# From Reading to Retrieval: Freeform Ink Annotations as Queries

*Gene Golovchinsky, Morgan N. Price, and Bill N. Schilit*
FX Palo Alto Laboratory, Inc.
3400 Hillview Ave., Bldg. 4
Palo Alto, CA 94304
+1 650 813-7361

{gene, price, schilit}@pal.xerox.com

## ABSTRACT

User interfaces for digital libraries tend to focus on retrieval: users retrieve documents online, but then print them out and work with them on paper. One reason for printing documents is to annotate them with freeform ink while reading. Annotation can help readers to understand documents and to make them their own. In addition, annotation can reveal readers' interests with respect to a particular document. In particular, it is possible to construct full-text queries based on annotated passages of documents. We describe an experiment that tested the effectiveness of such queries, as compared to relevance feedback query techniques. For a set of TREC topics and documents, queries derived from annotated passages produced significantly better results than queries derived from subjects' judgments of relevance.

## Keywords

Annotation-based queries, freeform digital ink, query expansion, query-mediated browsing, information appliances, information retrieval, relevance feedback, information exploration, digital libraries, user studies, empirical evaluation

## 1. THE PROBLEM

Digital library users tend to use computers and paper together when engaged in interactive information retrieval tasks. Library clients may read piles of search results using "paper copies, even when the material had been delivered electronically" [17], and many information analysts "make hardcopies of source materials, and mark up the pages" (Catherine C. Marshall, personal communication). This "search and print" reality contrasts with the "search and read" ideal held by digital library researchers (e.g., [3]). In general, we believe that users often find relevant documents online, make sense of them by printing them out and by marking them up on paper, and then return to a computer to iterate the search.

These transitions between retrieval online and working on paper are disruptive. Printing out piles of retrieved documents is inherently slow. Once the results are printed, the system typically loses all knowledge of the user's information needs. Furthermore, the system cannot observe people's use of paper to inform subsequent iterations of online retrieval.

Such barriers between retrieval and understanding can also interfere with the iteration inherent in information exploration. For example, when many apparently relevant documents are available, people resort to *information triage*: instead of narrowing their search as they read, they retrieve many documents, print them out in large piles, and then evaluate them on paper [13].

## 2. INFORMATION APPLIANCES[1]

Why do people print documents when they could read them online, avoiding these problematic transitions? One answer is that typical desktop computer interfaces for reading and understanding documents are awkward, and do not support existing work practices such as annotation [18].

We believe that knowledge workers need a new platform for working in digital libraries: *digital library information appliances* that support a wide range of document activities related to reading [24]. We are exploring this concept in the XLibris "active reading machine"[23].

### 2.1 Paper Document Metaphor

XLibris supports reading and annotation by employing the "paper document metaphor" (as opposed to the desktop metaphor) that imitates the physical experience of reading and marking on paper [23]. Both hardware and software support the affordances implied by the metaphor.

XLibris runs on off-the-shelf pen tablet displays (e.g., Figure 1). Devices that users can hold are much more comfortable to read from than displays, such as CRTs, which are fixed in space [7]. XLibris displays a full page at a time and devotes almost the entire display to the page. Readers mark anywhere on the page with a variety of ink and highlighter pens. As on paper, readers can quickly underline, circle, or highlight important passages, or jot down comments in the margin. We believe that the ability to make unstructured, freeform, idiosyncratic ink annotations is crucial for making sense of information.

[1] Jef Raskin coined the term "information appliance" for small systems intended to perform mainly one task (Nielsen, 1988).

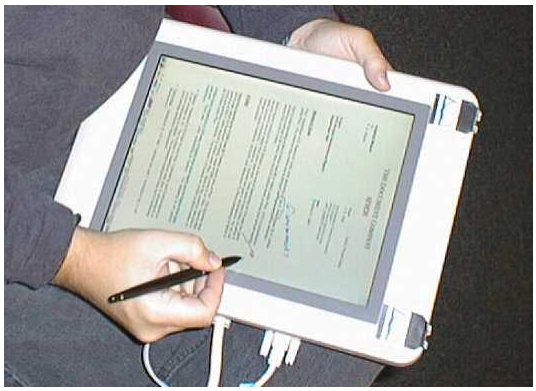

**Figure 1: XLibris on a Mutoh tablet: readers hold the image of a page in their lap and mark on it with a pen.**

XLibris also helps readers to find information—to transition from reading to retrieval—by deriving queries from users' freeform ink annotations [19]. People often use free-form ink annotations to help them understand their reading [1], [12]. In particular, information seekers use annotations to manage printed search results: they mark up relevant passages with highlights and brief marginalia [17], [13]. Thus, freeform ink annotations can reveal which aspects of the underlying documents are of interest, and by observing these marks, XLibris can discover people's interests without requiring additional effort from them.

XLibris interprets readers' freeform ink annotations as selections of underlying text, constructs queries from this text, runs the queries against a full-text database, and presents links to retrieved documents. This is a form of "queryless" information retrieval, where the system hides the query syntax from the user; another example may be found in VOIR [5], [4].

In the following sections, we briefly compare queries derived from annotations with relevance feedback queries, and then describe an experiment based loosely on the TREC interactive track that compared them.

# 3. QUERY EXPANSION

Relevance feedback and annotation-based queries are automatic techniques for query construction which attempt to use information about a user's interests to reduce the effort required to formulate queries, thereby improving retrieval performance. Relevance feedback supports the transition from examining search results back to searching; annotations-as-queries support the transition from reading tasks, such as extracting information from search results, to subsequent searching. Whereas relevance feedback requires users to expend additional effort to communicate their interests to the system, annotations-as-queries impose no such burden because queries may be derived from annotations made for purposes of understanding documents, rather than explicitly for retrieval.

Query expansion driven by relevance feedback may cause the system to select terms that, although statistically representative of the document, do not reflect accurately a user's interest in the document. Queries derived from passages selected by annotations, however, should reflect users' interests more accurately, and passage-level evidence has been shown to be important in document retrieval [21]. Thus we hypothesize that queries derived from annotations should outperform queries derived from explicit relevance judgements.

# 4. EXPERIMENT

We conducted an experiment to test the hypothesis that readers' annotations would capture the relevant aspects of a set of retrieved documents better than binary judgments of relevance on entire documents. We used a within-subjects one-factor experimental design. The control condition used relevance feedback queries based on subjects' relevance judgments, and the experimental condition used queries derived from annotations for that topic. Each subject read the same documents on the same topics (presented in randomized order).

We used residual recall and precision[2] [22], [10] as dependent measures. Although information exploration tasks tend to be precision-oriented [14], we believe that calculating recall is useful not only to reflect the needs of a broader set of search tasks, but also to assess the differences between query types. Measures of interactive search performance, such as time to retrieve the first relevant article, were not applicable because subjects did not search interactively, and did not receive any feedback on their performance.

## 4.1 Task

The experimental task consisted of reading, annotating, and judging six documents related to a particular topic. Subjects were asked to mark up the relevant passages in each document, and they were also required to make a relevance judgment for each document on a ten-point scale. A ten-point scale was chosen over binary judgments to reduce the likelihood of having subjects select no relevant documents for a particular topic (thereby rendering the comparison with automatic relevance feedback meaningless), and to make it easier for subjects to "hedge" their decisions. Subjects were allowed to change their judgments of relevance at any time during work on a particular topic.

Subjects used a version of XLibris specifically designed and instrumented for this experiment, running on the Mutoh 12P pen-tablet display [15]. The experimental system supported reading, page turning, annotation with pens and highlighters, and erasing as described in the "Paper Document Metaphor" section above. The interface also included buttons for moving among documents (bottom right of Figure 2), and buttons corresponding to the levels of relevance (bottom middle of Figure 2).

Prior to performing the experimental task, subjects were given a set of written instructions, and were shown how to use the experimental software. They were allowed to practice writing and annotating on some sample documents until they were comfortable with the interface. Following the training session, subjects were handed a description of the topic of interest (i.e., a

[2] Residual measures exclude previously retrieved documents from calculations of performance. We used residual measures to make comparisons of incremental retrieval typical of relevance feedback situations more meaningful. Rank freezing ([20] cited in [11]) was not used because subjects always read similar numbers of relevant and non-relevant articles for each topic (see the method section, below).

description of relevant documents), and were asked to read it prior to starting the experiment. Each subject performed the task on six topics. They were required to finish each topic before moving on to the next.

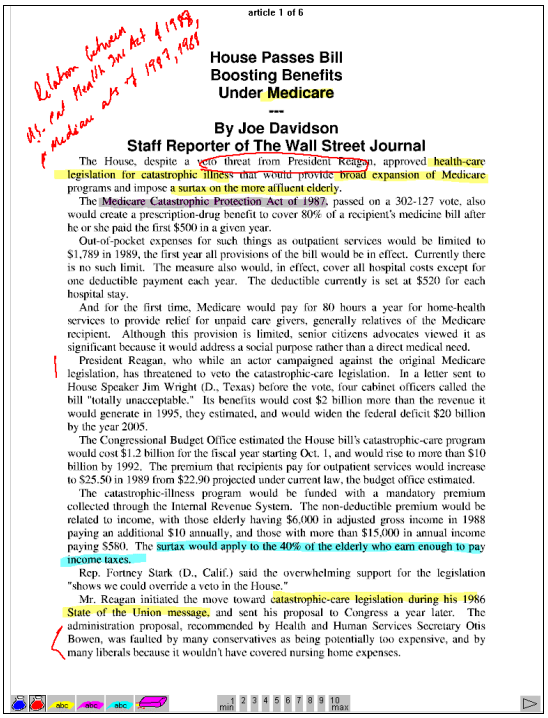


**Figure 2: The experimental interface**

## 4.2 Subjects

Ten volunteers (researchers at FXPAL and at PARC) participated in the experiment. They were not screened for reading ability or for annotation habits, and we did not inquire about their online searching experience as all searches would be done off-line, without any additional input from participants. Subjects were not paid for their participation, and could withdraw from the experiment at any time. Subjects were encouraged to take short breaks between topic sessions.

## 4.3 Method

We evaluated all queries against the 173,600-document collection of The Wall Street Journal articles from the TREC-3 conference [9] with the InQuery search engine [26]. Each TREC-3 search topic consists of a paragraph characterizing relevant documents, and has an associated list of relevant documents as judged by expert raters. The lists of relevant documents allow us to compute recall and precision. Although these measures reflect experts' understanding of the topic, rather than the subjects' understanding, they are still useful for comparing different types of queries.

We selected six topics for this experiment from the 50 topic descriptions from the TREC-3 conference [9], based on the following criteria. We evaluated a query based on each topic's description,[3] and selected topics for which the top ten documents returned by the query contained three to six relevant documents. In addition, we chose two topics with low overall numbers of relevant documents (25 each), two with moderate numbers of relevant documents (57 and 68), and two with high numbers of relevant documents (98 and 135).

For each topic, we selected the top three relevant and non-relevant documents for the subjects to read (we did not examine the documents manually before selecting them). We used this strategy to ensure that relevance feedback algorithms had enough relevant documents to perform reasonably, and to insure that enough non-relevant documents were present for the task to be non-trivial. These original documents were excluded from the residual precision and recall computations used to assess query performance. We formatted each document in Microsoft Word™, using a 12-point Times New Roman font for the body text (see Figure 2 for an example of a resulting page image).

Subjects' relevance judgments were converted into binary judgments of relevance required by the search engine as follows: for each subject and topic, relevance scores were averaged, and only documents with average or above-average scores were considered relevant. (This scheme may bias the comparison in favor of relevance feedback, as half of the documents were in fact relevant. For example, one subject complained that none of the documents presented for a particular topic was relevant, but used two different low ratings for those documents.) The documents judged (more) relevant were used to create a new query using InQuery's `#reldoc()` operator.

Subjects' annotations were recorded and subsequently converted to weighted-sum queries (InQuery's `#wsum()` operator) as described in the following section.

### *4.3.1 Deriving queries from annotations*

XLibris derives queries from users' annotations in three steps: it interprets free-form ink strokes as selections of text, groups strokes and accompanying word selections, and assigns weights to selected terms.

*Selection.* XLibris interprets annotations by using simple heuristics to classify each stroke and to determine which text is selected [23]. Circled phrases, underlines, and highlights select specific words, while broader circles and margin bars select nearby sentences. Other marks in the margin select no text because they may be handwritten comments that can stray far away from the relevant text, especially if the document has small margins. Other marks on the text select the words they touch.

XLibris selects words in three ways—as focus, context, or passage—based on the class of the stroke. Circled phrases, underlines, and highlights select specific words and phrases, so they form "focus" selections. Other words in those sentences form "context" selections.[4] Finally, broadly circled chunks of text and margin bars form "passage" selections.

---

[3] The text of the "description" field of each topic was used as an unstructured query, with the first three words of this text removed. The first three words of these topic descriptions are phrases such as "Document will contain," "Document will discuss," etc.

[4] Including context words that surrounded explicitly selected words helped disambiguate term sense [4].

*Ink Grouping.* Given groups of ink strokes, XLibris combines word selections by retaining the most explicit selection for each instance of a word on a page. By retaining only the most explicit selection for each word instance, XLibris ignores situations where several strokes are used to annotate a phrase or passage. By retaining selections across word instances, XLibris accumulates information about how often users select each word. For example, if a word was highlighted and then a margin bar was used to select the passage, XLibris would assign its weight based on the more specific highlight annotation. If that same word appeared elsewhere in the document and was selected by another annotation, XLibris would use the sum of the weights from the two selections.

*Weights.* XLibris constructs weighted queries out of these lists of selected words. Weighted queries allow XLibris to specify term weights explicitly, giving *a priori* importance to terms based on selection counts; the search engine may then use the statistical distribution of terms among documents in the collection to adjust these weights. The weight for each term is the sum of the selection weight over all selections for that term. Focus selections receive four points, passage selections receive two points, and context selections receive one point.

## 4.4 Results

Three dependent measures were used to compare performance between ink-based (weighted-sum) and relevance feedback queries. Differences in performance between the two types of queries were calculated for residual precision at ten documents, and for residual recall and precision at 100 documents. Positive values represent better performance for annotation-based queries.

Precision at ten documents is intended to represent the number of documents a user might examine in a typical browsing session to get a quick impression of the quality of the search results. Recall at ten documents is not a meaningful measure for large collections that may contain large numbers of documents relevant to any particular topic. Precision and recall at 100 documents are intended to represent the limits of what a user may reasonably be expected to examine interactively.

The experiment was a repeated-measures within-subjects design. Although the experimental design lent itself to a within-subjects multivariate analysis of variance (MANOVA), the data deviated significantly from the normal distribution, rendering the MANOVA inappropriate. Therefore, the paired Wilcoxon signed ranks test (see [10] for a discussion) was performed separately on three dependent variables: residual precision at ten documents, residual precision at 100 documents, and residual recall at 100 documents. Query type (markup vs. relevance feedback) was the independent variable, and topic (six trials) was used as a repeated measure.

A conservative alpha level (0.01) was used to compensate for potential inflation of significance due to multiple tests. Significant differences ($p<0.001$) between markup (M) and relevance feedback (RF) queries were found for all three dependent variables, as shown in Table 1. These results indicate that in the majority of cases higher residual precision and recall were observed for queries derived from ink annotations than from relevance feedback queries based on users' judgments of document relevance, allowing us to reject the null hypothesis. Figure 3 shows the distribution of scores for each topic; means above the top hashed line (difference of 0.10 or more) are "material" [25], [11].

| | Prec @ 10 | | Prec @ 100 | | Rec @ 100 | |
|---|---|---|---|---|---|---|
| | N | Z | N | Z | N | Z |
| M > RF | 36 | | 51 | | 51 | |
| M < RF | 15 | 4.71 | 8 | 5.73 | 8 | 5.23 |
| M = RF | 9 | | 1 | | 1 | |

**Table 1. Results of the paired Wilcoxon signed ranks test. In column 1, M represents performance of Markup queries and RF represents performance of Relevance Feedback queries. The differences for all dependent variables are significant at $p < 0.001$.**

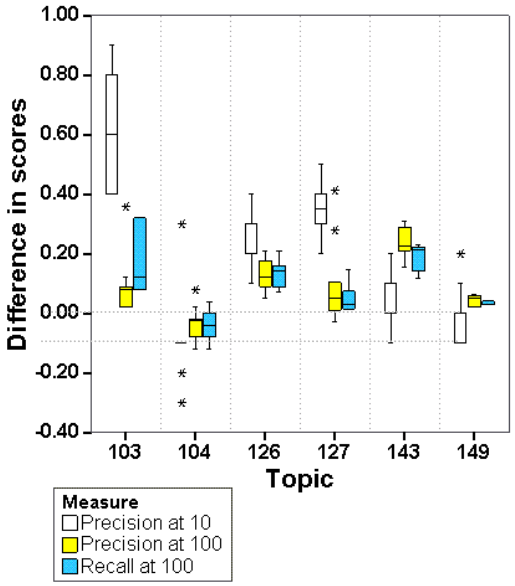


**Figure 3. Distribution of performance differences by topic.**

The results of this experiment indicate that queries derived from users' ink marks on documents can produce retrieval performance better than standard relevance feedback queries. These results are likely due to the differences between the statistical nature of relevance feedback, and the semantic nature of users' marks. While relevance feedback tends to select terms that discriminate well among documents, these terms do not always capture the aspects of documents that are interesting to users. When users select relevant passages manually, the system may be able to capture the latent information needs more accurately.

## 4.5 Exploratory Analysis

Although the results confirmed our research hypothesis, there were large differences in relative effectiveness among topics. We ran a number of exploratory analyses to gain insight into how users' behavior could account for the observed differences.

### *4.5.1 Differences among topics*

Between-topic variability is a well-known phenomenon in TREC experiments. In our study, some subjects commented that they had problems deciding on relevant documents for topic 104 because it required documents related to a particular date (the 1988 Medicare act), and that information was not always available to them. The poor performance of both methods for

topic 149 (Figure 4) suggests that the documents shown to subjects were not representative of other relevant documents.

Although we designed the experiment to compare two types of queries in a controlled manner, we also allowed subjects to interact with the system in whatever way seemed natural to them. Thus we can analyze their patterns of behavior to try to account for differences in performance seen in the hypothesis-testing analysis [6].

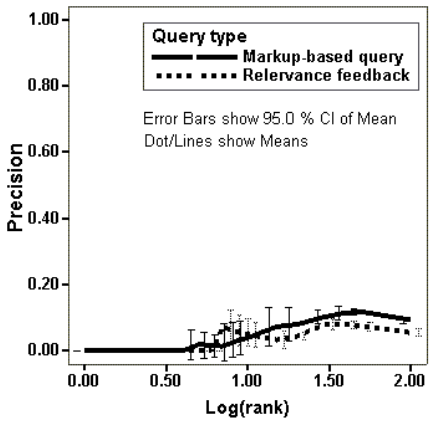


**Figure 4. Performance for topic 149.**

Our subjects performed two main activities in XLibris: they annotated documents and made relevance judgments. Thus, we were interested in consistency of annotations and of judgments of relevance, and in how these factors related to performance on the different topics.

*Annotations.* We recorded the sentences annotated by each subject, and measured the pairwise overlap for each topic. A one-way ANOVA on sentence overlap (normalized by the geometric means) with topic as the independent variable revealed significant differences among topics ($F[5,264]=13.4$, $P<<0.01$); post-hoc pairwise comparisons of means indicated that topics 104 and 149—topics with less overlap—were consistently grouped together (showed no reliable differences in means), as were topics 103, 126, 127 and 143. The partitioning was not perfect, however: the means for topic 126 and 149 were not reliably different. It is worth noting that our technique failed to outperform relevance feedback for topics 104 and 149.

*Relevance judgments*. We also explored the possibility that the degree to which subjects agreed on which documents were relevant would be reflected in their performance. Subjects made six relevance judgments for each topic, and we measured pairwise correlations among their relevance judgments. We then performed a one-way ANOVA on these scores, with topic as the independent variable. ANOVA indicated the presence of significant differences among topics ($F[5,264]=19.4$, $p<<0.01$); post-hoc comparisons of means indicated that topics 103, 126 and 127 were reliably different from topics 104, 143 and 149. The average correlation for the former group was 0.75, and only 0.36 for the latter.

From Figure 3, we can see that for topics with high correlations between relevance judgments (103, 126, and 127), subjects had obtained better performance overall with queries based on annotations, and the differences were larger at ten rather than at 100 documents. Recall and precision measured at 100 documents may be less sensitive to these judgments because of the large numbers of terms used in the queries.

Although these analyses are not conclusive, they do yield similar results. This suggests that the effects described here are probably real. It may be interesting to test these ideas further by analyzing relevance judgment data from other TREC experiments.

### *4.5.2 Query construction*

How many terms should be included in the queries derived from links? Harman [8] observes that after about 20 to 30 terms, marginal gains in precision tend to decrease with added terms, while the cost of computing these queries tends to increase linearly with the number of terms. In our prototype, we have not filtered the terms selected by users (except stop words, which are removed by the search engine). Thus queries of 200-300 terms were not unusual in this study.

We examined the effect that filtering the query terms would have on overall performance. We created several subsets of query terms derived from each subject's annotations. Terms were ranked by the weight*idf scores, and residual recall and precision were computed for top-ranked subsets of 10, 20, 30, 40, 50, 75, 100, 150, and 200 terms. Performance differed very little after 50 terms, as shown in Figure 5 for residual recall at 100 documents.

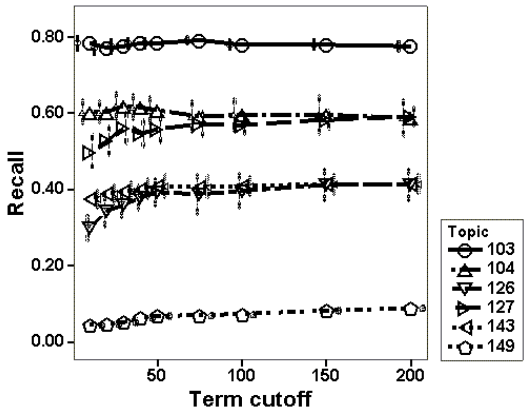


**Figure 5. Recall at 100 documents as a function of query size by topic.**

A possible advantage of the "annotations as queries" technique is that readers specify their interest at the word level rather at the more coarse paragraph or document levels. This experiment showed that word-level selections out-perform document-level judgments of relevance. Do word-level selections also out-perform paragraph-level selections? We compared word-level queries with queries computed from paragraphs (implicitly) selected by the annotations, and found no statistically-significant differences in performance in terms of residual recall and precision. Thus we believe that paragraphs identified by users' annotations can be useful sources of terms for relevance feedback. This conclusion, however, must be tempered by the observation that most newspaper paragraphs are only one to three sentences long.

## 5. FUTURE WORK

This experiment has shown the potential of free-form ink as a mechanism for eliciting relevance feedback. Is it also appropriate for interactive browsing on specific topics (e.g.,

margin links in [19])? The instructions given to participants in this experiment caused them to make many annotations while reading. Furthermore, their task was to read entire documents, rather than to browse or to search for information. In contrast, interactive browsing will involve a smaller number of annotations from which queries must be derived. In these situations, we expect query performance to be sensitive to the details of query construction. Will some weighting schemes produce better performance? Should explicitly marked phrases be weighted differently from annotated sentences or passages? How should context terms surrounding those phrases be handled? Finally, can individual differences in annotation strategies be harnessed to form better queries?

Although it is possible to expand this experiment to include more topics (to understand topic effects in more detail), we believe that evaluating these techniques in naturalistic settings will ultimately yield more useful results. A remaining challenge is to identify appropriate measures and techniques to evaluate complex information exploration interfaces in real-world settings where relevance judgments are not available.

Preliminary exploration suggests that the degree to which subjects understand a query topic can predict the relative performance of markup-based queries. If confirmed (perhaps through analysis of data collected in TREC interactive tracks), this topic analysis methodology may be useful for designing and evaluating future interactive information retrieval experiments whose dependent measures are based on experts' judgments of relevance.

## 6. CONCLUSIONS

We have described a novel interface for eliciting users' interests in information exploration tasks. This technique allows users to specify terms, phrases, or passages of interest by marking up documents with freeform digital ink. We performed an experiment that compared the effectiveness of queries derived from ink annotations to relevance feedback queries. Queries derived from users' ink marks produced significantly better results than relevance feedback queries for residual precision at ten documents and for residual recall and precision at 100 documents.

This work is one step toward a better understanding of the role user interfaces play in information exploration tasks. Information retrieval is becoming an increasingly common computer-mediated activity, but rarely is it engaged in for its own sake. Typically, retrieval is situated in other activities (e.g., reading, analysis, writing, etc.). The work described here is an attempt to make the transitions between retrieval and the task that motivates it more seamless, to situate retrieval in a broader human-computer interaction. We believe that systems for situated retrieval can take advantage of users' existing work practices to reduce the cognitive overhead of information retrieval while leveraging the increasing power and flexibility of search engines.

## 7. ACKNOWLEDGMENTS

We thank our participants for their time and effort, Jim Baker and Joe Sullivan for supporting this research, and the PDR group at PARC for many interesting discussions.

## 8. REFERENCES

[1] Adler, A., Gujar, A., Harrison, B.L., O'Hara, K., and Sellen, A. (1998) A Diary Study of Work-Related Reading: Design Implications for Digital Reading Devices. In *Proceedings of CHI98* (Los Angeles, CA, April 18-23), ACM Press, pp. 241-248.

[2] Callan, J.P. (1994). "Passage-level evidence in document retrieval." In *Proceedings of SIGIR '94*, Dublin, Ireland, ACM Press, pp. 302-310.

[3] Entlich, R., Olsen, J., Garson, L., Lesk, M., Normore, L., and Weibel, S. (1997) Making a digital library: the contents of the CORE project, ACM TOIS, 15(2), pp. 103-123.

[4] Golovchinsky, G. (1997) What the Query Told The Link: The Integration of Hypertext and Information Retrieval. In *Proceedings of Hypertext '97*(Southampton, UK, April 8-11), ACM Press. pp. 67-74.

[5] Golovchinsky, G. and Chignell, M.H. (1997) The newspaper as an information exploration metaphor. *Information Processing & Management*, 33 (5), pp. 663-683.

[6] Golovchinsky, G., Chignell, M.H., and Charoenkitkarn, N. (1997) Formal experiments in causal attire: Case studies in information exploration. *New Review of Hypermedia and Multimedia*. Vol. 3. (1997), Taylor Graham. pp. 123-158.

[7] Gujar, A., Harrison, B.L., and Fishkin, K.P. (1998) A Comparative Empirical Evaluation of Display Technologies for Reading. In *Proceedings of HFES '98* (Chicago, IL, October 5-9). pp. 527-531.

[8] Harman, D. (1992) Relevance Feedback Revisited. In *Proceedings of SIGIR '92* (Copenhagen, Denmark June 1992), ACM Press. pp. 1-10.

[9] Harman, D. (1995) Overview of the Third Text REtrieval Conference (TREC-3). D.K. Harman, ed. NIST Special Publication 500-225, Gaithersburg, MD, pp. 1-19.

[10] Hull, D. (1993) Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proceedings of ACM SIGIR '93* (Pittsburgh, PA, June 1993), ACM Press, pp. 329-338.

[11] Keen, E.M. (1992) Presenting results of experimental retrieval comparisons. *Information Processing & Management*, 28 (4), pp. 491-502.

[12] Marshall, C.C. (1997) Annotation: from paper books to the digital library. In *Proceedings of the 2nd ACM international conference on Digital Libraries*, ACM Press. pp. 131-140.

[13] Marshall, C.C. and Shipman, F.M. (1997) Spatial Hypertext and the Practice of Information Triage. In *Proceedings of Hypertext '97*, Southampton UK, ACM Press, pp. 124-133.

[14] Meadow, C.T. (1992) *Text Information Retrieval Systems*, Academic Press.

[15] Mutoh (1997) Mutoh America Inc. Available at www.mutoh.com

[16] Nielsen, J (1988). Hypertext '87 Trip Report. *ACM SIGCHI Bulletin* 19**,** 4 (April 1988), pp. 27-35.

[17] O'Day, V.L. and Jeffries, R. (1993) Orienteering in an Information Landscape: How Information Seekers Get From Here to There. In *Proceedings of INTERCHI '93*, ACM Press, pp. 438-445.

[18] O'Hara, K. and Sellen, A. (1997) A Comparison of Reading Paper and On-Line Documents. In *Proceedings of CHI97* (Atlanta, GA, March 1997), ACM Press, pp. 335-342.

[19] Price, M.N., Golovchinsky, G., and Schilit, B.N. (1998) Linking by Inking: Trailblazing in a Paper-like Hypertext. In *Proceedings of Hypertext '98* (Pittsburgh, PA, June 22-24, 1998), ACM Press. pp. 30-39.

[20] Salton, G. (1970) Evaluation problems in interactive information retrieval. *Information Storage & Retrieval*, 6(1), pp. 29-44.

[21] Salton, G., Allan, J. and Buckley, C. (1993) Approaches to passage retrieval in full-text information systems. In *Proceedings of SIGIR '93*, (Pittsburgh, PA, June 1993), ACM Press. pp 49-58.

[22] Salton, G. and Buckley, C. (1990) Improving retrieval performance by relevance feedback, *Journal of the American Society for Information Science*, 41(6), pp. 288-297.

[23] Schilit, B.N., Golovchinsky, G., and Price, M.N. (1998a) Beyond Paper: Supporting Active Reading with Free-form digital Ink Annotations. In *Proceedings of CHI98* (Los Angeles, CA, April 19-26), ACM Press, pp. 149-156..

[24] Schilit, B.N., Price, M.N., and Golovchinsky, G. (1998b) Digital Library Information Appliances. In *Proceedings of Digital Libraries '98* (Pittsburgh, PA, June 24-26), ACM Press. pp. 217-226.

[25] Sparck-Jones, K. (1974) Automatic indexing, *Journal of Documentation*, 30(4), pp. 393-432.

[26] Turtle, H. and Croft, W.B. (1990) Inference Networks for Document Retrieval. In *Proceedings of SIGIR '90*, ACM Press, pp. 1-24.