Journal of Digital Imaging

An Inductive Method for Automatic Generation of Referring Physician Prefetch Rules for PACS

Yasuhiko Okura, MS, Yasushi Matsumura, MD, PhD, Hajime Harauchi, BS, Yoshiharu Sukenobu, Hiroko Kou, MS, Syunsuke Kohyama, BS, Norihiro Yasuda, BS, Yuichiro Yamamoto, and Kiyonari Inamura, PhD

To prefetch images in a hospital-wide picture archiving and communication system (PACS), a rule must be devised to permit accurate selection of examinations in which a patient's images are stored. We developed an inductive method to compose prefetch rules from practical data which were obtained in a hospital using a decision tree algorithm. Our methods were evaluated on data acquired in Osaka University Hospital for one month. The data collected consisted of 58,617 cases of consultation reservations, 643,797 examination histories of patients, and 323,993 records of image requests in PACS. Four parameters indicating whether the images of the patient were requested or not for each consultation reservation were derived from the database. As a result, the successful selection sensitivity for consultations in which images were requested was approximately 0.8, and the specificity for excluding consultations accurately where images were not requested was approximately 0.7.

KEY WORDS: PACS radiology, algorithm, Database Management Systems, Information Management

THE PURPOSE FOR INTRODUCING high-speed networks or high performance servers in hospital-wide a picture archiving and communication system (PACS) is increased efficiency. Nevertheless a major problem is the relatively long response time at image display terminals. Response time is affected both by the large number of images generated by digital modalities and the large number of image display terminals installed in the hospital. The load on network or disk servers tends to be large. To shorten response time at terminals and reduce the server load, image prefetching that transfers image data from an image server to an intermediate server installed near an image terminal is often used. Requesting image data stored in the intermediate server provides relatively short response time and low degree of image server load. In particular, this approach is effective when images stored in long-term storage such as jukeboxes are requested.

However, prefetching has limitations. The data size of all images requested in a day is too large, especially for hospital-wide PACS, to be prefetched generally because limitations in network bandwidth or limited capacity of disk drives in the intermediate server. In addition, a physician may not have requested a patient's images, or the physician may request not only the most recent images, but also previous older images for a comparative diagnosis over time. Therefore, accurate selection of images requested at consultations each day is required for efficient prefetching.

In most commercially available PACS, rules for prefetching are generally available and are set manually. Those rules depend on the experience or capability of each operator or soft-

From the Graduate School of Medicine, Course of Health Sciences, and the School of Allied Health Sciences, Faculty of Medicine, Osaka University, Osaka, Japan; and the Departments of Medical Information Sciences and Radiology, Osaka University Hospital, Osaka, Japan; and the Medical Systems Development Division, NEC Corporation, Tokyo, Japan.

Correspondence to: Yasuhiko Okura, MS, Inamura Laboratory, School of Allied Health Sciences, Faculty of Medicine, Osaka University, 1-7 Yamadaoka, Suita, Osaka, 565-0871, Japan; tel: +81-6-6879-2570; fax: +81-6-6879-2574; e-mail: ookura@abox9.so-net.ne.jp.

Online publication 21 January 2003
doi: 10.1007/s10278-002-0033-4

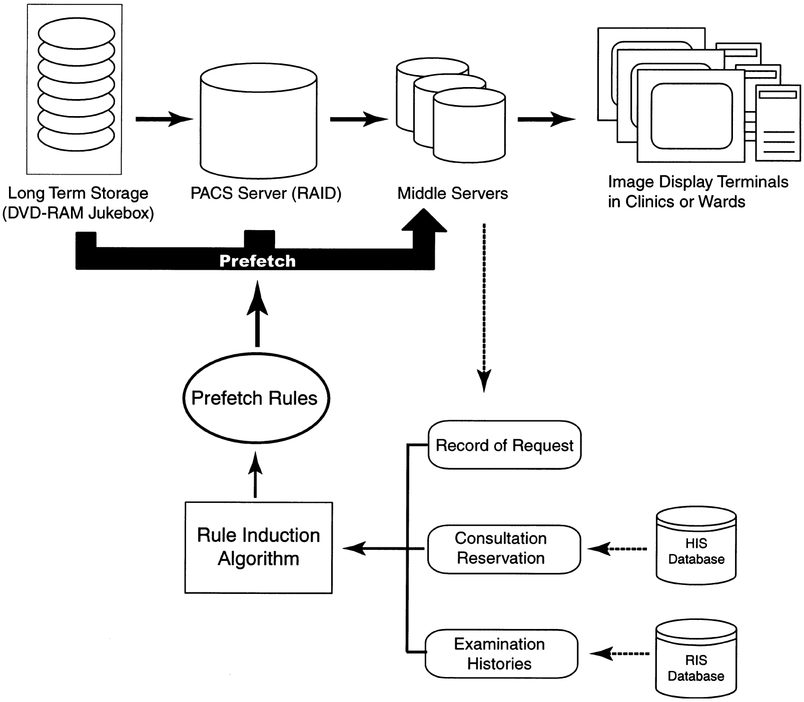

Fig 1. Schema of our method to derive prefetch rules from data obtained in the hospital.

ware specialist, and the accuracy of rules composed varies from hospital to hospital.

Recently, various methods that can be used to extract rules inductively from compiled data have been developed in the research field of machine learning and data mining outside the medical field.[10,17]

The purpose of this study was to develop an algorithm to build efficient image prefetch rules using the inductive method for a hospital-wide PACS. We present a method of automatic prefetch rule composition based on a historical survey of image requests and image examinations. Furthermore, this study examines our method using compiled data obtained from our PACS.

MATERIALS AND METHODS

Data Set

Three types of data from consultations at Osaka University Hospital from May 1, 2001, to May 31, 2001, were collected. The first data type was a log file recorded in the PACS image servers. This file recorded image requests from every image display terminal in the hospital. The second data type used radiological examination histories for patients from Aug 2000 to May 2001. Third, we acquired consultation reservation records, which were the records of scheduled clinic visit of patients, for the month of May 2001 from data files in the hospital information system (HIS). These included patient ID, consultation date, reserved department name, and purpose of the consultation. All logs or records were entered into a database using PostgreSQL version 7.1.3 on RedHat Linux version 7.1. The total numbers of PACS log records, examination histories, and consultation reservations entered into the database were 323,993, 643,797, and 58,617, respectively.

Selection of Consultations for Prefetching

The scheme of our inductive method to compose rules from actual records is shown in Figure 1. For each patient, we selected four parameters that were expected to influence whether the patient's study would be requested or not in the consultation: (1) the name of latest examination modality, such as computed tomography ("CT"); (2) the number of days since the latest examination; (3) the number of radiological examinations the patient received; and (4) the name of the department to which the patient was asked to report for the next consultation. If a patient had visited the surgery department and was told to go to the urology department the next day, this parameter was shown as "urology department." We extracted these four parameters on each consultation from acquired data to work out prefetch rules employing a program developed in the Perl language. The extracted parameters were stored in the database as a table named "mid_tbl_1."

Another table, "mid_tbl_2," containing data regarding whether a patient's study was requested or not in the consultation, was created from the acquired image request log by another Perl program.

We combined these two tables and constructed a table named "mid_tbl," consisting of four parameters and "the answer."

In this study, we used a decision tree algorithm to derive the prefetch rules and to compose SQL strings for selecting consultations. The decision tree algorithm can explicitly show the rules and the reasons the rules are selected. The processing speed achieved to construct decision trees and induce the rules was relatively fast. Because most of our PACS servers can utilize composed SQL strings for image prefetch, the speed suggested that the decision tree algorithm would be suitable for prefetching.

To compose the rules inductively from the "mid_tbl" table, we employed five decision tree induction algorithms, CHAID[8] E-CHAID[2] CART[3] QUEST[12] and C5.0 (RuleQuest Research Pty Ltd, St Ives, Australia). The first four algorithms were available in Answer Tree version 3.0 (SPSS Japan Inc., Tokyo, Japan). The C5.0 algorithm was also available in commercially available software. When C5.0 was used, we added an "-r" option to derive a rule in our study. The decision trees were induced from the data set of a part of the "mid_tbl" as learning data, and the rules for selecting consultation for prefetching were derived from induced decision trees by selecting the "leaves" defined as the consultation in which the images were requested. The rules were expressed as the form of a "WHERE" clause available in SQL strings.

To evaluate the rules, we used two indices. One was "sensitivity," which indicates the ratio of the number of consultations that actually requested a study ("true positive" cases) to the number of consultations selected by induced rules. This index reflects the degree of improvement in response time at a terminal. Another was "specificity," which is defined as the ratio of the number of accurately excluded consultations for prefetching ("true negative" cases) to the number of consultations where a patient's study was not requested. This value will show the degree of reduction of disk and/or network load of image servers during prefetching of study data from image server to intermediate server.

Error Significance Ratio

For image prefetching, two types of errors occurred: (1) a patient's study had been requested in a consultation but had not been selected for prefetching; (2) prefetched studies in a consultation had not been requested. The significance of these two types of error is different. Because the objective of introducing prefetching to PACS is to improve response times at terminals, the first type of error is more significant. We defined "error significance ratio" as the ratio of error significance of the second type of error to that of the first type. "Error significance ratio" was applied in every decision tree induction algorithm to differentiate misclassification costs for inducing the tree.

To find the optimized significance ratio between these two types of error, we examined the effect of the "error significance ratio" on the accuracy of induced rules. To assess the accuracy of the induced prefetch rule, ten folds cross-validation was used. This cross-validation divides the data into 10 sets of approximately equal size. In each experiment, a single set is used to evaluate the accuracy of the rule derived from the nine remaining data sets. Sensitivity and specificity were calculated using averaged numbers of consultations obtained from 10 experiments.

Data Size for Training

In addition, the minimum number of training data to induce accurate rules was studied. The data size for deriving a rule were changed, and the 10 independent data sets were chosen as training data from the collected data at random for each data size. For each data size, accuracy indices were also obtained from the averaged number of selected consultations.

Number of Studies Used to "Prefetch"

To carry out prefetch in PACS, one or more studies for each patient in every derived consultation must be selected. Because the number of selected studies affects the prefetch success rate, a further analysis was carried out to estimate the success rate. The hit rate, defined as the ratio of the number of prefetched studies to the number of requested studies, was measured using our data set.

RESULTS

The number of consultations for which patients had radiographic examination histories at the time of consultation in May 2001 was 43,269. In contrast, the number of consultation cases in which radiographic studies were requested from the image display terminals in the consultation rooms was 1,260.

Figure 2A and B shows both sensitivity and specificity as a function of the error significance ratio. Both indices were clearly influenced by the error significance ratio. A desirable error significance ratio might also depend on the practice of each PACS. In this study, we selected a ratio of 0.05, at which both indices showed a high value that was comparable in all four algorithms. In this case, sensitivity and specificity were approximately 0.7 and 0.8, respectively. When using algorithm C5.0, both indices changed with an extremely steep slope; we could not obtain an optimized error significance ratio.

As shown in Figure 3A and B, both indices were stable when the ratio of the numbers of all learning data sets accounted for more than 10% of our sample data evenly among the five

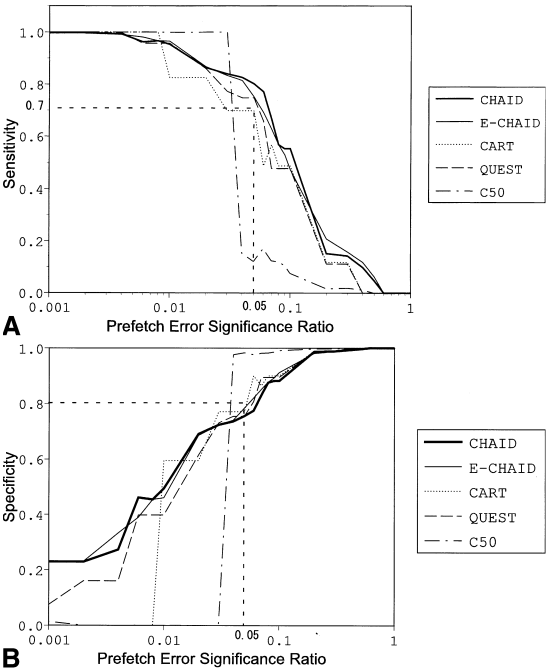

Fig 2. Two indices that show the accuracy of prefetch rules calculated depended strongly on the value of the error significance ratio. A: Sensitivity shows the ratio between the number of selectedconsultations and number of consultations where the patient's images were requested. B: Specificity shows the ratio of the number of not-selected consultations to the number of consultations in which the patient's images were not requested.

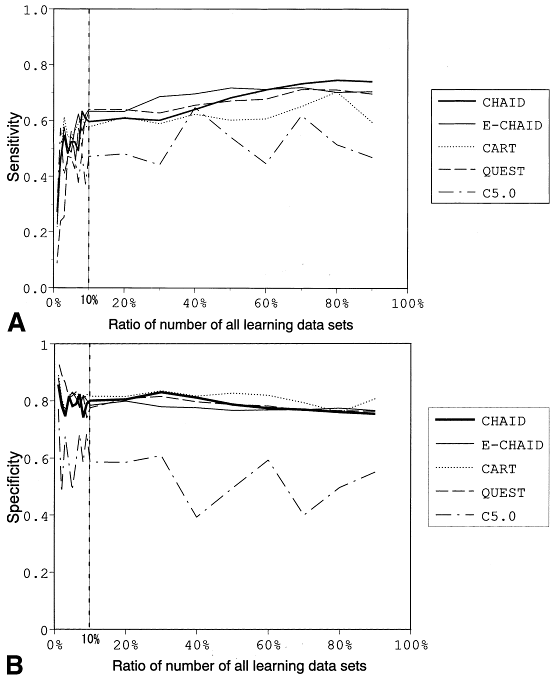

Fig 3. The prefetch rule accuracy indicated by (A) sensitivity and (B) specificity was sufficiently high when the ratio of the number of all learning data sets was 10% or larger.

methods. The number of days needed to collect the corresponding number of data was 5 days at our hospital.

The hit rate increased when the number of prefetched previous studies increased, as shown in Figure 4. When more than five previous studies were prefetched, the hit rate was over 0.9.

DISCUSSION

Although image prefetching itself is not a new idea,[9,11,13] the effectiveness of prefetching has recently been claimed to be augmented. One reason for this is the increase in the number of images for one study, as described below.

Multitiered archives that consisted of both a fast access device with small capacity and a slow access device with a large capacity were introduced virtually in many PACS to satisfy the dual demands for a short response time for displaying current images and a large capacity for long-term storage.[6,14] The increased amount of image data that is generated from multi-detector CT tends to use a large amount of fast access storage such as RAID. In the PACS layered storage system, an increase in the number of images indicates a decrease in the probability of accessing historical studies stored in fast access storage. Accordingly, the probability of fetching previous studies recorded on slow access storage is increasing. Namely, the response time at the terminal is being extended.

Image prefetching on PACS has been discussed in many earlier articles. Wilson et al. studied patient examination history statistically and then derived prefetch rules from the results.[16] Hu et al. noted a knowledge-based method for prefetching, and Bui et al. also described prefetching using multiple data sources.[4,7] Siegel et al. suggested that image prefetching should be carried out according to the last n studies in chronological order.[14]

In other articles, PACS or prefetch was mainly installed for radiologists. Where radiologists are concerned, images relevant to the current examination should be prefetched. To

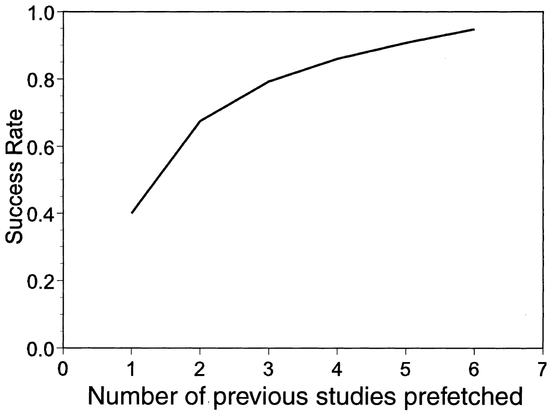

Fig 4. Number of studies prefetched for selected consultation versus the success rate of prefetching. To achieve an approximately 90% of success rate, five studies should be prefetched for selected consultation by prefetch rules.

select relevant images for prefetching, several methods have been studied. Donnelly used a "comparative region of interest" to find relevant images.[5] Andriole et al. employed metagroup categories to examine type mnemonics.[1]

It is important to note that, for hospital-wide PACS, current or old radiographic studies are not requested at every consultation, particularly in outpatient clinics. Only a few necessary images are requested at certain, limited consultations. Accordingly, it is important to avoid unnecessary prefetching to achieve maximal efficiency. Our strategy for prefetching involved two stages, the selection of consultation and the selection of studies for the selected consultation. As shown in Results, our method excludes most unnecessary prefetching. Because our method is limited to the selection of consultation, a further algorithm is required for prefetching. As discussed in other articles, the last *n* times selection of the stored study showed good performance.[14] As shown in Figure 4, our results were similar, and it is reasonable to apply our algorithm in the selection of studies to prefetch.

The inductive method has the advantage of generality when applying it to other types of PACS. Generally, it is difficult to compose efficient rules for prefetching, because of the variations in the study request patterns from various departments within a hospital. The method we developed solved this problem.

Nevertheless, several significant problems were found with our method. Although the number of parameters used to make rules was relatively small, we obtained sufficiently high values of sensitivity and specificity. However, the parameters themselves or the number of parameters needed to make efficient rules depends on the character of PACS in each hospital, and the selection of parameters will influence the accuracy of prefetch rules.

In addition, a method to optimize the error significance ratio should be considered. We carried out several experiments to find the reasonable error significance ratio. However, it is obvious that this method is too complex to apply to every PACS. As noted in the Results, the optimized error significance ratio may depend on the character of the PACS. A low ratio value should be used if a higher prefetched image probability is needed.

In spite of using different methods to induce a decision tree, the algorithms applied showed similar results, with the exception of C5.0. A detailed study of this phenomenon shown by C5.0 was beyond the scope of this paper. It is noteworthy that the shape of the graph for the CART results was stepwise while the graphs of CHAID, E-CHAID, and QUEST were slopes. Therefore, CHAID, E-CHAID, and QUEST are more suitable if the optimized error significance ratio is ambiguous.

In addition, a so-called incremental algorithm that induces a new decision tree from an existing decision tree and newly obtained training data described by Utgoff will be useful.[15] This "incremental" algorithm will enable practical application of use our methods.

REFERENCES

1. Andriole KP, Avrin DE, Yin L, et al: Relevant priors prefetching algorithm performance for a picture archiving and communication system. J Digit Imaging 13(Suppl 1):73-75, 2000

2. Biggs D, de Ville D, Suen E: A method of choosing multiway partitions for classification and decision trees. J Appl Statist 18:49-62, 1991

3. Breimen L, Friedman JH, Olshen RA, et al: Classification and Regression Trees. Wadsworth, Belmont CA, 1984

4. Bui AA, McNitt-Gray MF, Goldin JG: Problem-oriented prefetching for an integrated clinical imaging workstation. J Am Med Inform Assoc 8:242-253, 2001

5. Donnelly JT, Anderson Q: Automating clinically relevant prefetch processes in picture archiving and com-

munication systems solutions. J Digit Imaging 14(Suppl 1) 135-139, 2001

6. Dwyer SJ, Templeton AW, Martin NL, et al: The cost of managing digital diagnostic images. Radiology 144:313-318, 1982

7. Hu PJ, Wei CP, Liu Sheng OR: A knowledge-based patient image prefetching system: design, evaluation and management. Top Health Inform Manage 20:42-57, 1999

8. Kass G: An exploratory technique for investigating large quantities of categorical data. Appl Stat 29:119-127, 1980

9. Kishore S, Khalsa S, Stevens JF, et al: On-demand retrieval paradigm. Proc SPIE 2165:149-157, 1994

10. Kononenko I, Bratko I, Kukar M: Application of machine learning to medical diagnosis. In: Michalski RS. Bratko I. Kubat M eds. Machine Learning and Data Mining. Wiley, New York, 1997, pp 389-408

11. Lodder H, Poppel BM, Bakker AR: Prefetching: PACS image management optimization using HIS/RIS information. Proc SPIE 1446:227-233, 1991

12. Loh WY, Shih YS: Split selection methods for classification trees. Statistica Sinica 7:815-840, 1997

13. Osada M, Nishihara E: Implementation and evaluation of workflow based on hospital information system/radiology information system/picture archiving and communication system. J Digit Imaging 12(2 Suppl 1):103-105, 1999

14. Siegel EL, Reiner BI: Recommendations for image prefetch or film digitization strategy based on an analysis of an historic radiology image database. J Digit Imaging 11:94-99, 1998

15. Utgoff PE: Incremental induction of decision trees. Machine Learning 4:161-186, 1989

16. Wilson DL, Smith D, Rice B, et al: Intelligent prefetch strategies for historical images in a large PACS. Proc SPIE 2165:112-123, 1994

17. Zupan B, Demsar J, Smrke D, et al: Predicting patient's long-term clinical status after hip arthroplasty using hierarchical decision modelling and data mining. methods. Inf Med 40:25-31, 2001