

# Preliminary Work on Building a User Friendly Adaptive Clinical Documents Repository

Enriko Aryanto  
Stanford University  
121 Campus Dr. #3112A  
Stanford, CA 94305  
1-650-497-7306  
earyanto@stanford.edu

Yang Huang  
Stanford Medical Informatics  
X-215, 251 Campus Dr.  
Stanford, CA 94305  
1-650-725-6699  
huangy@stanford.edu

## ABSTRACT

This paper describes preliminary work on the experiment of building a user-friendly adaptive clinical document repository and the evaluation of the possible improvements on automated concept indexing using the feedbacks from users. The authors used the SAPHIRE indexing system to automatically identify related biomedical concepts in the clinical documents, radiology reports in this experiment. These concepts were represented as terms in the 2001 edition of the U.S. National Library of Medicine's Unified Medical Language System (UMLS) Metathesaurus. The results from this automated indexing then served as a baseline to be improved by the users of the system. Limited feedbacks on a subset of the documents were collected and the system automatically propagated revised scores of concepts to the whole set of reports. The revised indexing was then compared to the baseline using results from an independent manual indexing as the gold standard. Significant improvement was seen in the results with cutbacks in a few cases. Further experiments may be needed to find better score revise mechanism to improve the results in even more cases. At the same time, a user-friendly report review system was built to collect feedback from users of this system.

## General Terms

Algorithms, Performance, Design, Experimentation.

## Keywords

Clinical Information System, Information Retrieval, Relevance Feedback, User Interface Design.

## 1. INTRODUCTION

As the adoption of Electronic Medical Record (EMR) system in hospitals and healthcare organizations is growing, more and more organizations have large clinical document collection stored electronically. This opens new opportunities for physicians and researchers to carry out studies of medical records in large scale. As a core function of the medical document repository, the documents need to be well indexed to support effective and efficient search and retrieval.

MEDLINE [1] is a good example of a successful information retrieval system in biomedicine and health domain. All citations

To make digital or hard copies of all or part of this work for personal or classroom use or commercial use requires prior specific permission from the authors.

and abstracts in MEDLINE were manually indexed using a controlled vocabulary, Medical Subject Headings (MeSH) from National Library of Medicine (NLM). The indexing work was regarded as of very good quality; however, the manual indexing approach is not feasible to deal with the large amount of clinical documents for other organization and hospitals.

A large portion of clinical documents exists in the form of unstructured or semi-structured free text. Radiology report is one example. Tens of thousands of radiology reports are stored electronically. However, there is no effective way of doing query and retrieval based on the contents of report. It is especially difficult to do for the image part of the radiology reports. On the other hand, some interesting research work [2][3] have been done on the text part of the report. To enable the query and retrieval of the free text of the reports, they need to be properly indexed. Thus, some automatic indexing engines were developed to index clinical documents or medical journals with NLM's Unified Medical Language System (UMLS) [4][5] Metathesaurus. Previous studies show UMLS could capture more than 80% of biomedical concepts used in literatures [2]. SAPHIRE [6] is such an indexing engine developed by Dr. William Hersh at Oregon Health Sciences University. Study shows that such an automatic indexing engine could provide decent recall but suffer from low precision in indexing radiology reports.

Patient records are used daily by well-trained physicians and researchers to support clinical care, medical research and medication. Current systems work very well in retrieving the records of a specific patient or a set of patients identified by the metadata. However, they can make very limited use of the free text part of the records. The content of a clinical document repository is relatively stable, with daily add-ins for new documents. While manually index all clinical documents is prohibitively expensive, physicians' feedback on a portion of them is a by-product from their normal workflow. To know if a document repository can adaptively improve its indexing on all documents based on feedbacks on a portion of them, we developed a prototype to index radiology reports using an automatic UMLS indexing engine and then revise the ranking of indexed terms adaptive to physicians' relevance feedback.

The evaluations of relevance feedback [7][8] usually take two approaches. Either include the results judged as relevant in the evaluation or exclude those results. Both of them are biased. The first approach is overly optimistic since it counts help from

human in the result. The second approach is overly conservative since those good results returned by the system are excluded.

We took a different approach. In our evaluation, the collection was arbitrarily split into a training set and a test set. The relevance feedback is only available on the training set. The system took the feedback and revised the index rankings of documents in both sets. The evaluation was done only on the test set to show unbiased results.

## 2. DESIGN AND IMPLEMENTATION

Compared to the Web, clinical document repository has well-controlled collections. While the Web search engines can target at very high precision at significant cost of recall [9], a clinical information retrieval system needs a good balance of precision and recall for research and education purpose. On the other hand, the contents of such a repository is much less diversified in formats and semantics, and the users of such a system is much better well-trained and have a much more focused use of the system. Thus, the relevance feedbacks from physicians should carry more weights and should be more consistent with each other.

The goal of such a system is to provide organizations a scalable and reliable repository for clinical documents. The system provides a friendly web user interface help physicians inspect document more efficiently and give less disturbance to their normal workflow. By leveraging users' intelligence, the system should improve its indexing of documents constantly and provide users powerful retrieval capabilities with good retrieval precision and recall.

The overall architecture can be seen in Figure 1. The first part on the left side is a parser. The parser accepts different types of clinical documents in semi-structured text format and converts them to well-formed XML documents. This step enables latter use of metadata and more powerful queries with more granularity than treating all the contents as one unit. The parsed documents are stored in database in XML format and handed over to the indexing engine.

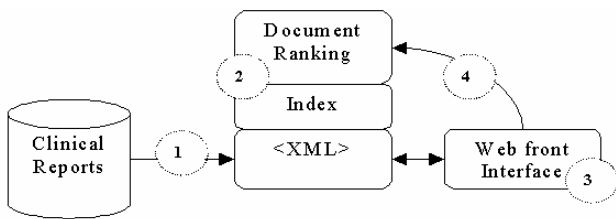


Figure 1. Overview of the document repository

The indexing engine is not fully completed at the time of this report. We used indexing results from SAPHIRE for the purpose of the experiment described later. The indexing engine is supposed to takes sections of documents and index them using terms from UMLS. The lexicon is based on NLM's Specialist Lexicon. Stemming and normalization of words are done by the Lexical Variant Generator package, which comes with the UMLS Metathesaurus 2001. The indexing results are stored in the database for later retrieval and improvement.

A web interface running on JSP-Servlet environment is designed to give users a web front end to work with indexed documents. As an example, a radiology report is parsed into up to five sections. The user is given the choice to select a specific report and section for inspection. After typing into the selection, the page with corresponding report and section is brought up along with the concepts that are assigned to it. If the number of concepts is 50 or less, all of them will be displayed. If that number is between 51 and 150, inclusive, the top 50% concepts (by score) will be displayed. And if that number is greater than 150, only the top 34% concepts are displayed. The control of the number of displayed concepts is mainly for usability reason. The low scored concepts are the ones less likely to be relevant. This cut off technique increases feedback efficiency while not sacrificing a great deal of valuable information. Also, after scores are updated based on feedbacks from one user, those concepts not displayed first time will have scores higher than those judged as irrelevant ones from previous inspection. Thus, originally low scored one will come out in later inspections.

Relevant?	Concept Name	Score	Original Phrase
<input type="checkbox"/>	renal adenocarcinoma	11.729	renal cell carcinoma
<input type="checkbox"/>	Sex Female	5.000	female with history
<input type="checkbox"/>	History <1>	.500	female with history
<input type="checkbox"/>	Medical History	.500	female with history
<input type="checkbox"/>	IN <2>	.500	female with history
<input type="checkbox"/>	RENAL PELVIS EPIDERMOID CARCINOMA	.500	renal cell carcinoma
<input type="checkbox"/>	KIDNEY, PELVIS, PAPILLOMA, TRANSITIONAL CELL	.500	renal cell carcinoma
<input type="checkbox"/>	renal cell carcinoma, metastatic	.500	renal cell carcinoma
<input type="checkbox"/>	Female XLA	.500	female with history
<input type="checkbox"/>	Female Organic Disorder	.500	female with history
<input type="checkbox"/>	female pseudo-Tumor syndrome	.500	female with history
<input type="checkbox"/>	Transsexuality with asexual history	.500	female with history
<input type="checkbox"/>	Transsexuality with homosexual history	.500	female with history
<input type="checkbox"/>	Transsexuality with heterosexual history	.500	female with history
<input type="checkbox"/>	Chiropractic consultation with history	.500	female with history
<input type="checkbox"/>	Phage with short tails	.500	female with history
<input type="checkbox"/>	renal cell carcinoma, recurrent	.500	renal cell carcinoma
<input type="checkbox"/>	Clear cell carcinoma of kidney	.500	renal cell carcinoma
<input type="checkbox"/>	granular cell carcinoma of the kidney	.500	renal cell carcinoma
<input type="checkbox"/>	stage, renal cell carcinoma	.500	renal cell carcinoma
<input type="checkbox"/>	Renal cell carcinoma - morphology	.500	renal cell carcinoma
<input type="checkbox"/>	RENAL CELL CARCINOMA REMOVAL	.500	renal cell carcinoma
<input type="checkbox"/>	BRONCHIODOLAR ALVEOLAR CELL CARCINOMA	.500	renal cell carcinoma
<input type="checkbox"/>	Squamous cell carcinoma of bronchus	.500	renal cell carcinoma
<input type="checkbox"/>	[M]squamous cell carcinoma NOS	.500	renal cell carcinoma
<input type="checkbox"/>	[M]transitional cell carcinoma NOS	.500	renal cell carcinoma

Figure 2. Relevance Feedback Screen

As shown in Figure 2, the radiology report and section names are listed on the first line, and the actual section text is displayed on the next line. There is a link following the report section on the first line to pop up a new browser window that displays the complete report (instead of just one section) that is intended to help the feedback evaluators get more sense of the fitting report context. The next section is showing the top 50% (26/52) of assigned concepts, which is consistent to our cut-off formula. Every concept row has a number and a checkbox in front of it. Since neutral feedbacks are not allowed, the corresponding concept is marked as relevant if the checkbox is checked and as non-relevant otherwise. The text field right above the submit button is used to indicate the last concept number evaluated. In most cases, this field should just be leaved untouched, since its default value is the last concept number displayed on the page. However, if the evaluators just wish to evaluate the top n concepts, they will have to use this field because unchecked checkboxes will trigger automatic non-relevant feedback to the corresponding concepts.

When the submit button is pressed, a request is sent to the top-level servlet. This servlet will then update the concepts score appropriately (the score update criteria are discussed in detail on the next section). Note that concepts in other radiology reports with the same section might be updated as well. The web browser is then redirected to the next report section. When it reaches the end of the report, the browser is taken back into the starting page, which inquires for the next report and section for inspection.

Figure 3. Relevance Feedback Starting Screen

Every single user feedback along with the user identification (user's IP address in the testing system) is kept in the database for recovery purpose. It assures the concepts score consistency by making possible undoing the score updates to a certain point if some destructing feedback entries accidentally get into the system.

The results were presented to physicians for inspection and feedbacks were collected. We also developed a simple heuristic to use the relevance feedback to change the ranking of the terms assigned to each document. The algorithm was shown in pseudo code in Figure 4.

```

if judged as relevant {
    for the same section of this report {
        score = score + (0.5 * (MAX-score));
        if (n == -1) n = 1;
        else n++;
    }
    for the same section of other reports of this type {
        if (n != -1) {
            score = score + (float)Math.log(1.0f+score*(float)n);
            n++;
        }
    }
}
if judged as irrelevant {
    only for the same section of this report:
    n = -1;
    score = score - 1/3*(MAX-score);
    if (score < 0.5) score = 0.5f
}

```

Figure 4. Pseudo Code of Score Update Algorithm

What the algorithm does is to give bonus to terms judged as relevant and give penalty to those judged as irrelevant. If a term is

judged as relevant to a certain section in the report by the user, a bonus is given to bring its score toward the maximum score, but not exceeding it. The lower the original score, the higher the bonus since the system should respect the authority of the user's judgment and make the term stand out by this feedback alone. On the other hand, if a term is identified as relevant to this section, it may be semantic suitable to represent the similar section in other reports of this type. Thus, a small bonus is given to it to differentiate it from other terms, if it appears in the result list from automatic indexing. It is obvious that this bonus should not be linear. Thus, the system keeps track of how many times it got bonus, later on, it get bonus less and less because of the log function. And after a number of times of updates, all scores need to be renormalized to be smaller or equal to the maximum score.

On the other hand, if a term is judged as irrelevant, we only want to penalize it in the section of this specific report, since a term irrelevant to this section says almost nothing about whether it is relevant to another report. The penalty is smaller than the bonus given to it in case it is judged as relevant, since we think relevance feedback is more valuable. In case two users don't agree on a term, all we know is that this term might be relevant. Thus, we still want it to get some bonus to differentiate itself with other un-judged terms, though the bonus is much lower than a term judged as relevant by both.

### 3. THE EXPERIMENT

The document set for this experiment consisted of forty-nine radiology imaging reports, chosen to represent the most imaging modalities (10 conventional chest X-ray, 10 head CT scans, 9 chest CT scans, 10 abdomen-pelvis CT scans, 5 head MRI and 5 whole body radio-nuclide bone scans). Each report has up to five sections, Procedure, History, Technique, Findings and Impression. These reports had been previously manually de-identified and then manually indexed using UMLS Knowledge Source Server (KSS) [10] by two domain experts knowledgeable about UMLS. The results of the manual UMLS indexing served as the gold standard in this experiment. This process has been described elsewhere [2].

The parser handed 230 sections of 49 reports to SAPHIRE to index. The indexing results from SAPHIRE indexing engine baseline algorithm [3] was used as initial indexing results from an automatic indexing system. SAPHIRE takes a lexical approach to match document text to UMLS terms. It generated a large number of concepts with different rankings since it allows partial match to boost retrieval recall. The ranking is given based upon how well the text in the reports match to the phrase of the concept, called concept name.

The results from above are saved in databases for users' inspection and use. We released our system to be tested by two physicians. However, due to time constraint and schedule conflicts, they were not able to give enough relevance feedbacks on selected reports. So we simulated users' use of this system by comparing the results from SAPHIRE against the results of gold standard directly. Terms listed as hits by both are judged as relevant ones, while those generated by SAPHIRE only are judged as irrelevant ones.

We randomly selected 7 of 10 conventional chest X-ray, 7 of 10 head CT scans, 7 of 9 chest CT scans, 7 of 10 abdomen-pelvis CT

scans, 3 of 5 head MRI and 3 of 5 whole body radio-nuclide bone scans as the training set. The remaining results are used as test set.

Relevance feedbacks of one round on the documents of training set were collected and the system automatically propagates revised scores of terms to the whole set of reports. The algorithm is described in the code in figure 4.

The revised indexing was then compared to the original results from automatic indexing using results from an independent manual indexing as the gold standard. The performance was measured in extrapolated average eleven-point precision.

#### 4. RESULTS

The result of above experiment is listed in table 1. It listed the average change in eleven-point precision of five different types of sections of fifteen radiology reports in the test set mentioned above. The second column is the average change in percentage. The last three columns shows the number of reports in test set which has a positive change, no change and negative change in eleven-point precision respectively.

Section Type	Average Change	Positive Change	Zero Change	Negative Change
PROCEDURE	246%	11	4	0
HISTORY	178%	8	7	0
TECHNIQUE	94%	5	7	0
FINDINGS	195%	12	0	3
IMPRESSION	212%	9	4	1

Table 1. Change in eleven-point precision of test set

Significant improvements have been seen in the above results with cutbacks in a few cases. The four section types except TECHNIQUE got improvements averaged about 200%. Also, more than half of sections in the test set benefited in the adaptive system.

TECHNIQUE section is usually short and some reports don't have this section. So the smaller improvement may due to inadequate amount of this type of sections in the training set.

There are also four cases in sections of FINDINGS and IMPRESSION, in which revised version performs worse than the original results. It was caused solely by bonuses given to some irrelevant terms. When the original scores of true relevant terms are not high enough, they can be topped by irrelevant terms with bonus. It implies that the bonus we gave for other similar sections of same report type is too high. It is understandable since these two sections are generally longer; the number of terms generated by the automatic UMLS indexing engine is larger. Thus, the possibility of getting bonus for the terms in these two sections is larger. The same bonus amount working for the other three sections may not fit very well with these two sections.

#### 5. DISCUSSIONS AND CONCLUSIONS

First, the encouraging result of this pilot study is by no means conclusive. Due to the concern of privacy protection and hospital regulation, we could only carry out the test over a small set of radiology reports in the project. A larger collection is definitely needed to test the algorithm, which impacts the result greatly.

Also, other types of clinical documents should be tested to see if the conclusion still holds.

Second, a clinical document repository is well controlled in content generation and daily use. It targets at physicians and researchers. Thus, the quality of relevance feedback can be guaranteed. This partially justified our decision to carry out a simple simulation. However, it is very possible for users of the system to give inconsistent judgment. Our score update algorithm made some effort to deal with it, however, it is mostly untested.

Third, the radiology reports are parsed and indexed by sections, thus, more powerful query can be done on better granularity. It is different from just indexing the whole report regardless the different semantics of different sections.

The reason that the system worked well partial lies in that different sections and different types of reports may have their own set of terms with suitable semantics. Terms proven to be relevant to one section is likely in this set and thus may be more likely to relate with another section as long as the terms are judged as relevant by the automatic UMLS indexing engine as well. However, under current algorithm, the most common terms get more bonuses than unique terms, which are actually more informative terms. This should be corrected by including the document frequency in the term weighting, which was not done partially because of the small collection size. The bottom line is that the algorithm should not rank common terms above those more informative however less frequent terms.

The above mechanism does not change the precision and recall if no cut-off score is used. Its strength lies in improving the precisions with different cut-off scores. So it is more likely to work with indexing engines with high recalls. To improve the recall as well, the feedback system should allow users to add in new relevant terms.

The two physicians participated in the test of the system confirmed the usability of this system. They also gave positive response to the user interface design.

#### 6. ACKNOWLEDGMENTS

We would like to thank Dr. Henry Lowe for providing us the document collection and many good discussions with one of the authors. Also, we want to thank Dr. Todd Ferris and Dr. Greg Garrison for their time in the evaluation of the system.

#### 7. REFERENCES

- [1] Fact Sheet MEDLINE®  
<http://www.nlm.nih.gov/pubs/factsheets/medline.html>.
- [2] Lowe HJ, Antipov I, Hersh W, Smith CA, Mailhot M. Automated Semantic Indexing of Imaging Reports to Support Retrieval of Medical Images in the Multimedia Electronic Medical Record. *Method Inform Med* 1999; 38: 303-7.
- [3] Hersh WR, Mailhot MF, Lowe HJ, Smith CA. Selective Automated Indexing of Findings and Diagnoses in Radiology Reports. *Journal of Biomedical Informatics*, Vol 34, No. 4, August 2001, pp 262-273
- [4] Lindberg DA, Humphreys BL, Mc Cray AT.

The Unified Medical Language System. *Method Inform Med* 1993; 32: 281-91

- [5] Humphreys, B., et al. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc* 1998; 5:1-11
- [6] Hersh WR, Leone TJ. The SAPHIRE server. *Proceedings of the 19<sup>th</sup> Annual Symposium on Computer Applications in Medical Care* 1995; 858-62.
- [7] Buckley, Singhal, Mitra, Salton, New retrieval approaches using smart: trec4, nist, 1996.
- [8] Salton G, Buckley C. Improving retrieval performance by relevance feedback. *Journal of the American Society of Information Science*, 41(4):288-297, 1990.
- [9] Brin, S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 1998.
- [10] McCray AT, Razi AM, Bangalore AK, Browne AC, Stavri PC. The UMLS Knowledge Source Server – a Versatile Internet-Based Research Tool. *Proc AMIA Fall Symp* 1996; 164-8.