Journal of Digital Imaging

Effective Metadata Discovery for Dynamic Filtering of Queries to a Radiology Image Search Engine

Charles E. Kahn Jr.

We sought to demonstrate the effectiveness of techniques to index radiology images using metadata discovered in their free-text figure captions. The ARRS GoldMiner™ image library incorporated 94,256 figures from 11,712 articles published in peer-reviewed online radiology journals. Algorithms were developed to discover metadata—age, sex, and imaging modality—from the figures' free-text captions. Age was recorded in years, and was classified as infant (less than 2 years), child (2 to 17 years), or adult (18+ years). Each figure was assigned to one of eight imaging modalities. A random sample of 1,000 images was examined to measure accuracy of the metadata. The patient's age was identified in 58,994 cases (63%), and the patient's sex was identified in 58,427 cases (62%). An imaging modality was assigned to 80,402 (85%) of the figures. Based on the 1,000 sampled cases, recall values for age, sex, and imaging modality were 97.2%, 99.7%, and 86.4%, respectively. Precision values for age, sex, and imaging modality were 100%, 100%, and 97.2%, respectively. Automated techniques can accurately discover age, sex, and imaging modality metadata from captions of figures published in radiology journals. The metadata can be used to dynamically filter queries for an image search engine.

KEY WORDS: Information retrieval, metadata, knowledge discovery, image library, information filtering, abstracting and indexing, search engine

INTRODUCTION

The ARRS GoldMiner™ search engine facilitates access to web-based, peer-reviewed radiological images.[1] To allow users to refine their searches, we sought to discover metadata—specifically, age, sex, and imaging modality—from the images' textual captions. These metadata allowed users to dynamically filter their searches to identify specific subsets of images. The current work assesses the accuracy and effectiveness of techniques to index radiology images using metadata discovered in their free-text figure captions.

MATERIALS AND METHODS

The ARRS GoldMiner™ image library incorporated 94,256 figures and their captions from 11,712 articles published in five online radiology journals and one e-learning web site.[1] All of the articles had undergone peer review and were made available by their publishers for "open access" to the general public. The image library captured figures as low-resolution JPEG "thumbnail" images that linked to the full-size images at the source web sites. The accompanying textual captions were mapped to concepts in a subset of the UMLS® Metathesaurus using MetaMap Transfer (MMTx) software (U.S. National Library of Medicine, Bethesda, MD). The images were indexed by these concepts and by textual keywords.

Patient age was parsed from caption text containing a phrase such as "43-year-old" or "10 months old." Age units of years, months, and days were recognized. The age in years was computed, and images were classified as infant (less than 2

From the Division of Informatics, Department of Radiology, Medical College of Wisconsin, 9200 W. Wisconsin Ave., Milwaukee, WI, 53226, USA.

Correspondence to: Charles E. Kahn Jr., Tel: +1-414-8052173; Fax: +1-414-2599290; e-mail: kahn@mcw.edu

Online publication 9 June 2007
doi: 10.1007/s10278-007-9036-5

years), child (2–17 years), or adult (18+ years). Entries were classified by sex using matching to a set of words; for example, an image's subject was classified as female if the textual caption contained the word "woman", "girl", or "female." Because all EURORAD articles are case reports, we identified the age and sex of subjects in their images from information in the body of the articles.

Images were assigned to one of eight modalities: computed tomography (CT), graphic (a chart or illustration), magnetic resonance (MR), nuclear medicine, positron emission tomography (PET), photography (optical imaging, including endoscopy and photomicrography), ultrasonography, and radiography. Assignment was made using a set of string-matching patterns. If more than one match was identified, the first (leftmost) matching term was used.

We explored the database entries for which age, sex, and/or imaging modality were identified. A random sample of 1,000 entries was reviewed manually to assess the accuracy of the assignments. The recall and precision of each of the values of age, sex, and imaging modality were computed. Recall is the fraction of all relevant documents in the database that have been retrieved; it measures how well relevant items are retrieved from the image library. Precision assesses the quality of the items actually retrieved: it measures the fraction of the retrieved documents that are relevant.[2]

RESULTS

Overall results

The patient's age was identified in 58,994 cases (63%); of those, 86% were adults (Table 1). The patient's sex was identified in 58,427 cases (62%); of those classified by sex, 53% were male (Table 2). In the entire database, entries were assigned an imaging modality based on their captions in 80,402 (85%) of cases. MR (26%), CT (24%), and radiography (13%) were the most frequently assigned modalities (Table 3). Recall and precision values are shown in Table 4.

Age

Of the 1,000 sampled cases, age was determined in 671 cases (67%), and all of these were assigned correctly. The 95% confidence interval [$CI_{0.95}$] for correct assignment, thus, has a lower bound of 99.9%. In five unclassified cases, the term "neonate" or "newborn" appeared, which would allow one to assign an age of 0 years. Several cases were unclassified because the age was expressed in weeks or in weeks of gestational age.

Table 1. Classification of Images by Age

Age Group	Number	Frequency (%)	Relative Frequency (%)
Infant (<2 years)	1,805	1.9	3.6
Child (2–17 years)	5,937	6.3	10.1
Adult (18+ years)	51,252	54.4	86.9
Total Classified	58,994	62.6	100.0
Unclassified	35,262	37.4	
Total	94,256	100.0	

The relative frequency expresses the percentage in each age group among those classified.

Sex

Six hundred sixty-seven (67%) of the sampled cases were classified by sex; all of these were assigned correctly. Thus, among those classified by sex, the $CI_{0.95}$ lower bound for correct classification is 99.9%.

Imaging Modality

Modality was assigned in 861 of the 1,000 sampled cases (86.1%); of these, it was assigned correctly in 829 cases (96.3%; $CI_{0.95}$, 95.0–97.6%). Through manual review of the caption text or the thumbnail image, the modality was assigned to 131 of the 139 initially unclassified entries. The errors in classification by imaging modality typically occurred because more than one imaging-modality term appeared in the figure caption. All

Table 2. Classification of Images by Sex

Sex	Number	Frequency (%)	Relative Frequency (%)
Male	30,809	32.7	52.7
Female	27,618	29.3	47.3
Total Classified	58,427	62.0	100.0
Unclassified	35,829	38.0	
Total	94,256	100.0	

The relative frequency expresses the percentage of each sex among those classified.

Table 3. Classification of Images by Imaging Modality

Modality	Number	Frequency (%)	Example text words
CT	22,458	23.8	CT; MDCT; computed tomography
Graphic	7,527	8.0	Diagram; drawing; bar chart
MRI	24,862	26.4	MR; T1-weighted; MRCP
Nuc Med	603	0.6	Nuclear medicine; scintigraphy; SPECT
PET	777	0.8	PET; positron; FDG
Photo	3,967	4.2	Photomicrograph; endoscopic image
US	7,782	8.3	US; Doppler; sonography
X-ray	12,426	13.2	Radiograph; X-ray; ERCP
Total Classified	80,402	85.3	
Unclassified	13,854	14.7	
Total	94,256	100.0	

Each image was assigned to one of eight imaging modalities, as itemized here. The "example text words" indicate some of the terms used to assign the imaging modality.

Table 4. Recall and Precision

Attribute	Value	Recall (%)	Precision (%)
Age Group	Infant (<2 years)	57.6	100.0
	Child (2–17 years)	92.6	100.0
	Adult (18+ years)	100.0	100.0
	All	97.2	100.0
Sex	Male	99.7	100.0
	Female	99.7	100.0
	All	99.7	100.0
Imaging Modality	CT	91.4	97.0
	Graphic	79.3	100.0
	MRI	91.9	99.2
	Nuclear Medicine	37.5	75.0
	PET	100.0	83.3
	Photo	65.7	100.0
	Ultrasound	91.0	92.2
	X-ray	79.9	96.1
	All	86.4	97.2

Effectiveness of information retrieval. The weighted average ("All") represents the mean value of recall and precision as weighted by the true number of cases in each category.

images classified as photographs (n=44) or graphics (n=73) were classified correctly.

Metadata by Source

The presence of discoverable metadata varied significantly by the source of each image (Table 5). In EURORAD, in which each article is a case report of an individual patient, the age and sex were available in all of the entries. From another source, only 13% of image captions had such information. Similarly, discovery of imaging modality ranged from 57 to 90% of cases, depending on the source.

User Interface

The user interface was designed to accommodate the filters. Each filter is implemented as a pull-down selection that shows the number of images for each available selection (Fig. 1). The

Table 5. Metadata by Figure Source

		Percentage of Images Classified		
Source	Number of Images	Age	Sex	Modality
AJR	33,721	85	85	90
American Journal of Neuroradiology	5,373	25	23	74
British Journal of Radiology	2,040	13	13	57
EURORAD	12,423	100	100	76
Radiology	20,024	36	34	82
RadioGraphics	20,675	45	44	88
Total	94,256	63	63	85

The ability to detect metadata varied significantly by source. For example, each article in EURORAD, the European Association of Radiology's e-learning resource, presented a single case and specified the subject's age and sex; however, imaging modality was identified in only 76% of image captions from this source.

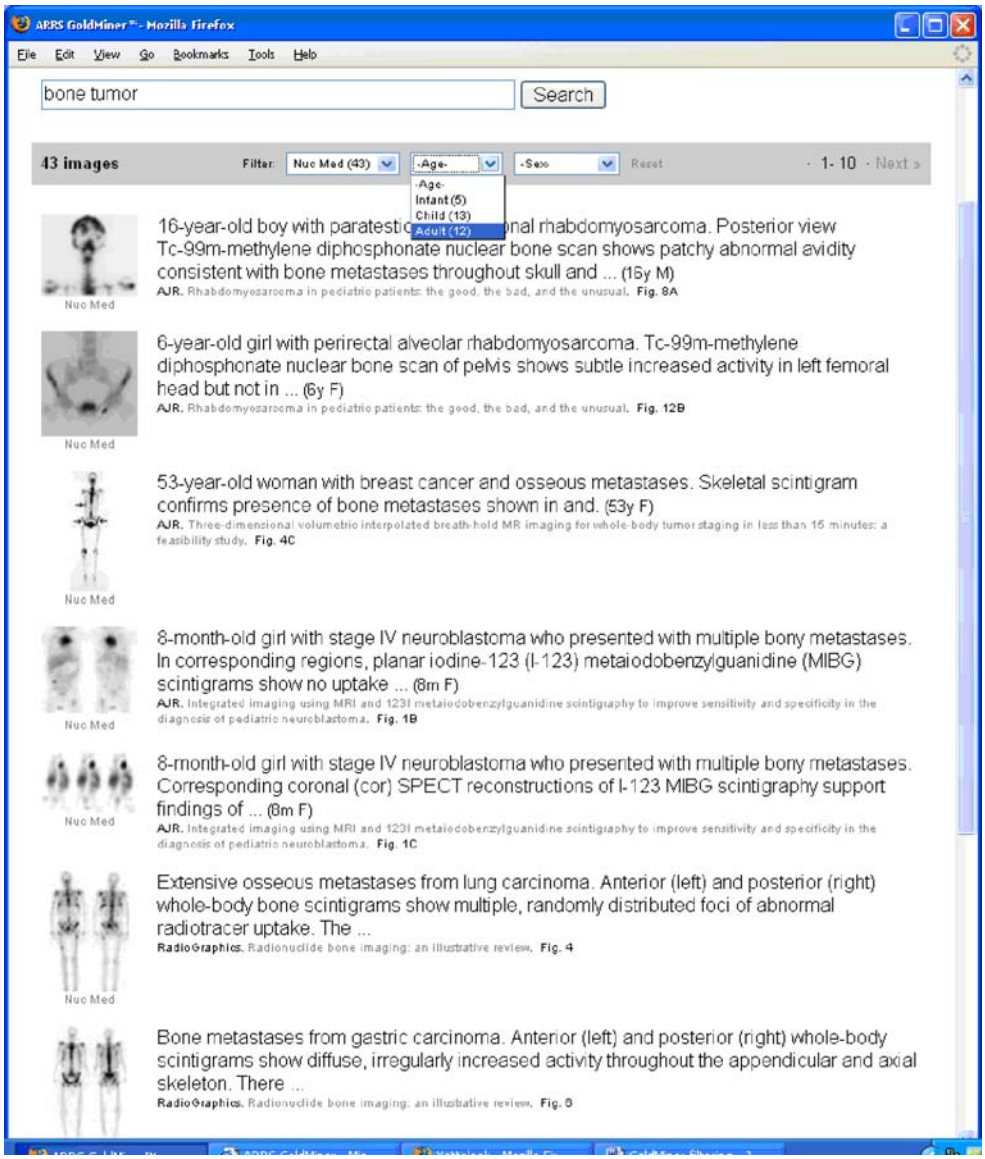

Fig. 1. The filters for imaging modality, age, and sex are seen in the grey bar at the top of the page; when the user clicks on one of these pull-down lists, the number of images in each category is shown. By using the imaging-modality filter, the user has limited the search to the 43 nuclear medicine ("Nuc Med") images from the 1,088 images retrieved in response to the query "bone tumor". The search can be refined further by filtering by age and/or sex. Here, the number of images in each age group is displayed. The filters can be applied in any order and modify the search results dynamically. The "Reset" link allows the user to reset all of the filters and return to the original, complete set of images that match the text query.

filters allow users to dynamically alter their search parameters.

DISCUSSION

Although much progress has been made, information retrieval remains a challenging task.[3] Searches can be aided significantly by the ability to modify or limit queries dynamically. In the current work, we explored the ability to discover and apply metadata about age, sex, and imaging modality from figure captions in peer-reviewed radiology journals. The metadata allowed useful and highly accurate filtering of concept- and keyword-based searches of a large radiology image library.

We plan to analyze those images that were not classified by modality to determine if there are

additional words or phrases that may help identify the imaging modality. Another potential approach that might improve the metadata's accuracy is to explore more intelligent syntactic analysis of the caption text when more than one modality term is present. For example, a caption that reads, "Unlike the CT, this ultrasound shows...", would currently be misclassified as a CT. The ability to identify negative or "comparison" clauses could reduce such ambiguity.

We also are exploring machine-learning techniques to identify the imaging modality directly from the images themselves. Although imperfect, our classification of images by modality is sufficiently accurate to allow the collection to serve as a training set for image analysis techniques. Our approach is in line with current efforts at content-based image retrieval.[4]

There was substantial variation in the quality of information available for filtering in the set of figure captions. Some sources presented information more completely and uniformly in their figure captions. Journal editors should consider approaches to aid in indexing the wealth of imaging content they publish through standardized formats for figure captions and the use of metadata. Image annotation[5] is an important component of efforts to derive the greatest possible information from biomedical text data.[6] Metadata annotation of biomedical images is part of the Annotation and Image Markup project in the National Cancer Institute's Cancer Bioinformatics Grid (caBIG) effort.[7] Several groups have been interested in improving metadata of medical images for use in clinical and educational applications.[8,9]

Our results suggest that relatively simple approaches can be applied successfully to discover metadata in a large image library and that the metadata can be used to filter concept-based searches by age, sex, and imaging modality. Readers can explore the search engine and image library online: ARRS GoldMiner™ is freely available at http://goldminer.arrs.org.

ACKNOWLEDGMENTS

This work was supported in part by the American Roentgen Ray Society.

REFERENCES

1. Kahn CE Jr, Thao C: GoldMiner: a radiology image search engine. AJR Am J Roentgenol 188:1475–1478, 2007

2. Hersh W: Evaluation of biomedical text-mining systems: lessons learned from information retrieval. Brief Bioinform 6:344–356, 2005

3. Singhal A: Modern information retrieval: a brief overview. IEEE Data Eng Bull 24:35–43, 2001

4. Antani S, Long LR, Thoma GR: Content-based image retrieval for large biomedical image archives. Medinfo 11:829–833, 2004

5. Goede PA, Lauman JR, Cochella C, Katzman GL, Morton DA, Albertine KH: A methodology and implementation for annotating digital images for context-appropriate use in an academic health care environment. J Am Med Inform Assoc 11:29–41, 2004

6. Cohen AM, Hersh WR: A survey of current work in biomedical text mining. Brief Bioinform 6:57–71, 2005

7. Saltz J, Oster S, Hastings S, et al: caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. Bioinformatics 22:1910–1916, 2006

8. Muller H, Rosset A, Vallee JP, Terrier F, Geissbuhler A: A reference data set for the evaluation of medical image retrieval systems. Comput Med Imaging Graph 28:295–305, 2004

9. Hersh WR, Muller H, Jensen JR, Yang J, Gorman PN, Ruch P: Advancing biomedical image retrieval: development and analysis of a test collection. J Am Med Inform Assoc 13:488–496, 2006