# Using Narrative Reports to Support a Digital Library

**Eneida A. Mendonça, M.D., James J. Cimino, M.D., Stephen B. Johnson, Ph.D.**
**Department of Medical Informatics, Columbia University, New York, NY, USA**

*The vast amount of information collected and stored in clinical systems can be a significant challenge in the integration of digital libraries and electronic medical records, especially the selection of clinical data to be used in the search, retrieval, and summarization processes. In this study, we describe the use of information retrieval measures with natural language processor output to identify critical information in narrative reports. Our hypothesis is that clinical data that occur often in narrative reports are less important to clinicians than findings that occur rarely. We used the information retrieval methods to analyze one year of discharge summaries. We then conducted a performance study, using physicians as subject. Results show that the methods can be used for filtering critical information from reports. Further studies need to be done on evaluation of the method based on an evaluation of the system performance in the context of a digital library.*

## INTRODUCTION

Digital libraries have been described by many as a "new way of carrying out library functions", encompassing new types of information resources, approaches to acquisition, methods of storage and preservation, approaches to classification and cataloging, modes of interaction, and more reliance on electronic systems and networks.[1] In the health care environment though, digital libraries should provide more than just access to different literature resources. The integration of digital library resources and clinical information systems can be an important step in the effective retrieval of scientific evidence, especially evidence that is personalized based on the context of individual characteristics.[2,3] Several applications have been developed with this intent. These applications have varied from a simple integration between clinical and bibliographic systems, allowing the user to access the retrieval system and select the desired information to be retrieved from the clinical system (e.g. Medical Desktop[4], and Meta-1 Front End[5]), to more complex systems, which use the patient record or clinical reports to anticipate the user's needs (e.g. Hepatopix[6], Psychtopix[7], Chartline[8], IQW[9], the Medline Button[10], and Infobuttons.[11]). However, the development of personalized access to a distributed digital library is not a straightforward task. One of the many challenges is to understand what information in the individual medical record is important to the user and therefore potentially useful in the search, retrieval, and summarization processes. Ranking an individual's data according to the clinical relevance may be an interesting solution to the problem.

Natural language processing techniques have been used to analyze free text reports in order to provide data for applications, such as automated encoding, decision support, patient management, quality assurance, outcomes analysis, and clinical research.[12-19] Data mining and knowledge discovery techniques have been used to interpret data from natural language processing output of narrative reports.[20]

In this paper, we describe the use of an information retrieval method, document frequency thresholding, to identify critical information in narrative reports. The method is based on the assumption that there is no user relevance assessment a priori about what renders a finding (or term) relevant or not relevant. Therefore, instead of concentrating on the relevance of particular terms, it considers the occurrence of terms in complete document collections. Document frequency is commonly used to index documents for information retrieval systems. The method has also been used in automatic feature selection methods for removal of uninformative terms.[21] The basic assumption in these processes is that rare terms are either non-informative, or not influential in global performance. The method is then compared to other information retrieval methods: inverse document frequency and TF*IDF weighting.

The hypothesis of our experiment is based on the assumption that rare terms in clinical reports have significant relevance to patient care. We expect, for instance, that the presence of systemic lupus erythematosus would be more significant than the juxta-position of terms that occur frequently, e.g., cutaneous rash, woman, and pain. We evaluated the possibility of using text reports, more specifically, discharge summaries to identify information of clinical importance in patients' medical records. In other words, our goal is to identify terms that are good discriminators of medical records. This information can then be used to feed information retrieval systems in order to make them more effective.

| | | |
|---|---|---|
| findings: | demographics | |
| | age | 35 year |
| | section | report summary |
| | sex | female |
| problem: | pain | |
| | bodyloc | abdomen |
| | region | right upper quadrant |
| | numunit | several month |
| | sectname | report summary item |
| | status | history |

**Figure 1**. Partial view of MedLEE putput for the sentence "This is a 35 year-old woman who presented with a several month history of right upper quadrant abdominal pain."

**Table 1.** Formulas used in this study.[22] (N = total number of reports, *tf* = number of occurrences of term *i* in document *j,* and $df_i$ = number of documents in the collection that term *i* occurs in)

$$tf_{i,j} = 1 + \log(tf)$$

$$idf\text{i} = \log(N/df\text{i})$$

$$weight_{(i,j)} = tf_{i,j} \cdot idf\text{i}$$

## METHODS

The MedLEE system[14], a general natural language processor, was run on a set of all discharge summary reports in a one year period (1998). Discharge summaries were chosen because of the breadth of the information contained in them (past and present histories, procedures, prescriptions, hospital course, etc.). The processor works basically as follows. The text report is fed into a preprocessor, which identifies terms and phrases, maps them to standard terms, and assigns them a semantic type. The parser uses a grammar to identify the structure of the sentence, based on semantic rules, and generates a structure that consists of findings (e.g., symptoms, diseases, procedures, medications) and descriptive modifiers, such as certainty, body location, degree, status, etc. (Figure 1) The parsed reports constituted the training set. To make the data useable by the information retrieval method, we flattened each report into a vector of findings. We performed this flattening by combining findings (e.g., pain) and modifiers (e.g., body location), and considering each finding-modifier a separate attribute. Findings were also considered separately without modifiers. We computed document frequency and inverse document frequency for each unique attribute in the training set. Document frequency $df_j$ is the number of documents in which a unique term $t_j$ occurs. Inverse document frequency *idfi* is the logarithm ratio of the total number of documents to the numbers of document in which term $t_j$ occurs. (Table 1)

We then performed a pilot study in order to assess the validity of the information retrieval method to identify findings of "high importance" in the patients' records. The test set consisted of 10 discharge summaries randomly selected from the text reports of a different year than the training set (1999). These reports were parsed by MedLEE, and the attributes identified. For each attribute in the documents, we computed the term frequency and the TF*IDF weighting. Term frequency $tf_{i,j}$ is a measure of the number of occurrences of a term $t_i$ in a document *j*. TF*IDF weighting $weight_{(i,j)}$ is the product of the $tf_{i,j}$ and the *idfi* measures. (Table 1) The subjects were three physicians. A brief explanation of the project was given to the subjects prior to the study. The attributes were then presented to each physician along with a brief description of the case. All physicians received the same reports to review. Study subjects classified each attribute as "very important", "somewhat important", or "not important" for the task of filtering clinical data.

We performed a descriptive analysis to measure performance. We use two reference standards based on physicians' answers: a) majority of physicians judged the attribute "very important", b) all physicians judged the attribute as having some importance. For the performance of the DF measure, different threshold frequencies $T_f$ were selected based on the descriptive analysis. A term $t_i$ was identified as positive if $df_i < T_f$. Sensitivity, specificity, and positive predictive value were calculated for the different thresholds. An estimate of the area under the ROC curve was then computed using the non-parametric *A´* statistic proposed by Pollack and Norman.[23] Bootstrapping was used to estimate the variance of this average *A´* measure. We plotted the sensitivities and specificities using receiver-operator characteristic (ROC) curve axes. To measure performance of the TF*IDF weighting method, different threshold weights $T_w$ were defined based on the distribution of all weights in the testing set. A term $t_i$ was identified as positive if $weight_{(i,j)} > T_w$.

## RESULTS

During the year 1998, 28,832 discharge summaries were entered in the clinical database at Columbia-Presbyterian Medical Center. These reports were used as the training set. The parsing and flattening processes of these reports identified 379,709 unique attributes. The document frequency varied from 1 to 26560 (mean 19.45, sd 212.00).

**Table 2.** Estimate of the ROC area

| | ROC area A'(95% CI) | |
|---|---|---|
| | DF | TF*IDF weighting |
| Majority of physicians judged the attributes very important | 0.66 (0.55 to 0.78) | 0.66 (0.56 to 0.75) |
| All physicians judged the attributes to be at least somewhat important | 0.71 (0.61 to 0.82) | 0.76 (0.66 to 0.86) |

In the test set, 1,214 unique attributes were identified. From those 843 (69.43%) occurred in less than 10% of the documents in the training set ($df_i$ < 2,883). Eighteen threshold frequencies $T_f$ were defined. Because of the high number of attributes with ($df_i$ < 2,883), we used thresholds with 1% increments up to 10% of the document number, and then with additional increments of 10% (thus, the increments were 1%, 2%,..., 10%, 20%,..., 100%).

The DF method performed well when discriminating important attributes. When studying individual physicians' answer, there was no significant difference among physicians' answers on filtering "very import" and "unimportant" attributes. When classifying "somewhat important" physician 1 did not differ from the others, but there was a significant difference between physicians 2 and 3 (p=0.026).

The sensitivity and specificity based on the two reference standards are plotted in Figure 2. Table 2 lists the areas under the ROC curves and confidence intervals. The inverse document frequency measure performed similarly.

TF*IDF weighting varied from 1.03 to 7.10 (mean 2.99, sd 1.28). Seventeen threshold $T_w$ points were defined. The performance of the TF*IDF measure was slightly better than the performance using document frequency. When studying individual physicians, there was a significant difference between physicians 2 and 3 (p=0.05) when classifying "somewhat important" attributes. Table 2 lists the areas under the ROC curves and the confidence intervals for this measure. The sensitivity and specificity based on the two references standards are plotted in Figure 2.

## DISCUSSION

Researchers have suggested that the integration of clinical systems with library resources can improve the access to scientific evidence by personalizing information retrieval based on the context of individual characteristics.[2,3] To be able to do this in an automated way, information should be found in the electronic medical record that identifies characteristics of individual patient.

The primary focus of this experiment was to explore the use of information retrieval methods to identify critical information in narrative reports. In single-term indexing theories, document frequency and inverse document frequency are based on the observation that a *less* frequently occurring term has *better* discriminating properties than a term that occurs *more* frequently. The term frequency measure is based on the observation that a high occurrence of a concept in a particular document indicates that the corresponding document is closely related to that concept. TF*IDF combines the two observations in a single measure. A high TF*IDF weight suggests that a term is a good descriptor.

Narrative reports (discharge summaries, radiology reports) contain an enormous amount of information that can be useful to this task if well classified. Our hypothesis in this experiment was supported by the results. Physicians tended to judge very important or somewhat important concepts that occur less frequently. The TF*IDF method performed slightly better when we considered an attribute positive if all physicians had judged that attribute as at least of some importance.

We believe that the sensitivities and specificities for the classification of important attributes were sufficient to encourage the use of the methods. Further evaluation needs to be done in order to understand the behavior of the methods when integrated to the electronic medical record and the digital library system. The definition of the threshold to be used will depend on the evaluation of the system performance using such measures. The use of frequency measures integrated in addition to semantic algorithms, and a knowledge base to support a digital library are being investigated[24]. The association of this method with some form of machine learning algorithm may also be appropriate.

One limitation of our approach is the use of only one type of narrative report. Most of the discharge summary reports contain a description of the patient's admission, past personal and family history, medications, hospital course, and follow up plan. However, the quality, quantity, and granularity of the

information may vary depending on who writes the notes, their clinical experience, and the time they have available to perform the task.

Our method would prove more interesting if it can be verified in other domains, such as radiology, which is quite large and divided into modalities (e.g., nuclear scanning, computerized axial tomography, and magnetic resonance imaging) and anatomic subdomains (e.g., abdomen, musculo-skeletal, neurological, etc.). Other domains include echocardiography, electrocardiography, and pathology. A broad use of narrative reports may illustrate the medical record data more precisely.

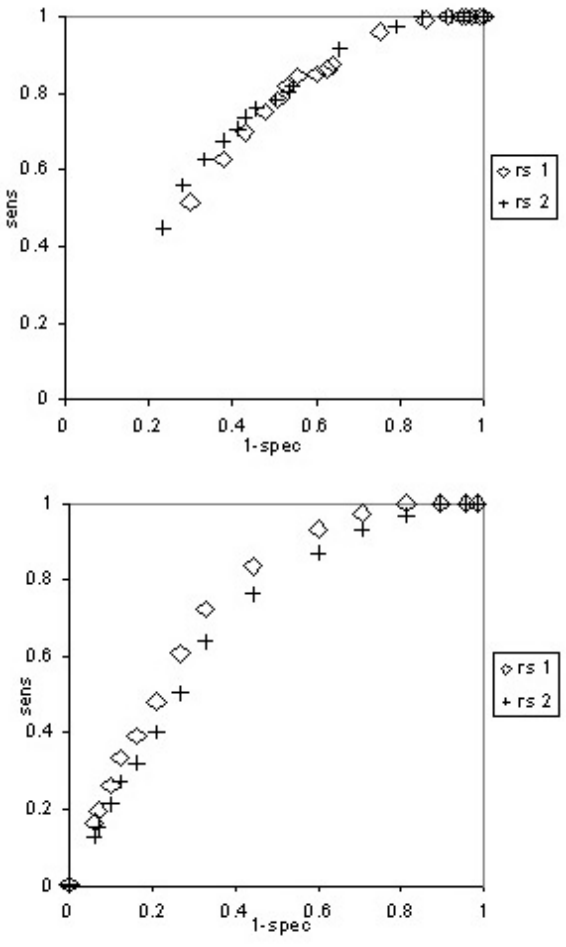


**Figure 2.** Sensitivity and specificity plotted on ROC axes. Graph 2A presents the data for document frequency. Graph 2B presents the data for TF*IDF weighting. (rs 1 = majority of physicians, rs 2 = all physicians judged the attributes of at least somewhat importance).

Another limitation of this study may be the small number of cases in the testing set. Each physician analyzed 10 cases. A larger set of cases would cover the possible spectrum of narrative reports and, perhaps, demonstrate a significant effect of TF*IDF weighting.

## CONCLUSION

We believe that it is possible to use information retrieval measures for filtering large amounts of clinical data on important findings. These data can be used to enhance the retrieval of literature pertaining to individual characteristics. Further studies need to be done on the validation of the method based on an evaluation of system performance within the context of a digital library.

### Acknowledgments

The authors thank Adam Wilcox for his assistance in the retrieval of text reports from the data warehouse. This publication was supported in part by National Science Foundation grant IIS-98-17434, by the Columbia Center for Advanced Technology in Information Management supported by the New York State Science and Technology Foundation, and by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil, grant 20057/95-5.

References

1. Fox EA, Akscyn RM, Furuta RK, Leggett JJ. Digital libraries. Communications of the ACM 1995; 38(4):23-8.
2. Cimino JJ. Linking patient information systems to bibliographic resources. Methods of Information in Medicine 1996; 35(2):122-6.
3. Hersh W. "A world of knowledge at your fingertips": the promise, reality, and future directions of on-line information retrieval. Acad Med 1999; 74(3):240-3.
4. Loonsk JW, Lively R, TinHan E, Litt H. Implementing the Medical Desktop: tools for the integration of independent information resources. Clayton PD. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care. 1991: 574-7.
5. Powsner SM, Miller PL. From patient reports to bibliographic retrieval: a Meta-1 front-end. Proc Annu Symp Comput Appl Med Care 1991; 526-30.
6. Powsner SM, Riely CA, Barwick KW, Morrow JS, Miller PL. Automated bibliographic retrieval based on current topics in hepatology: hepatopix. Computers and Biomedical Research 1989; 22(6):552-64.

7. Powsner SM, Miller PL. Automated online transition from the medical record to the psychiatric literature. Methods of Information in Medicine 1992; 31(3):169-74.
8. Miller RA, Gieszczykiewicz FM, Vries JK, Cooper GF. CHARTLINE: providing bibliographic references relevant to patient charts using the UMLS Metathesaurus Knowledge Sources. Frisse ME. Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care. New York: McGraw-Hill, 1992: 86-90.
9. Cimino C, Barnett GO, Laboratory of Computer Science - Massachusetts General Hospital. Standardizing access to computer-based medical resources. Miller RA. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care. Washington, D.C.: 1990: 33-7.
10. Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD, author. The MEDLINE Button. Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care. 1992: 81-5.
11. Cimino JJ, Elhanan G, Zeng Q. Supporting infobuttons with terminological knowledge. Masys DR. Proceedings/AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, 1997: 528-32 .
12. Sager N, Lyman M, Nhan NT, Tick LJ. Medical language processing: applications to patient data representation and automatic encoding. Methods Inf Med 1995; 34(1-2):140-6.
13. Baud RH, Rassinoux AM, Wagner JC *et al*. Representing clinical narratives using conceptual graphs. Methods Inf Med 1995; 34(1-2):176-86.
14. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association 1994; 1(2):161-74.
15. Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic reports. Work in progress. Radiology 1990; 174(2):543-8.
16. Gundersen ML, Haug PJ, Pryor TA *et al*. Development and evaluation of a computerized admission diagnoses encoding system. Computers and Biomedical Research 1996; 29(5):351-72.
17. Friedman C, Knirsch C, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. Proc AMIA Symp 1999; 256-60.
18. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Annu Fall Symp 1997; 829-33.
19. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000; 7(6):593-604.
20. Wilcox A, Hripcsak G. Medical text representations for inductive learning. Proc AMIA Symp 2000; 923-7.
21. Yang Y, Pedersen JO. A comparative study of feature selection in text reports. Proceedings of the Fourteeth International Conference on Machine Learning . Morgan Kaufmann, 1997: 412-20.
22. Manning CD, Schütze H. Topics in Information Retrieval. Foundations of Statistical Natural Language Processing. Fourth edition. Cambridge, MA: The MIT Press, 2001: 529-74.
23. Pollack I, Norman DA. A non parametric analysis of recognition experiments. Psychonomic Science 1964; 1:125-6.
24. Mendonça EA, Cimino JJ, Johnson SB, Seol YH. Accessing Heterogeneous Sources of Evidence to Answer Clinical Questions. J Biomed Inform 2001.