Document thumbnail visualizations for rapid relevance judgments: When do they pay off?

William Ogden, Mark Davis, Sean Rice
Computing Research Lab
New Mexico State University

In a recent experiment, Kaugars (1989) developed a document visualization system that displayed the results of a keyword text search as a set of small graphical representations (thumbnails) of the top 20 retrieved documents laid out in one window and a single document viewer in a second window (See Figure 1). Document thumbnails presented color-coded highlighting indicating positions and identities of search terms in the document, and the single document viewer showed a fish-eye view of the document focused on multiple regions defined by the sentences holding search terms. He compared this system to another more traditional system that represented documents as lists of short titles and a simple scrolled window document viewer. The users judged 8 sets of documents in accordance with relevance judgments supplied with 8 selected TREC-6 topics. The documents and topics were all pre-selected for the users who had no part in selecting query terms or interacting with the retrieval software. Kaugars found that people were faster and better at making relevance judgments for a fixed set of retrieved documents when using the thumbnail/fish-eye system.

There are at least two important aspects of the Kaugars interface that could have led to its superior performance, either the document-thumbnails view or the fish-eye document view (or both). The thumbnail view of the documents allow users to quickly scan the returned document set for instances of keywords and keyword collocations and their distribution in and between documents. It gives the user in a single glance information about which documents contain which keywords and which keywords are missing or how often they appear in the documents. Because the thumbnails retain the familiar shape and format of a document, the user can easily see how the keywords are distributed in the actual document. They could help users to locate information within a document and could help them answer the question “why was this document retrieved?”

Alternatively, the unique document viewer could have led to the superior performance. A fish-eye view presents the area of current interest in normal scale but as distance from the interest area increases, information scale decreases. In Kaugar’s document viewer, there could be multiple areas of interest, each defined by the presence of a search term in a sentence. These areas would be displayed in a normal font. Other sentences and paragraphs were shown in a smaller font. This allowed the user to very easily find and read relevant passages while ignoring intervening and mostly irrelevant text. Users could define new areas of interest in the document with a mouse click thereby returning those areas to a normal sized font. Because the documents were much smaller in size than the normal full sized scrolled view, much more of the relevant text fit within a single window on the computer screen perhaps making it easier to make relevance judgements.

From this experiment by Kaugars, we cannot determine the relative importance of these two interface features. The goal of the current study was to investigate the role of the thumbnail-document view feature, exclusive of the fish-eye document view feature. In addition, we wanted to test the potential advantage of the thumbnail view within the context of an interactive text retrieval task that engaged the user more than the relevance judgement task. Thus we attempted to replicate Kaugars result using an interactive, WEB version of a thumbnail document set viewer and using the prescribed TREC-7 interactive track methodology. The system (named J24 after the July 24th deadline for its completion) was built using the Unicode Retrieval System Architecture (URSA) text retrieval software library developed the Computing Research Lab at New Mexico State University. The system provides a WEB interface for entering search terms and displaying results with thumbnails for 10 documents at a time. The J24 system was compared to a control system, ZPRISE.

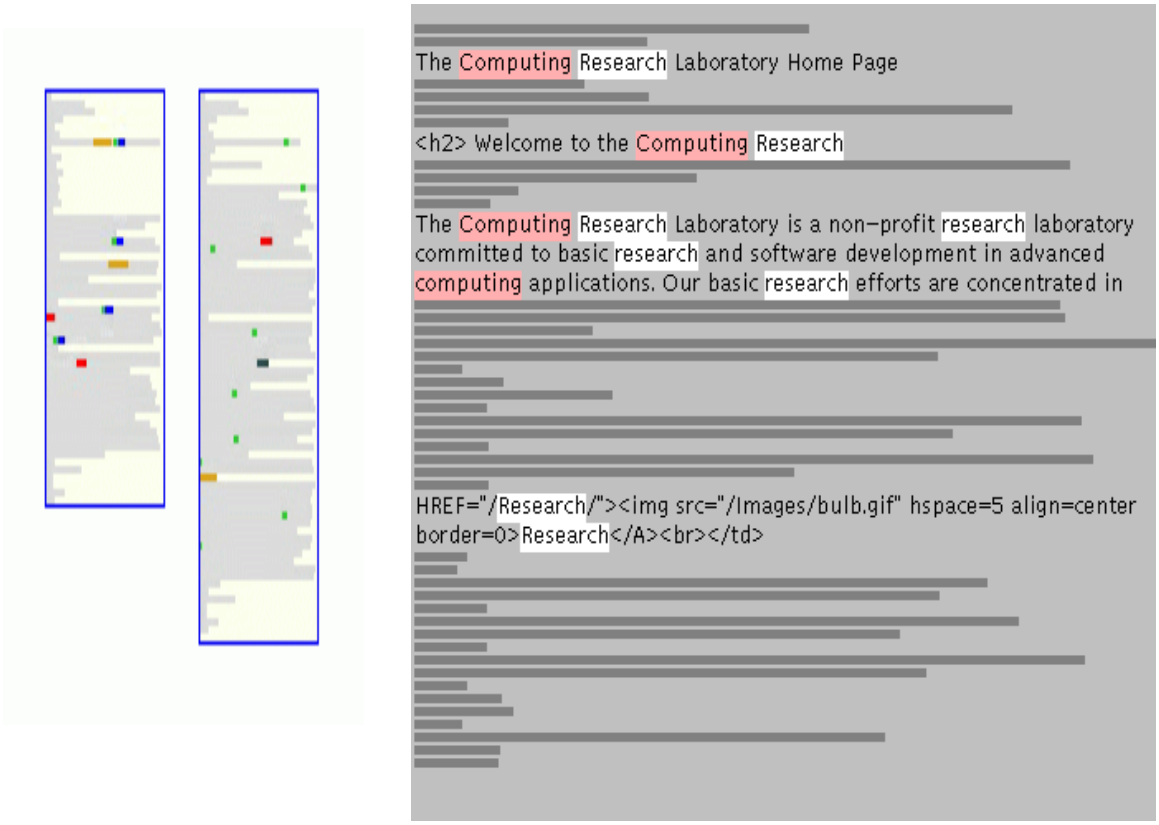


Figure 1. Document views: thumbnails (left) and multiple fish-eye (right)

J24: Interactive System Description

The experimental retrieval system for NMSU's interactive studies was J24, a newly developed retrieval engine based on NMSU's URSA (Unicode Retrieval System Architecture). The URSA project, funded under the US Government's TIPSTER text program, involves developing a Unicode retrieval engine that indexes UCS2 text, a 16-bit UNICODE representation. Like most text retrieval systems, URSA has two primary components, an indexing engine and a retrieval engine. For interactive studies, the J24 interface is dynamically generated by a rapid-configuration HTML generator that calls URSA libraries through a comprehensive API.

Converting texts to UCS2 is accomplished using a utility that can convert 104 native codesets to and from UCS2 representations. After conversion, the UCS2 texts are tokenized. URSA's tokenizer is language independent and incorporates several thousand rules that describe tokenization boundaries across languages, tokenization contexts and special handling for detecting sentence and paragraph boundaries. The collected rules embody knowledge about how, for example, "full-width" punctuation in Chinese serves to separate sentences, or how ellipsis (both vertical and horizontal) appears in text from Greek to Japanese. The tokenization process also can be configured to make use of highly language-dependent features like Chinese segmentation. Segmentation markup is analyzed as non-spacing characters in the UCS2 data stream. Other hooks are available for complex morphological analysis modules. Less intelligent backup methods are implemented for n-gram tokenization of CJK (Chinese-Japanese-Korean) languages, with variant configurations for separating Katakana, Hiragana and Kanji in Japanese, for example, prior to n-gram tokenization of Kanji sequences, or for breaking n-gram sequencing in Korean on boundaries between embedded Chinese characters (Hanja) and phonetic sequences (Hangul). In URSA, simple finite state morphological analysis routines for suffix stemming and term normalization have also been implemented for Spanish, French, English, German and Italian, reflecting the TREC language focuses of the CLIR and multilingual tracks for the past several years.

The indexing engine sequentially writes sorted postings to disk in batches as indexing progresses, then merges the resulting posting sets after completing a single pass through the document set. In URSA the final, merged postings are compressed using a simple, but fast integer compression scheme. This results in indexes that are only around 25% the size of the original text for non-phrase indexing of large (500Mb+) text collections. The converted UCS2 documents can be discarded at index time for additional space savings. The UCS2 tokens in the retrieval lexicon represent an extremely small fraction of the size of the postings, so the space cost of using UNICODE is very low. The time for converting the native codeset's of complex scripts is actually among the most costly of the URSA processes.

We implemented a comprehensive subset of the Zobel and Moffat (1998) octet weighting schemes for query and document term weighting, normalization and combination in URSA. The subspace implemented in URSA is characterized by the regular expression, [A-F][A-I]-[A-B][A-F][A-N]-[A-B][A-E]A, which represents 75,600 possible weighting schemes. For the J24 interactive experiments, the scheme AI-AFD-BCA was used. This scheme may be summarized as using an inner product combination of query and document term weights, with collection terms weighted by entropy, document terms weighted by the Okapi formulation (Robertson, et.

al., 1995), then normalized by the root of the number of unique terms in a given document. Query terms are simply weighted by their collection entropy in this scheme. This scheme was found by Zobel and Moffat (1998) to represent the best scheme for short (title) queries out of the multiplicity of schemes that they evaluated. Since interactive queries are typically very short, in part perhaps due to the experience of users with WWW search engines, the tuned octet formulation for short queries was a good choice. Query speeds in URSA are on the order of 1.2 seconds per query (using 350 "full" Trec queries as a test) when the index is on disk local to the HTTPD server daemon. Shorter queries consisting of just a few keywords take considerably less time, and are effectively instantaneous for interactive users, or at least as quick as any WWW search engine.

The user interface design based on the multi-frame HTML page is generated by a CGI process in response to a user request. The thumbnails are jpeg images generated in real-time for each retrieved document (we used the freely available libjpeg to compress in-memory bitmaps into jpegs), as are the modified documents with embedded font colorations. No special pre-processing is done to the collection to make thumbnails, summaries or highlighted documents available to the user—it's all done at the time of the query request. Javascript is used to interactively update images that indicate active documents as the user clicks on individual thumbnails or summaries in their respective frames. To conserve disk space at the server, a "reaper" process removes thumbnails and highlighted documents that are greater than 30 minutes old.

Interactive TREC studies presented a unique burden for a browser-based interactive system in that user relevance judgments and query timing had to be recorded. The J24 modification to support these features were significant and complex, requiring that state information be written out to the server each time a user preceded to another screen. Since the system is capable of presenting every document that matches a query in blocks of 10 at a time, the state information had to be used to dynamically change the interface as the user made choices, viewed individual documents and marked documents as relevant. The resulting system records initial query time, times for query reformulation, relevance judgments and observed documents as test subjects make use of the system. See the screen shot of the J24 interface on the last page of this paper.

TREC Interactive Results

The data we collected using the TREC-7 interactive methodology failed to replicate any advantage for the J24 thumbnail displays. Overall, there are no differences between J24 and ZPRISE in the time to perform the tasks, number of documents judged to have relevant aspects, the percent of those document that actually had relevant instances as judged by the NIST experts (Precision) or the percent of judged relevant instances that were found by our searchers (Recall).

The data does suggest that there may be interactions that cannot be analyzed with the Latin Square design used in the interactive methodology. For example looking at Figure 2, it appears that half the participants have better recall with J24 and the other half have better recall with ZPRISE.

However, the participants who have better recall with J24 are those that performed the second half of the tasks with J24 (Tasks 356, 366, 367 and 392), and the participants who have better recall with ZPRISE performed the second half of the task with ZPRISE. One can see from Figure

3 that the second half tasks produced higher recall scores for both J24 and ZPRISE and therefore it appears that the task has a bigger influence on recall than does system differences.

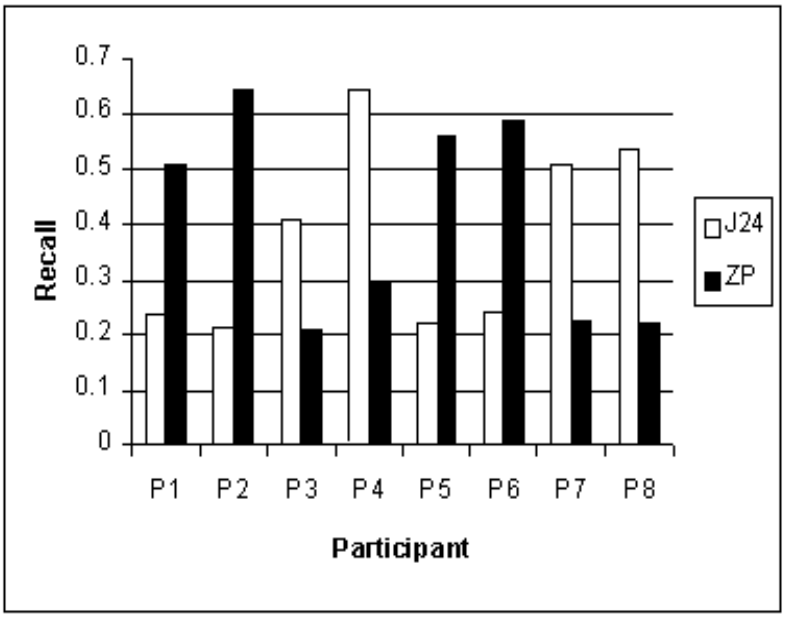


Figure 2. Average Recall for Participants and System

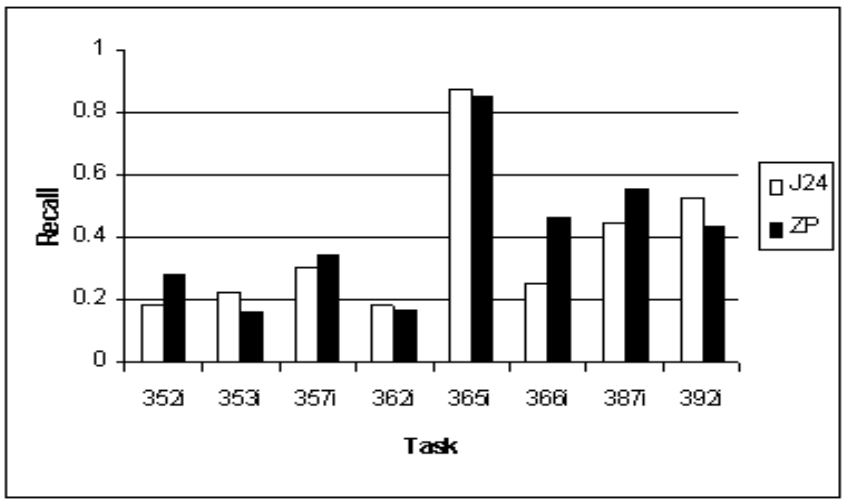


Figure 3. Average Recall for Tasks and System

Unfortunately, it is difficult to use the TREC-7 interactive data to tease apart task variables. Since different participants use different systems on the tasks with similar characteristics, we need more subjects using the same systems on these tasks. It is likely that there are some interactions between subjects, systems and tasks that are confounded with main effects in the Latin Square design.

Ongoing and Future Plans

Investigation of decision making during information retrieval tasks.

There are a number of types of decisions that users must make in an interactive text retrieval situation. From the selection of query terms, to the assessment of effectiveness of those terms in narrowing the search, to the assessment of document relevance either by keyword occurrence, reading a summary, or finding and reading relevant passages. The ability to make these decisions can vary widely from person to person depending on their experience with making these decision and their knowledge of the domain of the search. The number and types of decisions that need to be made vary from task to task and also depend on the domain and how relevant documents are distributed in the data. A decision making strategy that works well for one type of information request may not work for another, and a good searcher can easily shift to more appropriate decision making strategies as the situation changes.All of these decisions can either be aided by a well designed interface or hindered by a poor one.

Suppose that a single interface element, such as a document thumbnail display, aids the user with one decision making strategy, such as being able to quickly reject a non-relevant document because it does not contain a critical keyword. This interface element would therefore be effective only in tasks in which this is a viable and effective strategy and the user knows how to employ it. The interface is unlikely to show a major effect over a range of information retrieval tasks in which the affected decision making strategy plays a minor role in the overall performance of the user.

If we are going to study the effectiveness of user interfaces in interactive text retrieval then we must look at the effectiveness that interface elements have in aiding the decision making process in some very prescribed circumstances that may only sometimes contribute to the overall performance of the user and system.

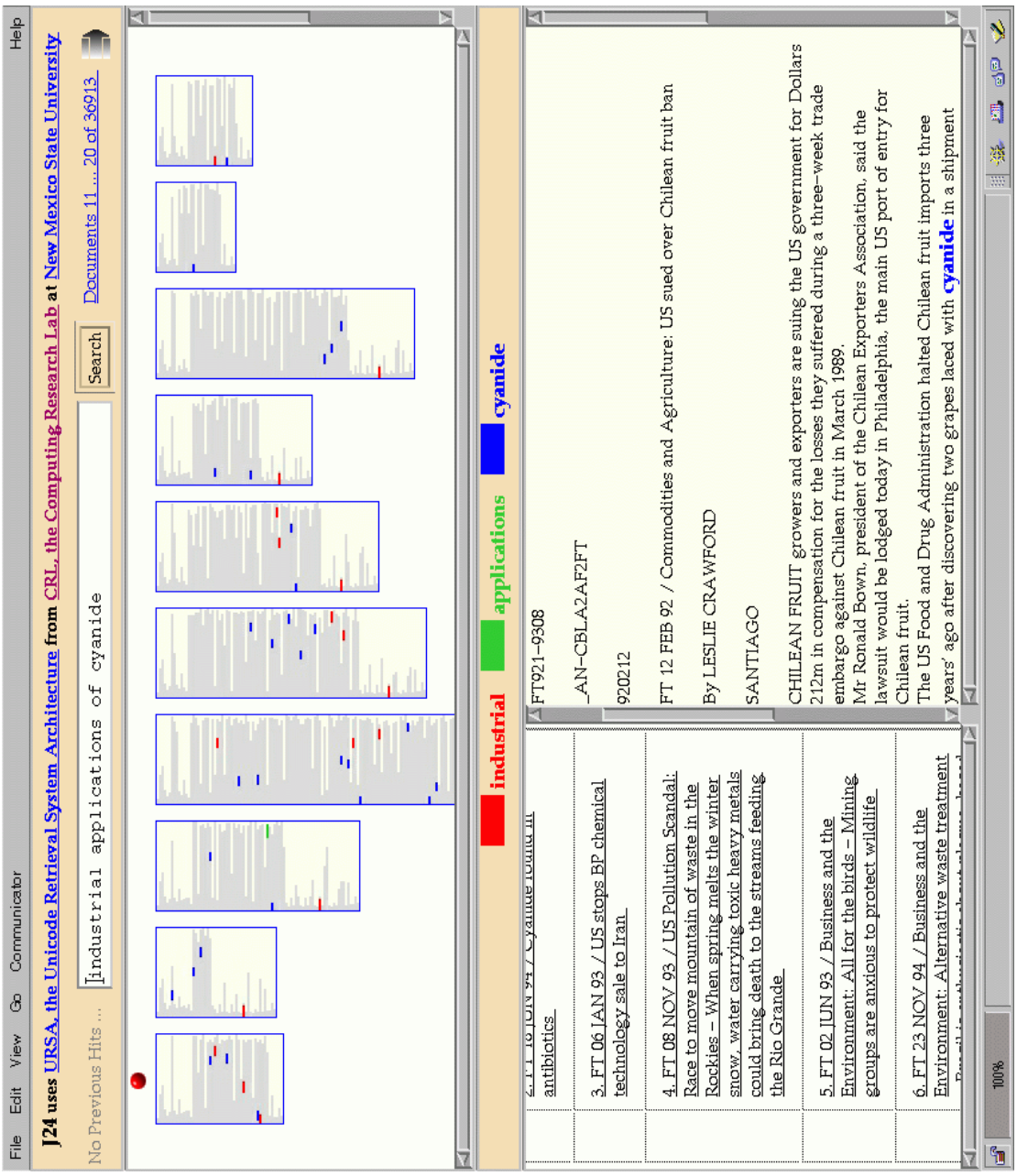
We are currently planning follow-up studies in which we will systematically vary the task characteristics we think are important. For example, instead of using the Interactive-TREC aspectual recall task which requires many stages of decision making, we are now using the relevance decision task used by Kaugars to more close examine the role document thumbnail displays have in users decisions about whether they have to read a document or not. Rather than trying to generalize the interactive evaluation to the set of “all” information retrieval tasks, we are beginning to think that it is better to characterize important interactive variables and their effect on users’ decision making abilities. We need to be able to identify the cognitive bottlenecks that prevent users from finding the relevant documents they seek and thereby discovering how best to build the interactive systems they will want to use

References

Kaugars, Karlis, J. 1998. A Hierarchical, Approach to Detail + Context Views. Unpublished Doctoral Dissertation. New Mexico State University.

Robertson, S.E., Walker, S., Beaulieu, M.M. and Gatford, M. Okapi at TREC-4. In The Fourth Text REtrieval Conference (TREC-4). D.K. Harman, ed. NIST Special Publication 500-236. 1996. PP. 73-96.

Zobel, J. and Moffat, A. 1998, Exploring the Similarity Space. SIGIR Forum, ACM Press. 32:1. PP. 18-34.



The J24 Interface