your query and contains significantly less material than average.

Figure 1 Output from the Centrifuser interface on the query of “angina”

friends of patients in a real health care settings (a waiting room outside of an operating room at a large metropolitan hospital). To conduct the evaluation we employed full video recording of user interactions and audio recording of subjects’ thinking aloud while using Centrifuser and the other search engines.[3] We also extended our approach to coding audio and video data to identify potential issues in the design of Web-based information retrieval systems.[4] We now describe an application of this analysis to the evaluation of Centrifuser and three commonly used Web-search engines. The objective of our work included assessing Centrifuser’s capability in addressing users’ information needs along the dimensions of content and user interface.

METHODS

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 described above.

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”. Then they were sequentially presented with their selected query results as displayed by the four systems (Centrifuser, Yahoo, Google and About.com) in random order. We asked the subjects 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 all 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 to tag comments on various aspects of the usability of the interfaces[3]. The scheme included

2:02:45 Oh this is good, because it gives you a lot of, the angina, you can find the definition, the cause, symptoms and treatments, I think it covers everything, it gives definition, gives symptoms cause and treatment
COMMENT – RANGE OF CONTENT AND COVERAGE OF MATERIAL
This sums up everything we want in a nutshell, I find the synopsis very useful
COMMENT – USEFULNESS OF CONTENT - SYNOPSIS
2:03:35 This is good here because it tells you the different kinds, tells you there are four articles
COMMENT – USEFULNESS OF CONTENT – ARTICLES
This "differences between documents", I'm assuming this will show me the different types of angina with heart attacks, this is not clear
POTENTIAL PROBLEM – UNDERSTANDING OF LABEL ("DIFFERENCES BETWEEN ARTICLES")

Figure 2 Excerpt of a coded transcript of a subject while examining the Centrifuser system

categories for subject comments regarding: 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[3,5]. 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.

RESULTS

Each hour of video data took about two to three hours for one experimenter to code and analyze. The coding was reviewed by a second research assistant, with minor disagreement being resolved during subsequent discussion. An excerpt from the coded transcript of a subject "thinking aloud" while interacting with Centrifuser is given in Figure 2 (coded comment categories are bolded and numbers indicate the corresponding time offset)

Qualitative Usability Analysis

In considering the subjects' comments, no one interface was found to be clearly superior, but rather certain features of the different interfaces and systems were identifiable as being useful or as being problematic by subjects. In order to obtain preliminary data on which features of the systems were identified as being useful or problematic, the coded comments were classified as being either positive or negative regarding the particular system feature (again, see Figure 2). For example, comments where the subject indicated there was a problem with regard to a specific feature of the system, such as poor navigation capability, would be classified as a negative comment regarding that feature. Table 1 summarizes the data coded this way.

Examination of Table 1 sheds light on which features of each of the systems tested were considered in a positive light by subjects and which were considered negatively. For example, subjects commented in a positive manner about the range of links provided by the About.com interface (e.g. links to related resources, news groups, articles etc.). Examination of the frequencies of positive and negative comments indicate certain patterns regarding what system features were found desirable or not across systems. While subjects liked About.com for its clarity of labeling and its range of linkages to broad resources, they were critical of the relevance of links that Google provided (several subjects commented that they felt Google did not filter their information request very well, providing links to many irrelevant sites), in the context of their specific health care question.

Regarding Centrifuser, the majority of positive comments were made regarding the content of information provided (in terms of its usefulness and understandability). Specific comments were made by several of the subjects regarding the perceived usefulness of having a synopsis made available to them in response to their queries (as illustrated in the excerpt in Figure 2). In order to further fine-tune Centrifuser, in-depth analyses of the "think-aloud" transcripts were conducted to pinpoint the nature of the users' comments for each of the coded and time- stamped issues that they raised. Comments related to content of the information provided were considered in

Features: Content and Usability	Search Engine							
	Centrifuser		Yahoo		Google		About	
	+'ve	-'ve	+'ve	-'ve	+'ve	-'ve	+'ve	-'ve
Content – Usefulness	10	1	2	1	1	4	3	1
Content – Overall understanding	9	3	1	1	1	1	4	1
Organization of Information	2			5	1		5	4
Understanding labels		5		2		3	7	
Navigational ability	3		1	1		1	1	2
Effort to find information	1	1		1		5	1	
Relevance of Links	1			1		9		
Amount of Information		2		6		2	1	4
Number of links available	2	1	1			4	3	
Range of information available	4		1		1		10	
Format/layout of information						1	2	1
Search capability		1					1	

Table 1 Frequency of positive (+'ve) and negative (-'ve) coded verbal comments made by subjects regarding the usability and content of the four systems tested

light of the verbatim transcripts, with a few subjects indicating 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 other subjects indicated the reading level of the synopsis in terms of medical content might need to be adjusted to take into account users with less knowledge of medical terms. Based on these types of in-depth analyses, modifications have since been made to Centrifuser. We are currently conducting a subsequent round of usability testing, under similar study conditions (with new subjects) to assess the impact of these changes. In addition, the newly modified system is now be tested over the WWW, with subjects being sequentially presented with Centrifuser and the three search engines, and then asked to rate them regarding their content, presentation and overall usefulness.

Comparative Ratings of all Interfaces

After examining the interfaces sequentially, subjects assigned numerical ratings to compare and rank the four interfaces. 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 3 shows the average score across the 13 subjects for each question and system combination.

About.com's human generated site was generally rated higher than the other system interfaces, with an emphasis on high quality content and range of different ways to access that information. Yahoo's human created hierarchy performed next best, consistently outscoring or equaling the remaining two systems. Yahoo performed least well, comparatively, in providing different information access mechanisms. 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 overall in the ratings; larger-scale evaluation is necessary to properly assess their differences, and is planned in future work.

DISCUSSION

In this paper, we have employed a usability engineering approach to the analysis of a new text summarization system. By collecting and

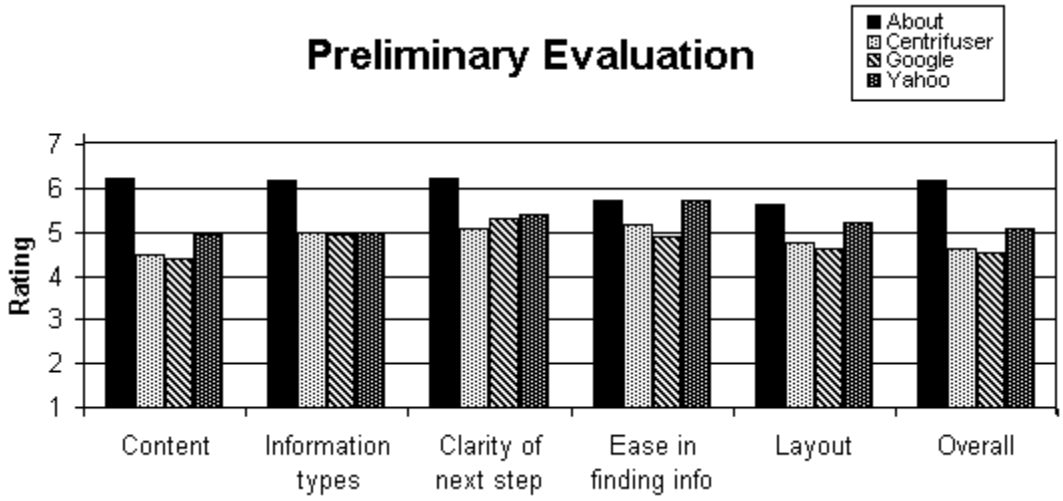

Figure 3 Quantitative Evaluation

analyzing both video and audio data on users interactions with the system, we have been able to characterize those aspects of Web interfaces that are useful to health information seekers. Additionally, by coding for categories of user comments, we have located specific areas where healthcare information systems can be improved. We have applied the results to modify Centrifuser, and have planned a second round of data collection with new subjects. In general, this iterative approach to data collection, analysis and reprogramming can lead to systems that are more acceptable in areas such as healthcare[6]. 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. However, the method employed in this paper can be successfully used to tease apart and identify which type of features work for users under different task conditions. Although we found general trends where one system was rated slightly higher than another on particular criteria, analysis of the "think aloud" data indicated that there are features of each of the interfaces tested which subjects preferred (e.g., the capability of providing users with a focused summary of multiple sources of information). Our current work aims at further teasing apart these factors to provide a rational foundation for the reverse engineering of new information systems (based on our analyses) that more closely match the information needs and requirements of users.

REFERENCES

1. Nielsen, J (2000). Designing web usability. Indianapolis: New Riders Press.
2. 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.
3. 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, 218-222.
4. Kushniruk, AW, Patel, C, Patel, VL, Cimino, JJ (2001). "Televaluation" of information systems: An integrative approach to the design and evaluation of Web-based systems. International Journal of Medical Informatics, 61(1), 45-70.
5. Kushniruk, AW (2002). Analysis of complex decision-making processes in health care: Cognitive approaches to health informatics. Journal of Biomedical Informatics, 34, 365-376.
6. Coble, JM, Karat, J, Orland, MJ, Kahn, MG. (1997). Iterative usability testing: Ensuring a usable clinical workstation. In Proceedings of the 1997 AMIA, 744-748.