*Research Paper* ■

# Ad Hoc Classification of Radiology Reports

DAVID B. ARONOW, MD, MPH, FENG FANGFANG, MD, W. BRUCE CROFT, PHD

**Abstract** **Objective:** The task of ad hoc classification is to automatically place a large number of text documents into nonstandard categories that are determined by a user. The authors examine the use of statistical information retrieval techniques for ad hoc classification of dictated mammography reports.

**Design:** The authors' approach is the automated generation of a classification algorithm based on positive and negative evidence that is extracted from relevance-judged documents. Test documents are sorted into three conceptual bins: membership in a user-defined class, exclusion from the user-defined class, and uncertain. Documentation of absent findings through the use of negation and conjunction, a hallmark of interpretive test results, is managed by expansion and tokenization of these phrases.

**Measurements:** Classifier performance is evaluated using a single measure, the *F* measure, which provides a weighted combination of recall and precision of document sorting into true positive and true negative bins.

**Results:** Single terms are the most effective text feature in the classification profile, with some improvement provided by the addition of pairs of unordered terms to the profile. Excessive iterations of automated classifier enhancement degrade performance because of overtraining. Performance is best when the proportions of relevant and irrelevant documents in the training collection are close to equal. Special handling of negation phrases improves performance when the number of terms in the classification profile is limited.

**Conclusions:** The ad hoc classifier system is a promising approach for the classification of large collections of medical documents. NegExpander can distinguish positive evidence from negative evidence when the negative evidence plays an important role in the classification.

■ **JAMIA.** 1999;6:393–411.

Classification or categorization systems assign class labels to documents. In contrast, information retrieval (IR) systems typically rank documents by their likelihood of belonging to a relevant class rather than categorizing them by whether they are or are not members of the relevant class. In the medical domain, more interest is often shown in the classification of clinical documents than in their ranking, because of the need to identify the status of every document (and the patients and medical events they represent) as opposed to locating a subset of the best documents in a collection.

Ad hoc classification is an approach taken when a user needs to sort a large number of documents into nonstandard categories. The classification is typically conducted only a limited number of times, since no long-standing information need is being addressed.

In today's health care system, the performance of hospitals, clinics, and providers is constantly being reviewed for both the quality and the cost of care. One

Affiliation of the authors: University of Massachusetts, Amherst, Massachusetts.

This work was supported in part by cooperative agreement EEC.9209623 from the National Science Foundation, Library of Congress, and Department of Commerce, and by grant N66001-94-D-6054 from Naval Research and Development.

Correspondence: David B. Aronow, MD, MPH, P.O. Box 9657, North Amherst, MA 01059. Reprints: Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, MA 01003. e-mail: ⟨david@aronow.com⟩.

Received for publication: 1/7/99; accepted for publication: 4/21/99.

of the major tools for this performance assessment is the review of patient medical records, to study correlations between treatment variations and outcomes, for example. We address medical document review as a classification task in this paper.

The text in clinical records is obtained from a medical provider's dictated notes or by direct data entry by the provider. In the case of the radiology reports used in our experiments, the text is a transcription of dictated interpretations of mammograms. This text is sufficiently different from typical IR test beds, such as news articles, that different indexing techniques may be needed. In particular, we look at techniques for indexing the frequent, significant negations in this text, such as "no evidence of . . ." and "no suspicious. . . ."

The product of medical record review for a large population of patients is generally statistical analysis of groups within the population, who are categorized according to the presence, absence, or value of specified clinical attributes. Since the records must be reviewed in some depth by the health care providers to ensure the accuracy of their categorization, the aim of an ad hoc medical document classification system is to reduce the manual effort required by reducing the number of records that the providers must review in full depth. The problem, then, is to help the person reviewing the records to define classes of interest and rapidly identify the records that require manual review.

Thus, we seek a classification system that specifies whether a record is definitely a member of a class, definitely not a member, or possibly a member. Then, only the records that are possibly members must be manually reviewed by the health care provider. The records that are definitely members of the class and those that are definitely not members of the class do not require manual review, as they may be tallied directly, without further review for the purpose of classification. The technique used for classification is a variation of relevance feedback, with enhancements for phrase indexing, negation indexing, and classification into three "bins."

## Background

Our approach in ad hoc classification is to use advanced information retrieval techniques to generate a profile that is able to categorize text documents. The profile is generated through a training process with the relevance feedback functionality of the Inquery retrieval system and a set of training documents and is stored as a query file consisting of weighted query concepts.

### Inquery

Inquery[1] is a full-text advanced information retrieval system developed by the Center for Intelligent Information Retrieval of the University of Massachusetts at Amherst. Using a probabilistic methodology, Inquery ranks retrieved documents according to their likelihood of being relevant to the information need represented by a query.[2]

Inquery is based on Bayesian inference networks,[3,4] which are directed acyclic graphs (DAG) in which nodes represent propositional variables and arcs represent dependencies. Leaf nodes typically represent propositions whose values can be determined by observations, while other nodes typically represent propositions whose values must be determined by inference from the values of the nodes on which they depend. A notable characteristic of Bayesian networks is that the certainty or probability of dependencies can be represented by weights on arcs.

The Bayesian network used by Inquery is a document retrieval inference network,[5,6] which consists of two component networks—one for documents and one for queries. The document network represents a set of documents with their term representation nodes at varying levels of abstraction, while the query network represents a need for information via query concept representation nodes. Each document node represents one of a variety of document elements (a word, a phrase, word proximity, proper names, dates, etc.) that is part of the proposition that a document satisfies a query concept. If a document term matches a query concept, then a belief value of the proposition, or a weight for the arc between the document term node and the query concept node, is calculated as follows:

$$w_{t,d} = 0.4 + 0.6 \times \frac{tf_{t,d}}{tf_{i,d} + 0.5 + 1.5 \dfrac{len_d}{avgdoclen}} \times \frac{\log \dfrac{N + 0.5}{docf_t}}{\log N + 1}$$

where $tf_{t,d}$ is the number of times term $t$ occurs in document $d$, $len_d$ is the length (in words) of the document $d$, $avgdoclen$ is the average length (in words) of documents in the collection, $docf_t$ is the number of documents containing term $t$, and $N$ is the number of documents in the collection. Inquery ranks the documents by their sum of the belief values, $w_{t,d}$, from high to low. This document ranking—i.e., retrieved documents—is the response to the query.

*DOB: xxxxxxx
MARITAL STATUS:
.RACE:
CLINICAL HISTORY: 42-year-old with history of fibrocystic changes, rt breast mobile mass upper outer quadrant.
BILATERAL FILM SCREEN MAMMOGRAM:
FINDINGS: Prior mammograms are not available for comparison. The breasts are heterogeneously dense. There is extremely dense fibrous tissue in the upper outer quadrants of both breasts. This lowers the sensitivity of mammography. B.B. was placed in the region of palpable abnormality and demonstrates dense breast tissue in this region. An occasional benign-appearing calcification is present in both breasts. No discrete mass, suspicious calcification or other secondary sign of malignancy is demonstrated.
IMPRESSION: No mammographic evidence of malignancy. Recommendation: comparison with patient's previous mammograms is recommended. The decision to biopsy any palpable abnormality should be based on clinical grounds.
D:xx-xxx-xx x:x-xxx-xx
xxx:xx
.EXAM DATE/TIME: xxxxxxx.xxx
.STANDARD TEXT USED:
.REPORT STATUS: VERIFIED
.EXAM STATUS: COMPLETE
.PREGNANT?: UNKNOWN
.REASON FOR EXAM: NO BRIEF COMMENT
42 yr old G0P0 with h/o FCBD. Approx 3-4cm upper outer quadrant R mobile breast mass. Routine retirement physical.
.PATIENT CATEGORY: OUTPATIENT
.REASON FOR ADDITIONAL VIEWS
1-90 LATERAL 1-SPOT
.ORDER COMMENT:
.CLINICAL IMPRESSION:
.CLINICAL IMPRESSION (XTND) 42 yr old G0P0 with h/o FCBD. Approx 3-4cm upper outer quadrant R mobile breast mass. Routine retirement physical.
.CLINICAL COMMENT:
.RADIOLOGY PROCEDURE: MAMMOGRAM, BILAT (SCREENING)(ACTIVE DUTY ONLY)

**Figure 1** Typical original mammography report. Person identifiers have been replaced by x's.

### Relevance Feedback

Relevance feedback improves the performance of a query by modifying the query on the basis of the user's reaction to retrieved documents. The user's judgments of relevance or irrelevance of some retrieved documents are used to find closely associated words and phrases that are added to the query, along with modification of the weighting of the query terms.

Any indexed term that occurs in a relevant document is a candidate for inclusion in the query profile. The candidate terms are first ordered by *rtf*, the number of times the term occurs in the relevant documents. The top 500 terms in that ranking are re-ranked according to a Rocchio formula[7,8]:

$$Rocchio = \alpha w_{query} + \beta w_{rel} - \gamma w_{irr} \qquad (1)$$

where $\alpha$, $\beta$, and $\gamma$ are experimental Rocchio parameters, and $w_x$ is the weight of the term in the query, relevant documents, and irrelevant documents. The weight of term $t$ in the relevant set is calculated as follows, using Inquery's belief function $w_{t,d}$[9]:

$$w_{relset}(t) = \frac{1}{|relset| \cdot \sum_{d \in relset} w_{t,d}}$$

where *relset* is the set of relevant documents and $|relset|$ is the number of relevant documents.

### Mammography Data

The test bed for this research consisted of screening mammogram reports from U.S. naval medical centers. As shown in Figure 1, the reports consist of several structured fields of numeric or controlled vocabulary values for person, time, and study identifiers as well as a number of unstructured text fields for clinical history, findings, and impressions.

Working in the medical domain in general, and with radiology reports in particular, presents several challenges to information systems developers. The first challenge, found throughout the medical domain, is that of unpredictable data quality. Medical data are generally not collected with an intention of extensive electronic manipulation, and unstructured clinical text data, in particular, are expected to be accessed via the coded and numeric data fields associated with them, rather than by the content of the text data itself.

In theory, every aspect of every health care activity could be documented in coded form. In practice, however, this is neither feasible nor always desirable. Coded or structured data often serve well to answer the health care questions of who, when, where, and what but often can not reveal the why of health care practice. The knowledge essential for understanding the rationale of health care decision making is usually embedded in unstructured text. Advanced IR and natural language understanding systems, rather than conventional database management systems, must be developed for this task.

### Related Work

Previous ad hoc classification work in this center concerned electronic medical record encounter notes. Also based on Inquery, the primary application was the automated identification of exacerbations of asthma from transcribed provider notes, both handwritten and dictated. This test bed presented different data quality and content challenges from radiology reports.[10–13]

Other centers have experimented with ad hoc classification of radiology reports and encounter notes. Hripcsak et al.[14] researched the detection of specific conditions in chest radiograms to serve as triggers for a clinical event monitor. Jain et al.[15,16] studied the experimental identification of suspected tuberculosis from chest radiograms and findings suspicious for breast cancer from mammograms. This work uses the MedLEE natural language processing system to extract facts from dictated unstructured text, permitting the reports to be classified according to specific clinical findings. De Estrada et al.[17] reported on the use of stereotypic phrases of known normal findings to screen for abnormal physical findings in COSTAR encounter notes.

Additional work on automated categorization has focused on assigning class labels from a predefined, standardized vocabulary of classes, such as MeSH terms, procedure codes, and diagnosis codes. Cooper and Miller[18] compared lexical and statistical methods of indexing unstructured medical record text to produce MeSH terms for MEDLINE searching. The classification categories are the MeSH terms themselves, and the statistical analysis is based on co-occurrence frequencies between MeSH terms and consecutive work phrases found in the documents. Yang and Chute[19] used more intensive statistical analysis, the nearest-neighbor approach, to categorize MEDLINE articles and surgical procedures. Larkey and Croft[20] used three statistical methods (nearest neighbor, Bayesian independence and relevance feedback) alone and in combinations to assign ICD-9-CM diagnosis codes to dictated discharge summaries. Hersh et al.[21] used other statistical methods, frequency of word occurrence, and principal component analysis, to predict procedure code assignment in emergency room dictations.

Automated classification of clinical documents, such as the assignment of codes or controlled vocabulary to unstructured text, is an area of vigorous medical informatics experimentation. A large variety of approaches, based on both natural language processing and information retrieval and both lexical and statistically intensive, are being actively researched.

## Methods

### System Overview

Our approach is the application of robust statistical IR technologies to ad hoc categorization questions. We want to identify both relevant and irrelevant documents, to provide positive and negative evidence to sort documents into three conceptual bins. Those documents for which there is insufficient evidence to qualify for inclusion in the user-defined category (positive bin) or exclusion from it (negative bin) are assigned to an uncertain bin. The technology is generic and can be readily ported across applications without intensive application-specific knowledge engineering or manual coding efforts.

The ad hoc classifier approaches document classification as a knowledge acquisition problem, constructing customized algorithms from captured domain expertise for each new user question. Operations can be separated into two phases—the training and testing phase and the run-time phase. It builds a classifier that is trained on sample positive and negative documents. The tools for extracting the evidence are completely generic and can be used to construct any number of document classification algorithms. For a given classification question, the user constructs and evaluates a classifier (training and testing phase), then ap-

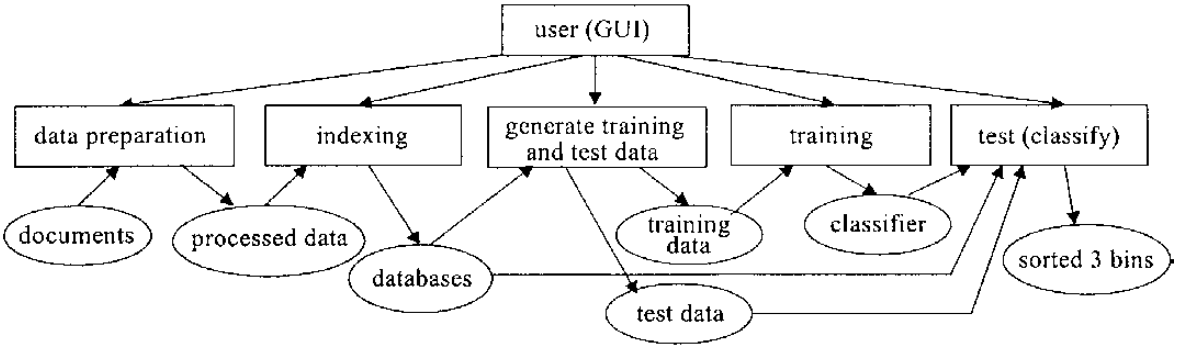



**Figure 2** Overview of the ad hoc classifier system.

plies the classifier to documents (run-time phase). Figure 2 presents an overview of the classifier.

The data preparation component normalizes the documents of a collection with standard vocabularies and abbreviations and tokenizes noun phrases (including negations) to allow their handling as distinct concepts. Most negation words (e.g., "no," "not," "nor," "neither," "without") are considered stop words by Inquery and are not indexed when a collection is built. For effectively distinguishing the relevant positive features from relevant negative features, a program called NegExpander expands negative words across conjunctive phrases and tokenizes them as individual negative terms. When data preparation is complete, the collection is indexed by Inquery to build databases for both the document file and the concept file.

To facilitate the generation of training and test data, a program called RelHelper presents a table of concepts found within the collection, from which the user identifies a group of concepts that are relevant to the task. The selected concepts are used as seeds to train a classifier, with which the user then identifies sets of retrieved documents that are clearly relevant and irrelevant to the classification question. These documents are extracted from the collection and are used to build the training and test collections. Inquery's relevance feedback module[22] uses the training collection to construct the classification algorithm.

When classified, documents are assigned a likelihood-of-relevance score and thresholds are set to specify which scores go into the positive bin, the negative bin, and the uncertain bin. To test a classifier, the query profile is run on the test collection and an evaluation program reports true categorization rates, false categorization rates, and the number of documents in every bin. To improve the profile, the user can either reset the parameters of the relevance feedback or run an enhancement program to automatically adjust the weights of query terms. In the next sections we discuss in greater detail the techniques used in the ad hoc classifier.

### Data Preparation

Despite the fact the naval medical centers supplying data use identical clinical information systems (CHCS) and used the same software utilities to extract their data, there were widespread data quality problems within and among sites. These problems concerned the replacement or expansion of CHCS field names with local institution field names, replacement or expansion of CHCS controlled vocabulary values with local terminologies, unpredictable use of upper case letters in some blocks of text, null fields, and duplicate records. Thus, an extensive effort was required to analyze, normalize, and structure the data before they could be used as a system development and research test bed.

In addition to preprocessing of the data to normalize the report structure, the general medical data challenge was largely addressed with the expansion of a small number of common medical abbreviations. This processing is primarily applicable to the "reason for exam" section of the reports, which is typically a transcription of a telegraphic, handwritten note from the patient's primary physician. This expansion allows more of the text to be handled appropriately by NegExpander, as described below. Examples of the abbreviation expansion are YO to "year old," H/O to "history of," CA to "cancer," and S/P to "status post."

Two challenges are particular to radiology and similar test reports in which expert observations are documented—absent findings and modifier permutations. These document types include extensive documentation of absent findings through the use of negation and conjunction. This is appropriate in interpretive results reporting to specifically document the fact that

untoward findings were looked for and not found. Our approach to this challenge is discussed in detail below. Finally, permutations of limited sets of modifier words are widely used in standard descriptive phrases. These were grouped according to their head nouns and bundled into single noun phrase concepts so that the different arrangements of modifiers are not interpreted as different concepts.

The user's classification interest creates the classification categories—what is relevant and what is irrelevant. The question can be defined narrowly or widely, according to the information need. Defining relevance narrowly leads to a broad definition of irrelevance.

In our classification task, we defined relevance broadly. We sought to classify those screening mammogram reports that include findings of calcification according to whether the radiologist recommended urgent and diagnostic radiographic or surgical procedures versus those with recommendations for continued routine screening appropriate for the patient's age. Thus, we accepted as irrelevant only those normal reports with explicit instructions for routine follow-up. Reports in which no specific recommendations were made, either for urgent follow-up or for routine screening, were excluded from these experiments.

This study question represents a prime potential application of the classifier as an automated health care quality assurance monitor. The ad hoc classifier creates a classification profile designed to detect specific evidence of suspicious conditions in the unstructured text portions of mammograms. Vast numbers of mammograms can be automatically classified for risk of these conditions according to user-defined confidence levels. For high-risk cases, coded data, such as diagnosis and procedure codes, could be accessed in other institutional information systems and automatically reviewed for occurrences of codes signifying appropriate follow-up for the suspected conditions. In such an application, cases without evidence of appropriate follow-up would then be manually reviewed.

### Phrase and Negation Indexing

Noun phrase concepts are recognized by a set of rules applied to the text after it has been part-of-speech tagged by Jtag.[23] Specifically, a sequence of any number of nouns or a sequence of any number of adjective words followed by a head noun are treated as concepts. Noun phrase concepts are used as the seed for finding the most relevant documents, because they play a leading role in characterizing the content of a text.[24]

The challenge of documenting absent findings is addressed by a new function called NegExpander. Our need is to represent differently in Inquery instances of positive evidence and of negative evidence that contain the same keywords and to expand that representation across all components of conjunctive phrases. To do this, we detect in the text the occurrences of a set of negation words ("no," "absent," "without," etc.) and conjunctions, use Jtag to identify noun phrases in the conjunctive phrases and replace the original text with tokenized noun phrase negations, which are themselves regarded as concepts.

For example, the text "NO OTHER SUSPICIOUS MASSES, SUSPICIOUS CALCIFICATIONS OR SECONDARY SIGNS OF MALIGNANCY ARE SEEN" is NegExpanded to "NO_OTHER_SUSPICIOUS_MASSES, NO_SUSPICIOUS_CALCIFICATIONS OR NO_SECONDARY_SIGNS_OF_MALIGNANCY ARE SEEN." Inquery no longer confuses "no suspicious massess" with "suspicious masses" in indexing, retrieval, or classification.

Combining initial abbreviation expansion with NegExpander results in the phrase "NO H/O LESIONS OR CA" being replaced by "NO_HISTORY_OF_LESIONS OR NO_CANCER"

Notice that standard Boolean logic expansion of NOT(A OR B) to NOT(A) AND NOT(B), and NOT(A AND B) to NOT(A) OR NOT(B) is not necessary, because both "and" and "or" are stopped by Inquery.

Figure 3 presents the text section of the mammography report (displayed in its original form in Figure 1) after it was normalized and NegExpanded. The new text is bolded and italicized for easier identification. NegExpander found the following concepts in this document: history of FCBD, upper outer quadrant, **right** mobile breast mass, routine retirement physical, history of fibrocystic changes, **right** breast mobile mass, upper outer quadrant, prior mammograms, extremely dense fibrous tissue, upper outer quadrants of both breasts, sensitivity of mammography, region of palpable abnormality, dense breast tissue, occasional benign appearing calcification, **no_discrete_mass**, **no_suspicious_calcification**, **no_other_secondary_sign_of_malignancy**, **no_mammographic_evidence_of_malignancy**, previous mammograms, palpable abnormality, and clinical grounds.

Although NegExpander works very well for this document, identifying all noun phrase concepts, its overall precision in identifying concepts correctly is only 93 percent. We randomly selected ten documents and used NegExpander to find the contained concepts. The documents were then manually reviewed to identify noun phrase concepts. Some errors were due to incorrect part-of-speech tagging by Jtag, and some errors were made by NegExpander itself.

<TEXT>
<REASON_FOR_EXAM>
NO_BRIEF_COMMENT.
42 ***year*** old G0P0 with ***history of*** FCBD. Approx 3-4cm upper outer quadrant ***right*** mobile breast mass. Routine retirement physical.
</REASON_FOR_EXAM>
<INTERPRETATION>
CLINICAL HISTORY: 42 year old with history of fibrocystic changes, ***right*** breast mobile mass upper outer quadrant.
BILATERAL FILM SCREEN MAMMOGRAM:
FINDINGS: Prior mammograms are not available for comparison. The breasts are heterogeneously dense. There is extremely dense fibrous tissue in the upper outer quadrants of both breasts. This lowers the sensitivity of mammography. B.B. was placed in the region of palpable abnormality and demonstrates dense breast tissue in this region. An occasional benign appearing calcification is present in both breasts. ***No_discrete_mass***, ***no_suspicious_calcification*** or ***no_other_secondary_sign_of_malignancy*** is demonstrated.
<IMPRESSION>
***no_mammographic_evidence_of_malignancy***. Recommendation: comparison with patient's previous mammograms is recommended. The decision to biopsy any palpable abnormality should be based on clinical grounds.
</IMPRESSION>
</INTERPRETATION>
</TEXT>

**Figure 3** Normalized and NegExpanded mammography report.

NegExpander hashes these concepts into a concept table and counts their frequencies. The concepts, sorted by frequency, are presented to the user and used by RelHelper as relevant concepts. For each document, NegExpander tokenizes not only the negative concepts but also the positive concepts to create a concept document. The concept document, which contains only minimal SGML tags and the collection concept list, is used for training the classifier.

### Classification Techniques

To generate the classification query profile we use the stand-alone version of Inquery's relevance feedback module[25] running with a list of document identifiers and their relevance/irrelevance judgments (the relevance file), against a training collection. The training collection is built with RelHelper and indexed with Inquery, using only those documents included in the relevance file.

#### Using RelHelper to Build Training and Test Collections

RelHelper is a new application that assists a domain expert in the creation of training and test collections by facilitating domain knowledge acquisition. Document judgment can be done manually, but RelHelper provides a point-and-click interface that speeds this review and scoring by providing caches of documents that are increasingly either relevant or irrelevant to the classification question.

Using RelHelper, the expert selects from all the concepts extracted from the collection, a set of key concepts that are most relevant to the classification question. RelHelper generates a seed query with the selected concepts,[26] runs that query on the collection, and presents the documents ranked in relevance to the seed query. The domain expert then reviews the documents, either high or lower on this list, and makes relevance judgments on them. The user's time is optimized by allowing the user to focus on those documents most likely to contribute useful positive and negative evidence to the classifier, without having to review and judge all the documents or select them randomly.

The seed is expanded to review and score more documents. RelHelper uses the already judged documents and relevance feedback to add new concepts, rich in positive and negative evidence, to the seed concepts. Default values are used for Rocchio parameters (eq. 1), such as

$$Rocchio = w_{query} + 2w_{rel} - 0.5w_{irr}$$

The first 100 terms ranked by Rocchio weight are added to the seed query of user-selected concepts. Previous query terms are always included in the new query, because they have been shown to be generally more reliable than automatically generated query terms.

The output from RelHelper is a relevance file, made up of document identifiers, relevance judgments, and the most recent query formulation, which is used to generate training and test collections from the document collection.

### Using Relevance Feedback to Generate the Profile Query

After a training collection has been built using RelHelper, the ad hoc classifier uses relevance feedback to generate the query profile, i.e., the classifier. Relevance feedback extracts relevant features (words, phrases, proximities, etc.), calculates weights for every feature, and constructs the query profile in the Inquery format. The weights for both the relevant and irrelevant documents use experimentally derived Rocchio parameters:

$$Rocchio = 6w_{rel} - 2w_{irr}$$

where $w_{query}$ is eliminated by a zero value for α because there are now no original query terms when running relevance feedback for training.

Ideally, relevance feedback would return a sufficient number of ''good'' concepts so that the retrieval result includes all relevant documents and no irrelevant documents. However, we need to categorize every document in a collection, not just the relevant ones; the irrelevant documents must also be retrieved and ranked. To achieve this, we run relevance feedback twice, once for the relevant documents and once for the irrelevant documents. The Inquery *NOT* operator is applied to the terms from the irrelevant run, which gives an opposite weight (i.e., $1 - w$). The two query profiles are then merged, ranking the relevant documents high and irrelevant documents low.

An experimental merging weight parameter, $r$, is used to increase the weights for those terms from the relevant run and decrease the weights for those terms from the irrelevant run. Thus, the merged formula of a document is

$$w_d = \left(r \cdot \sum_{i \in rel-run} R_i\right) + \left((1 - r) \cdot \sum_{j \in irr-run} R_j\right) \quad (2)$$

where $R_i$ is the Rocchio weight for the term $i$ from the relevant run, $R_j$ is the weight for term $j$ from the irrelevant run, and $r$ is the merging weight. If $r = 1$, then the terms derived from the irrelevant run are ignored. If $r = 0$, terms from the relevant run are ignored. After merging the query profiles, a test classification is run with the test collection.

### Three-bin Classification

The product of the ad hoc classifier is a three-category sort, with each document being assigned to either a positive, uncertain, or negative bin. Distinctions between the bins are based on user-selected thresholds for the desired correct and incorrect rates of document assignment to the bins. The cutoff values are derived for the thresholds from a ranking of relevant and irrelevant documents in the training collection by their document weights. We use logistic regression[27,28] to smooth the document weights and compute the cutoff values.

Initially, for each position in the document weight ranking (from top to bottom for the positive cutoff value), we compute a precision, called ''observed precision,'' which is a ratio of the number of relevant documents to the number of documents from the top-ranked document to that position.

In logistic regression we use a likelihood,

$$L = \sum_i \log \frac{e^{\beta 0 + \beta 1 \cdot Wd_i}}{1 + e^{\beta 0 + \beta 1 \cdot Wd_i}} \quad (3)$$

where $L$, β0, and β1 need to be initialized and optimized. To initialize these parameters, we group the documents and compute a mean weight and mean relevant rate for each group, such as

$$MWd_k = \frac{\sum_{i \in kth-bin} Wd_i}{N_k}$$

where $Wd_i$, from equation 2, is the weight for the document $i$, and $N_k$ is the number of documents in $k$th group. Similarly, we have the mean

$$Mrel_k = \frac{Nrel_k}{N_k}$$

where $Nrel_k$ is the number of relevant documents in the $k$th group. From these groups we find a range, $l < MWd_k < h$, and a mean of document weights from the range $(l, h)$, initially set arbitrarily at $l = 0.2$ and $h = 0.8$:

$$MWd_{(l,h)} = \frac{\sum_{i \in (l,h)} Wd_i}{|(l, h)|}$$

We then estimate the parameters as

$$\beta 1 = \log \frac{Mrel_l}{1 - Mrel_l} - \log \frac{Mrel_h}{1 - Mrel_h} \quad (4)$$

$$\beta 0 = \log \frac{Nrel_{(l,h)}}{N - Nrel_{(l,h)}} - \beta 1 \cdot MWd_{(l,h)} \qquad (5)$$

where $N$ is the number of documents and $Nrel_{(l,h)}$ is the number of relevant documents in the ranking.

β0 and β1 are initialized at zero and are optimized by iterative updates until $L$ becomes stable. The iterative updates use values of β1 (eq. 4) varying in the range of β1 ± epsilon, ε, with subsequent update of β0 (eq. 5) to find any new maximum $L$ using equation 3. With the optimized parameters of β0 and β1 we do the regression for every document position in the ranking, such as

$$Pr(Wd_i) = \frac{e^{\beta 0 + \beta 1 \cdot Wd_i}}{1 + e^{\beta 0 + \beta 1 \cdot Wd_i}}$$

$Pr(Wd_i)$ is called the probability of the document $i$ to be relevant, so that the cutoff value, called a fit precision, at position $j$ is

$$C_j = \frac{\sum_{i=1}^{j} Pr(Wd_i)}{j}$$

If the user chooses a threshold of 90 percent, the three-bin sort program sets the cutoff value to be the weight of the document whose fit-precision is closest to 90 percent. We similarly compute the cutoff values for the negative bin, doing the regression reversely, from bottom to top.

Classification is the application of the query profile to a collection and the sorting of its documents into three bins. Documents whose weights are greater than the positive cutoff are assigned to the positive bin, and those with weights less than the negative cutoff are assigned to the negative bin. Documents whose belief values fall between the cutoffs are assigned to the uncertain bin.

Because the cutoff values may overlap, the ad hoc classifier compares each document weight to the positive and negative cutoffs to ensure that the weight is greater than both cutoff values for the positive bin and less than both cutoff values for the negative bin. For example, the user may seek a positive threshold of 90 percent and a negative threshold of 90 percent. The positive cutoff value for the threshold may be 0.45, while the negative cutoff value for 90 percent may be 0.46. Overlap may occur if the training collection is too small, if it significantly fails to represent the general document collection, or if the user-specified degree of accuracy is too low. Usually, increasing the desired positive threshold can eliminate cutoff overlap. If this is not the case, distinguishing characteristics between relevant documents and irrelevant documents in the training collection are insufficient, because of unsuitable training parameters, faulty relevance judgments, too small a number of training documents, or insurmountable similarities between the two classes of documents.

### Automated Profile Enhancement

The optional classifier enhancement program runs a trained query profile against the training collection, analyzing the contribution of every query term of every misclassified document (a relevant document in the negative bin or an irrelevant document in the positive bin) and every document in the uncertain bin. The enhancement program increases the weight of a profile term if it occurs in a relevant uncertain or negative document, and decreases the weight of a term if it occurs in an irrelevant uncertain or positive document. Term weights can be modified many times, with a reweight strength that is selected by the user. There is no net change in term weight if the number of increments and decrements of equal strength are equal.

Users may repeat enhancement multiple times if significant document misclassification continues. Enhancement improves performance of a classification profile with the training collection, but that improvement can not be guaranteed for the classification of a test collection, because of overtraining. In this situation, the profile is too highly customized to one particular collection, the training collection, and does not perform well with any others.

## Evaluation

The performance of the trained profile is evaluated with a benchmark test collection and its relevance judgments. If a classification profile does not meet the desired cutoffs, the user may either run the enhancement program to automatically modify query terms in the profile, manually edit the profile, or do another round of training. The further training may include increasing the number of training documents and modifying the type and number of features or other parameters used by the relevance feedback.

### Test Run

The classified documents are actually sorted into six bins—three bins for relevant documents and three for irrelevant documents, shown in Table 1. True positive means that a relevant document is sorted into the positive bin, while false positive means that an irrelevant

*Table 1* ■

Six Bins for Classification Testing

| | Relevant Documents | Irrelevant Documents |
|---|---|---|
| Positive | True positive | False positive |
| Uncertain | False uncertain | False uncertain |
| Negative | False negative | True negative |

*Table 2* ■

Six Classification Contingency Table Values

| | Relevant Documents | Irrelevant Documents |
|---|---|---|
| Positive | *a* | *b* |
| Uncertain | *c* | *d* |
| Negative | *e* | *f* |

document is sorted into the positive bin. Similarly, true negative means that an irrelevant document is sorted into the negative bin, and false negative means that a relevant document is sorted into the negative bin. Since there are no uncertain documents in the benchmark test collection, all documents assigned to the uncertain bin are false classifications.

The three-bin sort program displays the number of documents assigned to each bin and calculates true and false document categorization rates. The user can easily access any bin and any documents in the bins for review of relevance judgments. The classifier interface allows the user to view the query terms present in any document as well the contributions of these terms to the weight of a document. The interface also allows the user to re-sort the documents manually, move a document from a bin to another bin, or run the enhancement program to automatically modify the query profile.

## Evaluation Measures

Performance evaluation concerns both the true positive and the true negative rates. The effectiveness of this classification may be represented by six contingency table values, shown in Table 2.

Several effectiveness measures can be defined for the classifier using these values. The standard IR metrics of recall and precision for the positive bin are precision of positive bin

$$PRC_{pos} = \frac{a}{a + b}$$

and recall of positive bin

$$RCL_{pos} = \frac{a}{a + c + e}$$

Because the three-bin sort is not a binary classification, we use a single, summarized measure to evaluate the classifier, a modified *F* measure.[29,30] The *F* measure provides a weighted combination of recall and precision for each of the positive and negative bins separately, which are then averaged into one value.
The *F* measure for the positive bin is defined in the terms of table values:

$$F\beta_{pos} = \frac{(\beta^2 + 1)\cdot PRC_{pos}\cdot RCL_{pos}}{\beta^2\cdot PRC_{pos} + RCL_{pos}} = \frac{(\beta^2 + 1)\cdot a}{(\beta^2 + 1)\cdot a + b + \beta^2\cdot(c + e)}$$

The parameter β ranges between 0 and infinity and controls the relative weight given to recall and precision. A β of 1 corresponds to equal weighting of recall and precision. With β = 1, we have $F1_{pos} = 2a/(2a + b + c + e)$. If it should occur that *a*, *b*, *c*, and *e* are all 0, we define $F1_{pos}$ to be 1 to avoid division by 0.

Similarly, the *F* measure for the negative bin is defined as:

$$F\beta_{neg} = \frac{(\beta^2 + 1)\cdot PRC_{neg}\cdot RCL_{neg}}{\beta^2\cdot PRC_{neg} + RCL_{neg}} = \frac{(\beta^2 + 1)\cdot f}{(\beta^2 + 1)\cdot f + e + \beta^2\cdot(b + d)}$$

and $F1_{neg} = 2f/(2f + e + b + d)$.

The final measure is an average of $F1_{pos}$ and $F1_{neg}$:

$$F1 = \frac{F1_{pos} + F1_{neg}}{2} = \frac{a}{2a + b + c + e} + \frac{f}{2f + e + b + d}$$

Thus, the single metric, *F*1, serves to summarize classifier performance, encompassing both precision and recall for both positive and negative bins.

## Results and Discussion

Three series of experiments were performed to evaluate the ad hoc classifier. The first experiments concern refinement of profile features, the second concern refinement of collection features, and the last concern NegExpander. Our benchmark document set included 421 relevant and 256 irrelevant documents from the calcification recommendation classification task. All experiments used a 90 percent cutoff for both positive and negative bins.

### Experiments to Refine Query Profiles

We do not expect that all relevant documents will be ranked above all irrelevant documents, allowing perfect performance in the three-bin sort. However, we conducted experiments toward this goal to determine the number, type, and weight of terms that should be added to the profile to optimize its classification performance. For both the training and test document collections, we randomly selected 100 relevant and 100 irrelevant documents to avoid the confounding influences of the size of collections and ratio of relevant and irrelevant documents. We conducted experiments for all 15 relevance feedback parameters and present in this paper the three most significant.

#### Number of Terms Added to the Profile

The first experiment was designated to find an optimal number of profile terms for refining classifier features and to retrieve all documents from the test collection. Figure 4 shows the change in the number of retrieved documents and in the *F* measure as the number of single terms in the profile increases. Both dependent variables initially improved: At 40 single

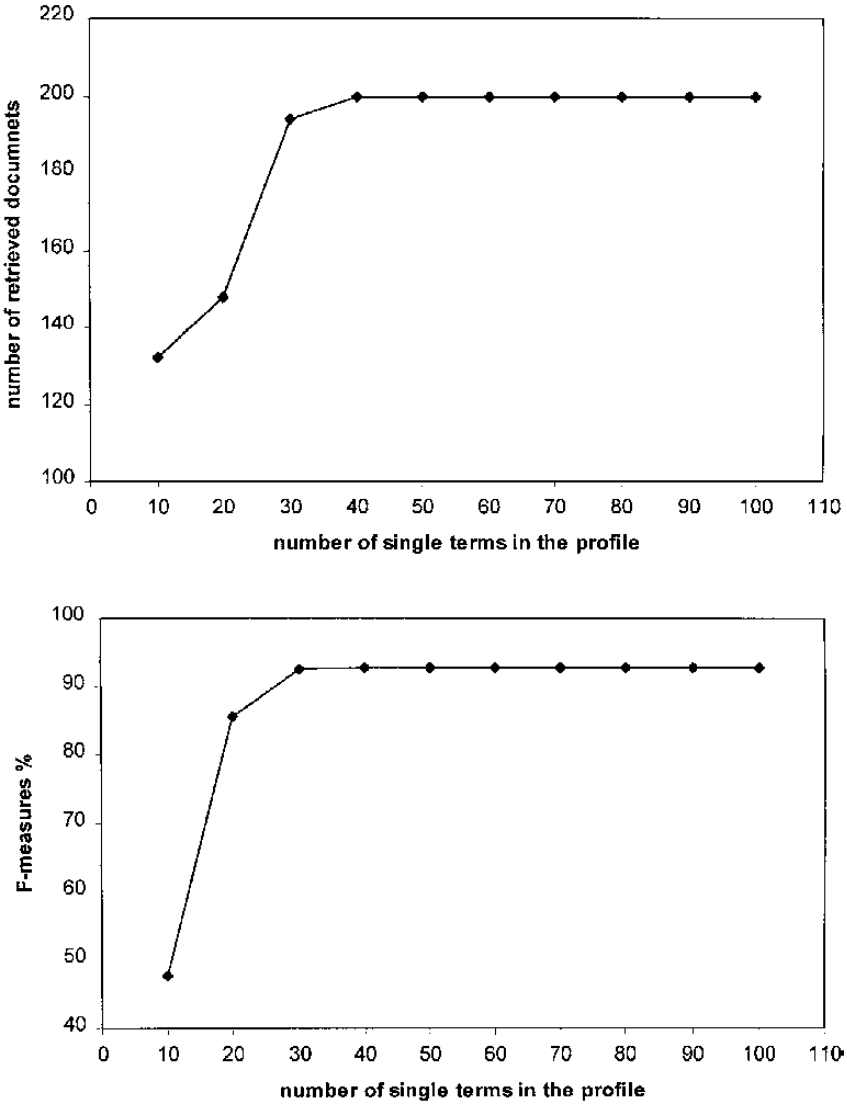



**Figure 4** Documents retrieved (*top*) and *F* measure (*bottom*) as functions of the number of single terms added to the query profile.

terms all 200 test documents are retrieved and the highest *F* measure provided. After this point the *F* measure stops improving, because relevance feedback cannot identify other significant terms to add to the profile, although the actual number of terms added to the profile continues to increase.

Similar experiments were conducted with each type of query feature, and in all cases profile performance leveled off after the addition of a certain number of terms to the profile. Neither pairs of single terms nor NegExpanded phrases performed alone better than the single terms.

We experimented with combinations of term types—single terms with ordered pairs, and single terms with unordered pairs. Performance was initially degraded adding either to single terms. Additional unordered pairs (40 pairs or more) yielded performance slightly better than single terms alone, while additional ordered pairs did not. Considering that addition of unordered pairs may improve the performance for other test collections, the combination of 40 single terms and 40 unordered pairs was used in subsequent experiments. Experiments were also conducted varying the proximity of the members of the term pairs, but there was no improvement as proximity varied.

In other experiments, combining 40 single terms with any number of noun phase concepts does not improve the profile. We had expected the addition of phrases to improve performance, but the experimental results demonstrate that phrases are not an important feature in this collection. Experiments with the Rocchio weighting parameters established best performance—*F* measure 93.3 percent, with $\beta = 6$ and $\gamma = 2$—although several other settings also improved classifier performance.

### Document Merging Weight

The merging weight parameter between relevant and irrelevant documents, *r*, was varied, as shown in Figure 5. A higher merging weight parameter yields higher *F* measures, as when the value is in a range (0.7, 1). Test documents that contain more terms from the relevant document training set are appropriately ranked higher, while those test documents containing more terms from the irrelevant document training are ranked lower. This suggests that the contribution of evidence from relevance feedback of irrelevant documents should be minimized.

### Profile Enhancements Based on Misclassified Documents

Automated reweighting of query terms in misclassified documents may be repeated as many times as a user desires. This enhancement improves performance of a classification profile with the training collection, but that improvement can not be guaranteed for classification of a test collection, because of overtraining. The profile becomes too highly customized to one particular collection, the training collection, and does not perform well with any others.

Figure 6, *top*, shows the performance of multiple enhancements run on the training collection, with three reweight strengths—0.1, 0.2, and 0.5. Best performance was obtained on the tenth running of the enhancement program and a reweight strength of 0.2.

The ten sequential enhanced profiles with reweight strength of 0.2 were then run on the test collection (Figure 6, *bottom*). Moderate repetition of enhancement on the training collection produces improved classification of the test collection, the highest *F* measure, 93.8 percent, at four enhancements. However, continued enhancement of the training collection degrades the performance of the profile on the test collection because of overtraining.

## Experiments to Refine Training and Test Collection Features

The following experiments used the relevance feedback parameters that performed best in previous experiments, while varying the size of the training and test collections and the proportion of relevant and irrelevant documents in each collection. Specifically, a combination of 40 single terms with 40 unordered pairs was included in the profile, with Rocchio parameters of $\beta = 6$ and $\gamma = 2$ and the merging weight $r = 0.9$. There was no automated enhancement of the profiles during training.

### Training Collection Size

We created 13 training sets, increasing the number of documents from 60 to 300 by intervals of 20, with equal numbers of relevant and irrelevant documents. The test collection contained 100 relevant and 100 irrelevant documents. For each training set, a classification profile was created and run on the test collection.

Figure 7 shows *F* measures as the training collection size increased. Performance improved from 89.1 to 94.1 percent as the collection size grew from 60 to 220, although not completely smoothly. After 220 training documents, however, there were no further improvements and the *F* measure declined. This suggests that a training saturation, at about 220 documents, may have been reached for this test collection. It is also possible that the parameter settings for relevance

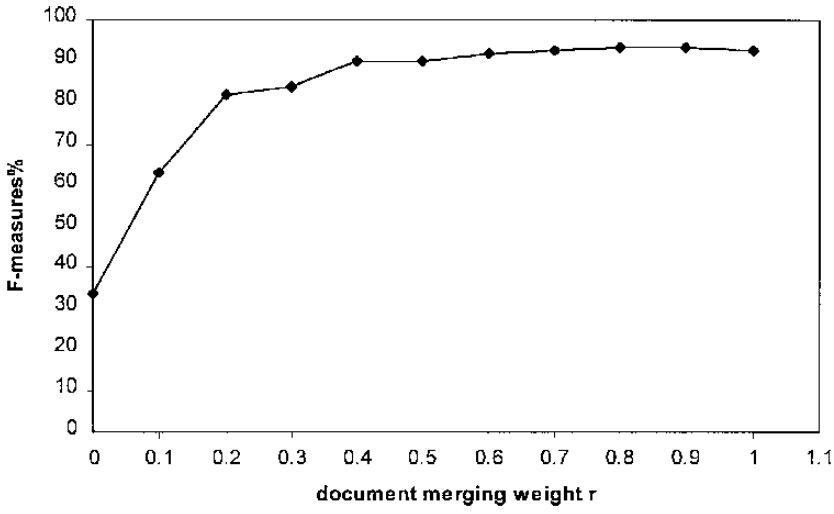
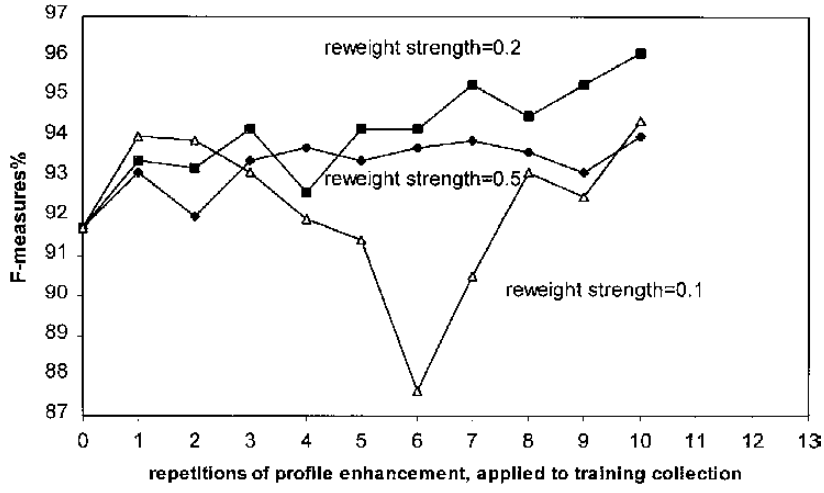



**Figure 5** *F* measure as a function of document merging weight *r*.



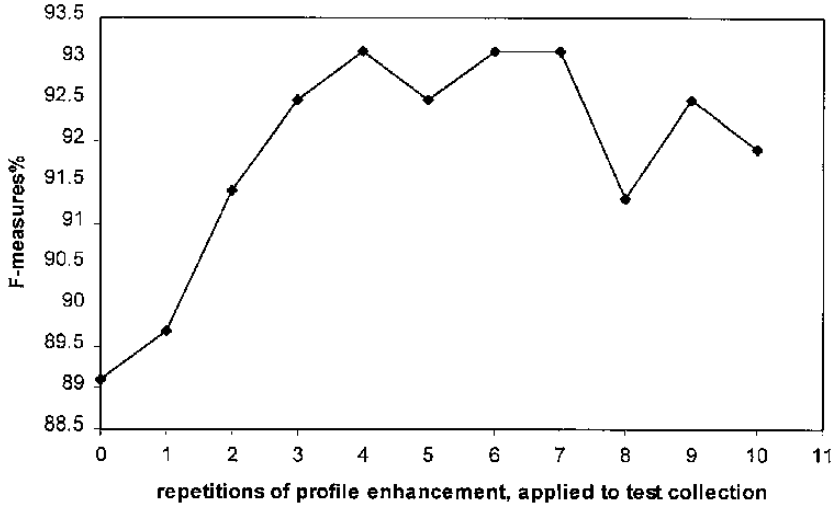




**Figure 6** *F* measure as a function of iterations of the enhancement program on the training collection (*top*) and the test collection (*bottom*).

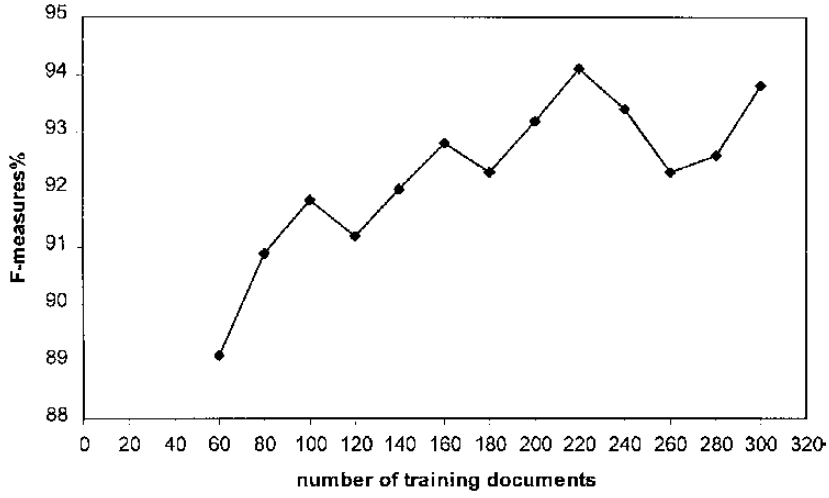



**Figure 7** *F* measure as a function of increased training collection size.

feedback may not be suitable for larger training collections. This is investigated in a later experiment.

### Proportion of Relevant and Irrelevant Documents

To study the impact of the proportion of relevant and irrelevant documents in the training and test collections on classifier performance, we conducted three experiments varying the proportions of relevant and irrelevant documents in the training or test collections, or both.

In the first experiment the proportion of relevant and irrelevant documents in the training collection was varied, using the same test collection (equal balance of 100 relevant and irrelevant documents) as the previous experiments. We first fixed 100 irrelevant documents in the training collection and increased the number of relevant documents from 0 to 100. Then we fixed 100 relevant documents and increased the irrelevant number from 0 to 100. Figure 8, *top*, shows that a near-equal balance of relevant and irrelevant documents in the training collection performed best for a balanced test collection.

In the second experiment we used the profile trained with an equal balance of 100 relevant and 100 irrelevant documents and varied with proportions of the test collection. We first fixed 100 relevant documents and increased the irrelevant documents from 0 to 100, and then fixed 100 irrelevant documents and increased the relevant documents from 0 to 100.

Figure 8, *middle*, shows that the profile created from equal numbers of relevant and irrelevant documents performed well for a wide range of test collection relevance proportions. *F* measures were consistently greater than 90 percent for either relevant or irrelevant document counts as low as 40 in the test collection.

The third experiment varied the proportion of relevant and irrelevant documents in the training and test collections, which were kept at equal size and proportions. For example, when the training collection contained 50 relevant and 100 irrelevant documents, the test collection did likewise.

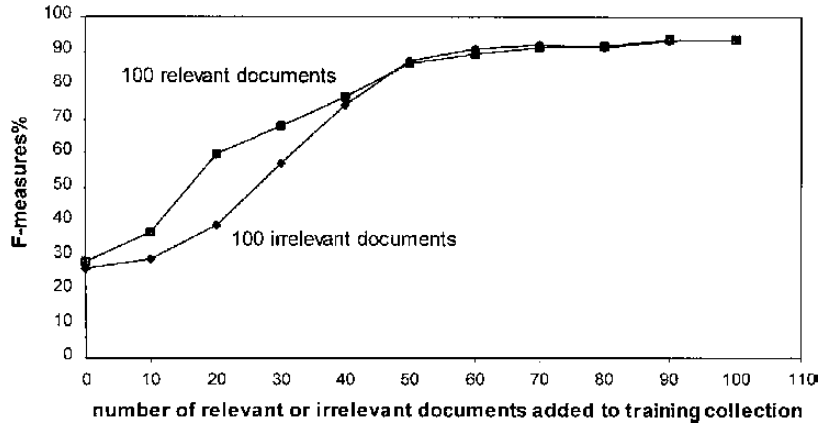
Figure 8, *bottom*, shows similar results. When the proportion of a training or test collection is very unbalanced, even if the training and the test collections have exactly the same proportions, performance suffers. For ad hoc classification tasks, then, the proportions of relevant and irrelevant documents should be close to balanced in the training collection, so that both relevant and irrelevant documents are fairly represented and contribute relevance evidence equally.

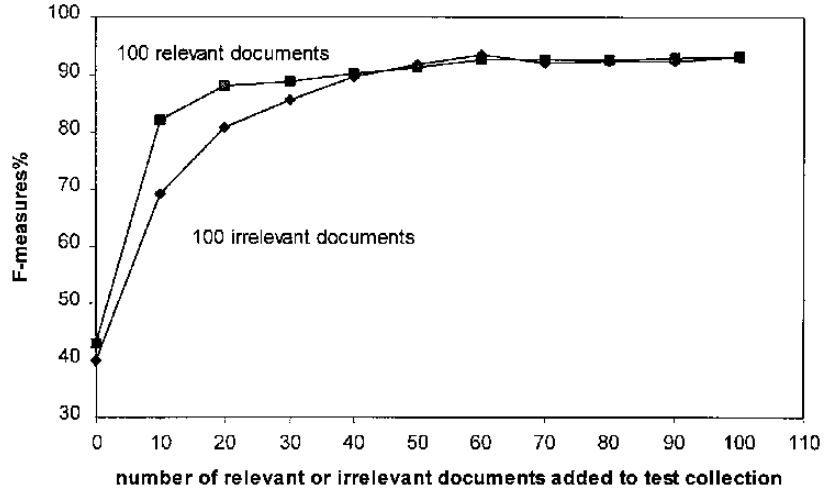
### Test Collection Size

To study the ability of a profile to classify a large document collection, we designed an experiment that used a small training collection and much larger test collections. Forty relevant and 40 irrelevant documents were randomly selected for the training collection. From the remaining 381 relevant and 216 irrelevant documents we created seven test collections, whose proportions of relevant and irrelevant documents are listed in Table 3.

We used two sets of training parameters. The first set was the same as in the previous experiments. The second set arbitrarily comprised 50 single terms and 30 unordered pairs, with $\beta = 2$ and $\gamma = 8$. No other experiments were run for these parameters for this training collection.

Figure 9 shows that increasing the size of the test collection from that close to the size of the training collection degrades classifier performance. A better initial



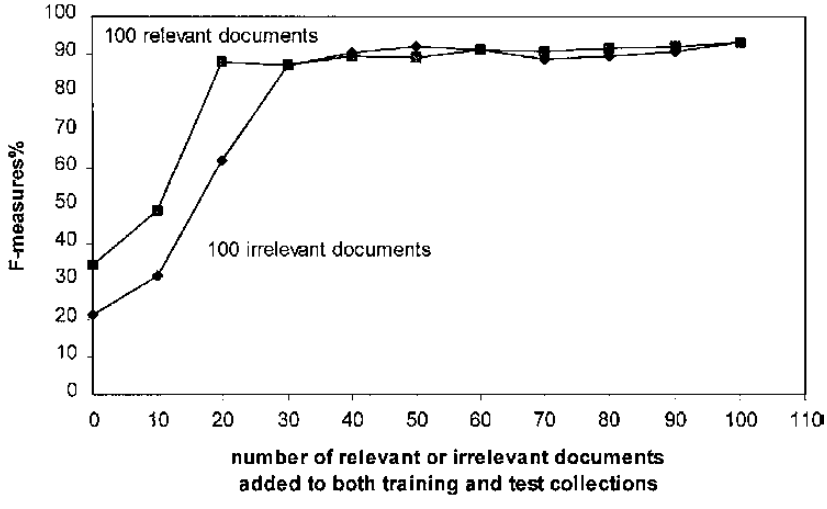




**Figure 8** *F* measure as a function of proportions of relevant and irrelevant training documents (*top*) and test documents (*middle*) and as a function of equally unbalanced proportions of relevant and irrelevant training and test documents (*bottom*).

*Table 3* ■

Seven Test Collections of Increasing Size

| | No. of Relevant Documents | No. of Irrelevant Documents | Total |
|---|---|---|---|
| Test 1 | 50 | 50 | 100 |
| Test 2 | 53 | 87 | 140 |
| Test 3 | 105 | 141 | 246 |
| Test 4 | 163 | 216 | 379 |
| Test 5 | 216 | 216 | 432 |
| Test 6 | 320 | 216 | 536 |
| Test 7 | 381 | 216 | 597 |

classification profile degrades more slowly, as shown in the plot of the second set of parameters: The *F* measure degrades from 89.7 to 79.1 percent, while the total collection size increases from 100 to 597. We did not have sufficient numbers of benchmarked documents to determine when the *F* measure levels off with increasing test collection size, as we would expect.

### Negation Sensitivity

In evaluating the data quality on the ad hoc classifier performance, we conducted an experiment in which NegExpander was applied differently to the training and test collections. We constructed a set of ''negation-sensitive'' training and test collections, by extracting every unique NegExpanded term from the 421 relevant and 256 irrelevant documents and identifying those terms that occurred in only irrelevant documents. From the whole document set we then extracted those documents that either were benchmarked as irrelevant documents and contained NegExpanded terms but no opposites or were benchmarked as relevant documents and contained opposites of the NegExpanded terms in the list but no NegExpanded terms. For example, the opposites of the NegExpanded terms ''no_suspicious_microcalcifications,'' ''no_suspicious_findings,'' and ''no_nipple_abnormalities'' are ''microcalcifications,'' ''findings,'' and ''nipple abnormalities,'' respectively.

This process yielded 97 relevant and 157 irrelevant documents to make up the ''negation-sensitive'' collection, from which we randomly extracted 40 relevant and 40 irrelevant documents for the training collections and used the remaining 57 relevant and 117 irrelevant documents for the test collections. With these documents we prepared two training collections, one processed by NegExpander and the other not, as well as two test collections, again, one processed by NegExpander and the other not.

This experiment has two parts—one with ''negation-sensitive'' collections, the other one with ''normal'' collections acting as controls. The ''normal'' collections were made of 40 relevant and 40 irrelevant documents for the training and 57 relevant and 117 irrelevant documents for the test, all extracted randomly from 421 relevant and 256 irrelevant documents. With these documents, again, we prepared two training collections, one processed by NegExpander and the other not, as well as two test collections. These experiments used merging weight $r = 0.9$, Rocchio $\beta = 2$ and $\gamma = 8$, and varied the number of single terms added to the profile from 10 to 100.

Figure 10, *top*, shows that NegExpander improved performance slightly in the ''normal'' collection, while Figure 10, *bottom*, shows a large difference between the performance in the ''negation-sensitive'' collection without NegExpander compared with the performance with NegExpander at the point of 10 terms contributing to the profile. Three NegExpander terms are found in the ten-term profile: ''no_suspicious_microcalcifications,'' ''no_suspicious_findings,'' and ''no_nipple_abnormalities.'' Increasing the number of terms in the profile, although adding more NegExpanded terms to the profile, reduced the differences in performance. When more than 40 terms were added to the profile, the differences became unclear.

NegExpander improves classifier performance for the ''negation-sensitive'' collection when the number of terms in the profile is limited and there is critical negative evidence to be considered. If the irrelevant documents of the collection contain most of the negative terms (negative evidence) and the relevant documents contain most of the positive terms (positive evidence), NegExpander might contribute even more significantly to classifier performance.

## Conclusion

Automated classification of clinical documents is an active field of medical informatics research. A number of centers are experimenting with multiple methodologies and several document types to discover ways to access and make usable the clinical information and knowledge embedded in unstructured electronic text.

The ad hoc classifier is designed to provide a text classification capability in a fully automated fashion given a collection of relevance-judged training documents. When the system is trained with a sufficient number of representative documents, the trained classifier can sort a large collection of documents into the three bins of user-defined ad hoc classes, permitting expert re-

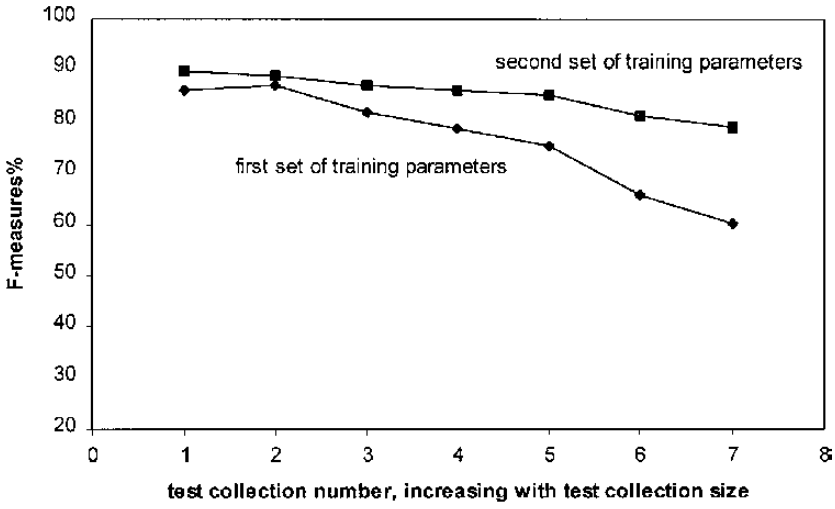
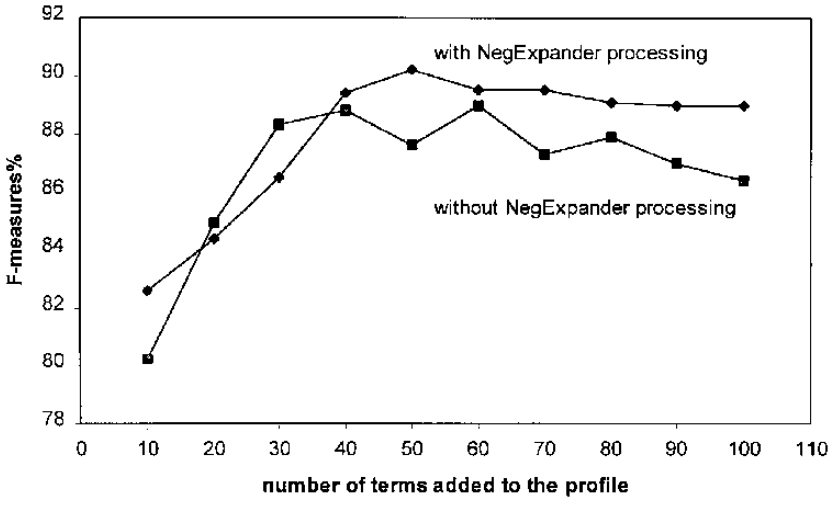



**Figure 9** *F* measure as a function of increasing size of the test collection.



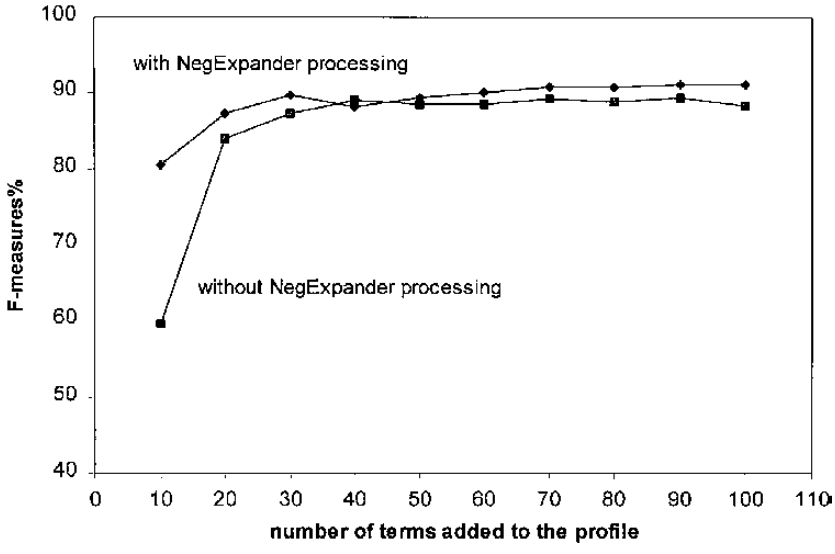




**Figure 10** *F* measure as a function of NegExpander and number of profile terms in "normal" collections (*top*) and in "negation-sensitive" collections (*bottom*).

view of documents to be limited to those in the uncertain class. NegExpander is an approach that can distinguish positive from negative evidence and may play an important role in classification when consideration of negative evidence is critical.

A key challenge of information systems research is to keep the work relevant. One aspect of this is the ability to transfer lessons from the idealized world of experimentation to the real world of health care delivery. For the classification task that we undertook (separation of reports with recommendations for urgent versus routine follow-up for mammograms with abnormal findings), we excluded documents with no follow-up recommendations and those in which recommendations were insufficiently explicit. This eliminated the majority of naval medical center reports and is an important limitation of our work. In the real world, interpretive test reports are not presorted to provide only the abstract ideal cases for automated classification. Transfer of the ad hoc classifier to a production environment will require extensive additional development in this area.

Another challenge is the availability of benchmarked ''gold standards'' against which experimental information systems can be evaluated. Despite the restrictions discussed above, the relevance judgments made in our research were not always straightforward. Ambiguity exists in the natural history of a disease process, the clarity of its manifestations, the abilities of the observers, and the skill of the observers in using language to record their findings. These factors add to the difficulty in judging the relevance of some reports, even with the ''relevant'' concepts highlighted by RelHelper. This may have contributed to training collections not being completely representative of the test collections, which would impair system performance. Hripcsak et al.[31] have recently studied the reliability of physicians in extracting facts from radiology reports in order to create benchmark standards for information systems research. Only one document reviewer participated in our research, and the validity or consistency of those relevance judgments was not reviewed.

Continued development of ad hoc classification systems for clinical documents needs to proceed in two additional areas. The first concerns making use of the internal structure of clinical documents. As evident in Figures 1 and 3, radiology reports contain many identifiable data fields. In our experience these fields are used in highly unpredictable ways. However, they may be able to be manipulated to be more useful in classification.

We constructed two subfields, ⟨REASON FOR EXAM⟩ and ⟨IMPRESSION⟩, in the unstructured text portions of our data, anticipating that we may want to differentially consider the evidence they contain. We have yet to explore this, although parallel work at our center, in automated ICD-9-CM code assignment to hospital discharge summaries, has taken advantage of building internal structure into clinical documents.[20]

The second area for further development of statistically based classification systems concerns increasing the degree of natural language understanding used by the classifier. We were disappointed in the impact of NegExpander's detection and expansion of negative evidence across conjunctive phrases. Although we incorporated only a small number of negation variants to detect, with no analytic logic in their implementation, we expected a more dramatic effect. We believe this type of special handling of test interpretations will be important in classification. Improving the performance of NegExpander would require incorporation of more natural language processing techniques, such as better sentence segmentation and a syntax parser.

The authors would like to acknowledge the contributions to this research of James Allen, for his work on relevance feedback, Xu Jinxi, for his work on Jtag, and Leah Larkey, for her work on logistic regression.

*References* ■

1. Callahan JP, Croft WB, Harding SM. The INQUERY retrieval system. Proceedings of the Third International Conference on Database and Expert Systems Applications. New York: Springer-Verlag, 1992:78–83.
2. Robertson SE. The probability ranking principle in IR. J Documentation. 1997;33:294–304.
3. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, Calif.: Morgan Kaufmann, 1988.
4. Charniak E. Bayesian networks without tears. AI Magazine. 1991;12(4):50–63.
5. Turtle HR, Croft WB. Efficient probabilistic inference for text retrieval. In: Proceedings of the 1991 Recherche d'Informations Assistee par Ordinateur (RIAO '91) Conference. 1991:644–61.
6. Turtle HR, Croft WB. Evaluation of an inference network–based retrieval model. Applied Computing Machinery (ACM) Trans Inf Syst. 1991;9(3):187–222.
7. Rocchio JJ Jr. Relevant feedback in information retrieval. In: Salton G (ed). The SMART Retrieval System: Experiments in Automatic Documents Processing. Englewood Cliffs, N.J.: Prentice-Hall, 1971:313–23.
8. Buckley C, Salton G, Allan J. The effect of adding relevance information in a relevance feedback environment. ACM Special Interest Group on Information Retrieval (SIGIR). 1994:292–300.
9. Allan J, Ballesteros L, Callan JP, Croft BW, Lu ZH. Recent experiments with INQUERY. Proceedings of the Fourth Text

Retrieval Conference (TREC-4). Gaithersburg, Va.: National Institute of Standards and Technology, 1995:49–63.
10. Aronow DB, Cooley JR, Soderland S. Automated identification of episodes of asthma exacerbation for quality measurement in a computer-based medical record. Proc 19th Annu Symp Comput Appl Med Care. 1995:309–13.
11. Aronow DB, Soderland S, Ponte JM, Feng F, Croft WB, Lehnert WG. Automated classification of encounter notes in a computer-based medical report. In: Greenes RA, Peterson HE, Protti DJ (eds), Proceedings of the Eighth World Congress on Medical Informatics. Edmonton, Alberta, Canada: Healthcare Computing & Communications Canada, 1995:8–12.
12. Lehnert W, Soderland S, Aronow D, Feng F, Shmueli A. Inductive text classification for medical applications. J Exper Theoret Artif Intell. 1995;7:49–80.
13. Aronow DB, Feng F. Ad hoc classification of electronic clinical documents. Digital Library [online magazine]. Jan 1997. Available at: http://www.dlib.org/dlib/january97/medical/01aronow.html.
14. Hripcsak G, Friedman C, Aldeson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med. 1995:681–8.
15. Jain NL, Knirsch CA, Friedman C, Hripcsak G. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. Proc 1996 AMIA Annu Fall Symp. 1996:542–6.
16. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc 1997 AMIA Annu Fall Symp. 1997:829–33.
17. De Estrada WD, Murphy S, Barnett GO. Puya: A method of attracting attention to relevant physical findings. 1997 AMIA Annu Fall Symp. 1997:509–13.
18. Cooper GF, Miller RA. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. J Am Med Inform Assoc. 1998;5:62–75.
19. Yang Y, Chute CG. An application of Expert Network to clinical classification and MEDLINE indexing. Proc 18th Annu Symp Comput Appl Med Care. 1994:157–61.
20. Larkey LS, Croft WB. Combining classifier in text categorization. ACM SIGIR. 1996:289–97.
21. Hersh W, Leen T, Rehfuss S, Malveau S. Automatic prediction of trauma registry procedure codes from emergency room dictations. Medinfo. 1998:665–9.
22. Haines D, Croft WB. Relevance feedback and inference networks. Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval. ACM SIGIR. 1993:2–11.
23. Xu J, Broglio J, Croft WB. The Design and Implementation of a Part-of-Speech Tagger for English. Amherst, Mass.: University of Massachusetts, 1994. Technical report IR52.
24. Jing Y, Croft WB. An Association Thesaurus for Information Retrieval. Amherst, Mass.: University of Massachusetts, 1994. Technical report IR47.
25. Allan J, Callan JP, Croft WB, et al. INQUERY at TREC-5. In: Harman D (ed). Proceedings of the Fifth Text Retrieval Conference (TREC-5). Gaithersburg, Va.: National Institute of Standards and Technology, 1996.
26. Turtle H, Croft WB. Inference networks for document retrieval. In: Vidick J-L (ed). Proceedings of the 13th International Conference on Research and Development in Information Retrieval. ACM SIGIR. 1990:1–24.
27. Wegman EJ, Wright IW. Splines in statistics. J Am Stat Assoc. 1983:78:351–65.
28. Silverman BW. Some aspects of spline smoothing approach to nonparametric regression curve fitting. J R Stat Soc. 1985; 46B:1–52.
29. Lewis DD, Gale WA. A sequential algorithm for training text classifiers. ACM SIGIR. 1994:3–12.
30. Lewis DD, Schapire RE, Callan JP, Papka R. Training algorithms for linear text classifiers. Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval. ACM SIGIR. 1996:298–315.
31. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. J Am Med Inform Assoc. 1999:143–50.