Knowledge Extraction From Radiology Report Using Text Mining

Ms.Abhilasha T. Borkar, Prof. Archana A. Nikose
Bhagwati Chaturvedi college of Engg. Nagpur

Abstract

Text Mining is the discovery by computer of new, earlier anonymous information by automatically extracting information from different written resources. A key aspect is the linking together of the extracted information together to form new facts or new hypotheses to be explored. The aim of this project to extract the information and used that information in radiology report. The system is consisting of three main modules: the medical finding extractor, report retriever, and image retriever. In this paper we are going to explain proposed method for implementing the system. first medical finding extracting module we have apply the natural language processing algorithm which automatically extract the medical finding and their modifiers which is used for structuring the radiology report. The structure report will act as a intermediate result for final result. This structuring of the free text report generally avoids the gap between user and reports, and make the information contained in the report is easily accessible. The next module is the image feature extraction which is used to extracting the feature for exact match. and the last module which is query analyzer module that is report retrieval module will take user query as a input and will generate the exact match like target image with associated report for the user enter query. the overall evaluation test is good.

1. Introduction

Text mining, sometimes alternatively referred to as text data mining, nearly equivalent to text analytics, refers generally to the process of derives high-quality information from text. High quality information is normally derived through the deriving of patterns and trends means such as statistical pattern learning. Text mining generally involves the process of structuring the input text such as parsing, along with the addition of some derived linguistic features and the removal of others, and consequent insertion into a database and deriving patterns within the structured data, and finally evaluation an interpretation of the output. Elevated quality in text mining usually refers to some combination of relevance, innovation, and interested information. Typical text mining tasks include text classification, text clustering concept or thing extraction, sentiment analysis, document compression, and entity relation modelling i.e. learning relations between named entities. With the advances in medical technology and wider adoption of electronic medical record systems, large amounts of medical text data are produced in hospitals and other health institutions daily. These Medicinal texts include the patient's medicinal history, medical encounters, instructions, improvement notes, test results, etc. The quantity of medical literature continues to produce and deliberate. However interest in the field of biomedical research is ready with route of time because of frequently changing in human genome resulting in the information overfilling in form of online publication.

In radiology reports are in free text format and usually unrefined, there is a great barrier between the radiology reports and the medical professionals like radiologists, physicians, and researchers, making it difficult for them to retrieve and use useful information and knowledge from the reports. Generally the information is not accessible; it cannot be used for other related applications. Therefore, to provide the essential information to the medical professionals and make use of the information, text mining in the radiology reports provides a solution to the problem and we are presenting a methodology like for extracting useful information from medical journals papers as well as Radiology report. In this Project we are applying some techniques of data mining to extract useful pattern from huge amount of clinical data.

2. Architecture Of proposed system

Our proposed system consist of three main modules: text extractor, image extractor and text assisted report extractor as shown in fig2.1 the first module is the text extractor module which is focused on free text radiology report. By using the steps of natural language processing we have structure that report for further processing.

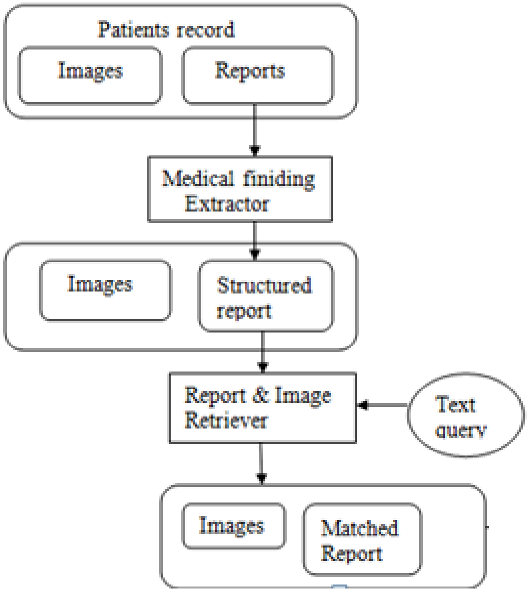


Fig2.1 System Architecture of proposed system

The second module which is generally focused on to the radiology images by using the various techniques of image feature extraction i.e. CBIR is used to retrieve the images from the various set of images. the third module is report extractor in this user query is entered like some text or phrase the proposed system giving the exact match for the user entered query means it gives the image associated report. the input to the proposed system is either text or image also.

2.1 Natural Language Processing

A natural language processor is developed that automatically structures the important medical information e.g., the reality, properties, location, and diagnostic analysis of findings contained in a radiology free-text document as a formal information model that can be interpreted by a computer program. In this project we have given input to the system is a free text report from a radiologic study. The system requires no reporting approach changes on the part of the radiologist. Arithmetical and machine learning methods are used comprehensively throughout the system. In System graphical user interface has been developed that allows the selecting the some free text radiology report. Various aspects of the difficult problem of implementing an automated structured reporting system have been addressed; Extensible Mark-up Language is rising as the chosen syntactic pattern for representing and distributing these structured reports within a clinical environment. Early successes cleave to out hope that related statistically based models of language will allow deep understanding of documented reports. The success of these statistical methods will depend on the accessibility of large numbers of high-quality training examples for each radiologic subarea. The suitability of condition. However, a huge percentage of this information is amorphous, taking the form of free text, and is therefore difficult to search, sort, analyze, summarize, and automated structured reporting systems will eventually depend on the results of widespread evaluations. In our project we have used the various steps for natural language processing as shown in fig.1.1.0 the Architecture of NLP.

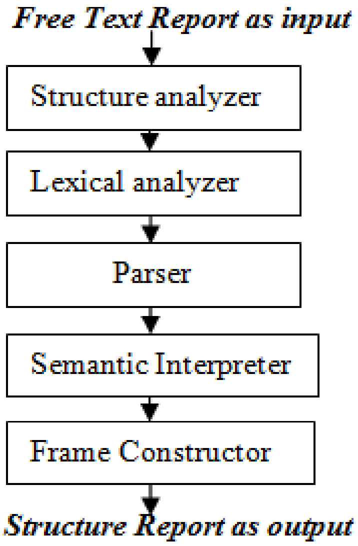


Fig 1.1.0 Architecture of NLP

3. Description about Radiology report

Radiology reports include a great deal of information that characterizes a patient's medical present. Previous studies have demonstrated the potential benefits of structured medical data for medical training, research, and teaching. In the medical setting, structured reports can be used to help categorize and improve the presentation of the medical record. For example, if a radiologist is interested only in a given clinical occurrence, he or she may select to retrieve only those reports in which the relevant anatomy or findings are described. Accurate extraction of abrasion size information from radiology reports can allow a system to automatically construct a growth timeline for an

1. Automatic structuring requires deep understanding because it is desirable to translate all relevant Information into structured form.

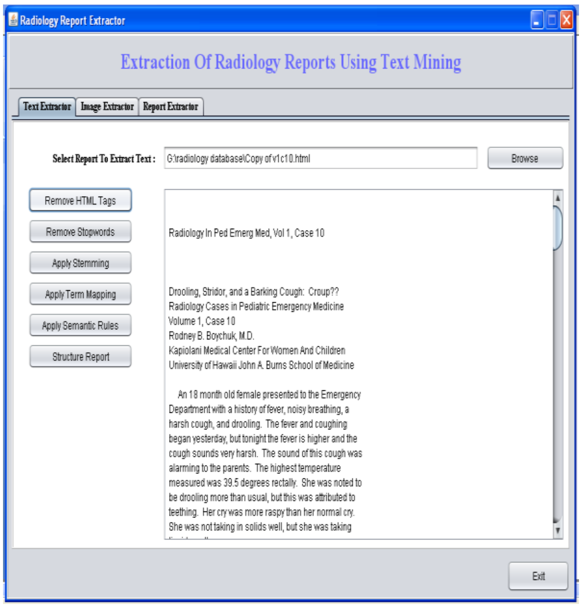
2. Automatic structuring must deal with ungrammatical writing approach. Shorthand and telegraphic Writing approach are common in radiology reports. A natural

3.The glossary is large. Large numbers of complex medical terms, suitable names, product names, abbreviations, and staging codes are used in radiology reports. Hundreds of explanatory adjectives are used that are not found in any common electronic medical glossaries.

4. There is an alleged knowledge between the writer and reader. The radiologist knows the World of the referring physician and vice versa. Subsequently, details are often left out because they are assumed to be common knowledge.

4. Working of medical finding extractor module

As shown in fig 1.1.0 the Architecture of NLP. The task of automatically structuring radiology reports can be divided into various subtasks. The first is removing HTML tags from models of the targeted information contained within the report. The second subtask is removing stop word from the medical report. Stop words are the commonly occurring words. Several medical natural language processing systems are currently being researched and are in various stages of clinical testing. The third subtask is stemming. Stemming means reducing a word to its base (or stem). For example, the words 'writing', 'wrote' and 'written' all have the stem 'write'. A stemmer acquires a word, or a list of words, and constructs the stem, or a list of the stems, of the input. Stemming is useful when you are doing any kind of text-analysis: when you are concerned about the inside of a text, the different times of verbs, and the different conclusion for singular Dictated report. Thus, ordered reports serve as a key for building medical multimedia digital libraries. The fourth subtask is semantic rule. In which the parser parses each sentence and outputs the typed dependency tree, which shows the syntactic relations between finding extractor selects the findings and thei rmodifiers according to a set of semantic rules. As shown in fig 3.1 the free text report as a input to our proposed system. We have developed a system that automatically structures the important medical information document as a formal information model that can be interpreted by a computer program. The input to the system is a free text report from a radiologic study done by expertise. the system enforces no constraints on the transcription style of the radiologist .the system outputs a formal representation of the important information contained in the free-text includes information on the existence, properties, document in the form of a report. The desired output frame for This sample sentence is shown in fig2.3 note that report consist of an abdominal Analysis and a set of examination on them. in this paper we focus mainly on the last subtask involved automatically structuring radiology reports like natural language processing we examine our proposed system in terms of overall architecture structure analyzer ,lexical analyzer, parser, semantic or syntactic interpreter, and report constructor.



Extraction Of Radiology Reports Using Text Mining

Select Report To Extract Text : G:\radiology database\Copy of v1c10.html

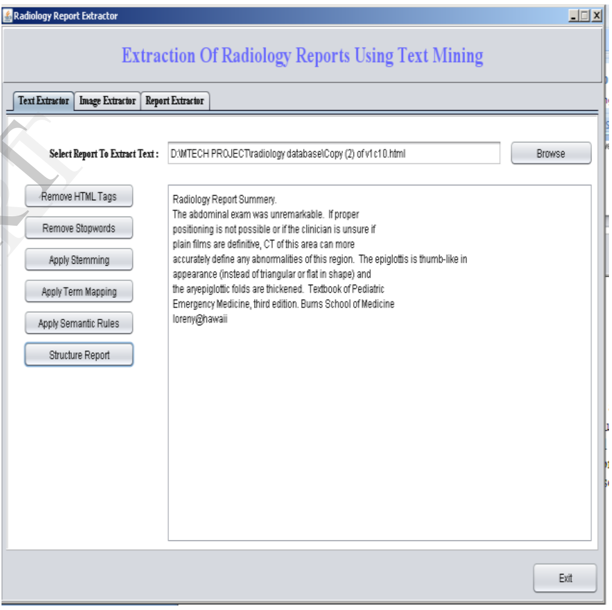
Radiology In Ped Emerg Med, Vol 1, Case 10

Drooling, Stridor, and a Barking Cough: Croup??
Radiology Cases in Pediatric Emergency Medicine
Volume 1, Case 10
Rodney B. Boychuk, M.D.
Kapiolani Medical Center For Women And Children
University of Hawaii John A. Burns School of Medicine

An 18 month old female presented to the Emergency Department with a history of fever, noisy breathing, a harsh cough, and drooling. The fever and coughing began yesterday, but tonight the fever is higher and the cough sounds very harsh. The sound of this cough was alarming to the parents. The highest temperature measured was 39.5 degrees rectally. She was noted to be drooling more than usual, but this was attributed to teething. Her cry was more raspy than her normal cry. She was not taking in solids well, but she was taking

fig3.1 free text Report

.



Extraction Of Radiology Reports Using Text Mining

Select Report To Extract Text : D:\MTECH PROJECT\radiology database\Copy (2) of v1c10.html

Radiology Report Summery.
The abdominal exam was unremarkable. If proper positioning is not possible or if the clinician is unsure if plain films are definitive, CT of this area can more accurately define any abnormalities of this region. The epiglottis is thumb-like in appearance (instead of triangular or flat in shape) and the aryepiglottic folds are thickened. Textbook of Pediatric Emergency Medicine, third edition. Burns School of Medicine
loreny@hawaii

Fig3.2 structure report

5. Image feature extraction module:

The image feature extraction module involves extracting the features like shape, location, and color texture of the images so that it will help to give the similar match for the result. The image feature extraction module then uses other image processing techniques to select other features. The extracted features form a structured image, which can be used for further image mining tasks, such as medical image classification. As shown n fig2.1 the architecture of image feature extraction module in which the query image given by the user the feature extraction will extract the features of query image and that feature will match to the target images feature in the database so that it will help to find the exact match For implementing this module we have used content base image retrieval method (CBIR), image classification etc. we use the location information from the type and location modifiers in structured report to select the area of interest in the image. When the patient of abnormality regions are produced from image mining procedures, we use the shape, size and intensity information from exact match of structured report to resize the area of interest, draw contours, and segments the abnormality region in the brain. The following Steps of image feature extraction are as follows.

5.1 Colour Extraction

Many image display devices allow only a limited number of colors to be simultaneously displayed. Usually, this set of available colors, called a color palette, may be selected by a user from a wide variety of available colors. Such device restrictions make it particularly difficult to display natural color images since these images usually contain a wide range of colors which must then be quantized by a palette with limited size. This color quantization problem is considered in two parts: the selection of an optimal color palette and the optimal mapping of each pixel.

5.2 Canny Edge detector:

The purpose of edge detection in general is to significantly reduce the amount of data in an image, while preserving the structural properties to be used for further image processing. Several algorithms exists, we have focuses on a particular one developed by John F. Canny (JFC) in 1986 Even though it is quite old, it has become one of the standard edge detection methods and it is still used in research .

The aim of JFC was to develop an algorithm that is optimal with regards to the following criteria:

1. Detection: The probability of detecting real edge points should be maximized while the probability of falsely detecting non-edge points should be minimized. This corresponds to maximizing the signal-to-noise ratio.
2. Localization: The detected edges should be as close as possible to the real edges.
3. Number of responses: One real edge should not result in more than one detected edge.

The Canny Edge Detection Algorithm

The algorithm runs in 5 separate steps:

1. Smoothing: Blurring of the image to remove noise.
2. Finding gradients: The edges should be marked where the gradients of the image has large magnitudes.
3. Non-maximum suppression: Only local maxima should be marked as edges.
4. Double thresholding: Potential edges are determined by thresholding.
5. Edge tracking by hysteresis: Final edges are determined by suppressing all edges that are not connected to a very certain (strong) edge.

5.2.1 Smoothing

It is inevitable that all images taken from a camera will contain some amount of noise. To prevent that noise is mistaken for edges, noise must be reduced. Therefore the image is first smoothed by applying a Gaussian filter. The kernel of a Gaussian filter with a standard deviation of σ = 1.4 is shown in Equation (1). The effect of smoothing the test image with this filter shown

$$B = \frac{1}{159} \cdot \begin{bmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 4 \\ 2 & 4 & 5 & 4 & 2 \end{bmatrix}$$

Eq.1

5.2.2 Finding gradients

The Canny algorithm basically finds edges where the grayscale intensity of the image changes the most. These areas are found by determining gradients of the image. Gradients at each pixel in the smoothed image are determined by applying what is known as the Sobel-operator. First step is to approximate the gradient in the x- and y-direction respectively by applying the kernels shown in Equation (2).

$$K_{GX} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

$$K_{GY} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Eq.2

The gradient magnitudes (also known as the edge strengths) can then be determined as an Euclidean distance measure by applying the law of Pythagoras as shown in Equation (3). It is sometimes simplified by applying Manhattan distance measure as shown in Equation (4) to reduce the computational complexity. The Euclidean distance measure has been applied to the test image. The computed edge strengths are compared to the smoothed image in Figure 3

4.3 Texture Extraction

Texture feature extraction is the procedure of generating descriptions of a textured surface in terms of measurable parameters. The extracted features represent the relevant properties of the surface, and may be used with a classifier. It is commonly agreed that textural features play a fundamental role in classifying textured surface and texture segmentation.

4.3.1 Use of Texture Analysis

1. Segment an image into regions with the same texture, i.e. as a complement to gray level or color.
2. Recognize or classify objects based on their texture.
3. Find edges in an image, i.e. where the texture changes.
4. Shape from texture.
5. Object detection, compression, synthesis.

4.4 Feature Extraction:

Feature extraction is the quantification of texture characteristics in terms of a collection of descriptors or quantitative feature measurements, often referred to as a feature vector. Texture features and texture analysis methods can be loosely divided into two categories –

4.4.1 Statistical approach:

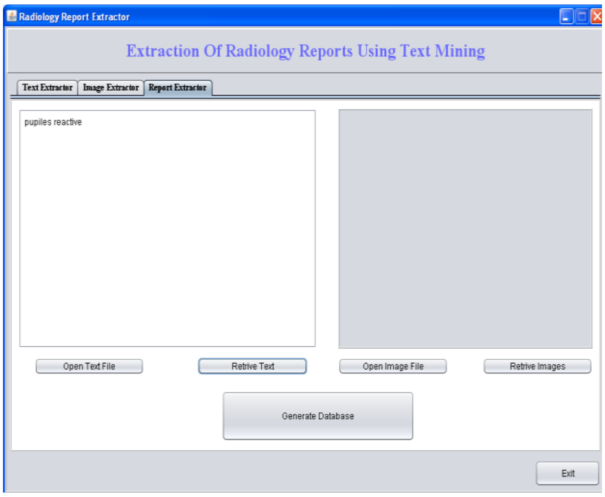
Statistical method is more easily handled here. This method analyze the spatial distribution of grey values, by computing local features at each point in the image, and deriving a set of statistics from the distributions of the local features. With this method, the textures are described by statistical measures. One commonly applied and referenced method is the co-occurrence method, introduced by Haralick [Haralick73]. In this method, the relative frequencies of grey level pairs of pixels separated by a distance *d* in the direction è combined to form a relative displacement vector *(d,* è*)*, which is computed and stored in a matrix, referred to as grey level co occurrence matrix (GLCM) P. This matrix is used to extract second-order statistical texture features. Haralick suggests several features describing the two

$$P_{ij} = \frac{V_{ij}}{\sum_{i,j=0}^{N-1} V_{ij}}$$

Dimensional probability density function *pij* which is nothing but the *normalization equation. eqn 1*
The features used are Entropy, Contrast, Dissimilarity, and Homogeneity. These are well handled with the previously formed normalization equation.One of the goals of radiology image mining is to detect any abnormality of the body part examined. For Stomach images, There is murmur or not and pupils reactive detection is one of the major tasks We use structured report scan of severe stomach flu in our project to help to extract features of any serious detection in the stomach . If there is such abnormality in the stomach, the detailed information about it is depicted in the structured report in terms of medical finding and its modifiers. In the example shown, “type”, “duration”, and “location” are modifiers for medical finding “epigastrium”.When they appear in the structured report as modifier values they entail the shape, the location and sometimes the size information of the epigastrium or daiheria as well. Other modifiers like “size” also entail image feature related information and are helpful for the image feature extraction module.

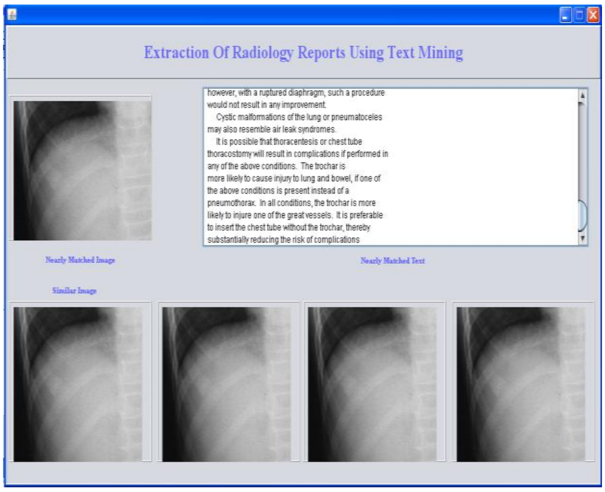
6. Report and Image Retrieval module:

The report and image retrieval module same as the medical finding extraction module instead of taking the full radiology report as input the query is entered by the user like word or any phrase like “there is murmur in abdominal or not” the report retrieval module analyze the query means it searches the image and structure report in the database and give the exact match which mostly needed by the user.



pupiles reactive

Fig.6.1 entering the query



however, with a ruptured diaphragm, such a procedure would not result in any improvement.
Cystic malformations of the lung or pneumatoceles may also resemble air leak syndromes.
It is possible that thoracentesis or chest tube thoracostomy will result in complications if performed in any of the above conditions. The trochar is more likely to cause injury to lung and bowel, if one of the above conditions is present instead of a pneumothorax. In all conditions, the trochar is more likely to injure one of the great vessels. It is preferable to insert the chest tube without the trochar, thereby substantially reducing the risk of complications

Fig.5.2 matching results

The above fig6.1 consists of a text box where a text query can be entered related to the observation required. This will give the two kind of match nearly match and similar match. Clicking on retrieve text the window that shows the radiology image and the associated text description of the image and also taking the image as a query it gives match result from the database.

6. Conclusion

In this project, we have proposed a text mining system to retrieve and use the information in radiology reports. The system consists of four main modules: medical terms extractor, report and image retriever, and text-associated image Feature extractor. The medical finding extraction module will automatically extract medical findings and associated modifiers to structure brain CT radiology reports. The construction of the free text Reports avoids the gap between users and report database, makes the information enclosed in the reports readily accessible. It also supply as intermediate result to other components of the system. The retrieval module analyze user's query and will returns the reports and images that match the query. The user query input will be in text as well as image also. In future point view we will try to generate the report from the image that is automatic machine translation.

7. References

[1]Automatic Structuring of Radiology Free-Text Reports Ricky K. Taira, PhD • Stephen G. Soderland, PhD • Rex M. Jakobovits, PhD RASNA 2001

[2] Caroline Lacoste, Joo-Hwee Lim, Jean-Pierre Chevallet, and Diem Thi Hoang Le have propose Medical-Image Retrieval Based on Knowledge- Assisted Text and Image Indexing. IEEE Transactions On Circuits And Systems For Video Technology, Vol. 17, No. 7, July 2007.

[3]Tianxia Gong1, Ruizhe Liu, Chew Lim Tan1, Neda Farzad, Cheng Kiang Lee, Boon Chuan Pang, Qi Tian, Suisheng Tang, and Zhuo Zhang have propose Classification of CT Brain Images of Head Trauma. Springer-Verlag Berlin Heidelberg 2007.

[4] A.Kannan , Dr.V.Mohan, Dr.N.Anbazhagan have proposed Image Retrieval Based on Clustering and Non Clustering Techniques using Image Mining. Int J Engg Techsci Vol 1(1) 2010,54-61.

[5] Ceyhun Burak Akgül, Daniel L. Rubin,2 Sandy Napel, Christopher F. Beaulieu, Hayit Greenspan, and Burak Acar have present the Content-Based Image Retrieval in Radiology: Current Status and Future Directions. Journal of Digital Imaging. Published online 08 April 2010.

[6] Tiamxia Gong, Chew Lim Tan, Tze Yun Leong, "Text Mining in Radiology reports", IEEE, 2008

[7] William Claster, Subana Shanmuganathan Text Mining of Medical Records for Radiodiagnostic Decision-Making journal of computers, vol. 3, no. 1, january 2008.

[8] D.T. Heinze, M.L. Morsch, and J. Holbrook, "Mining Free-Text Medical Reports.", Proceedings of the AMIA Annual Symposium, 2002, pp. 254-258

[9] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning Generating Typed Dependency Parses from Phrase Structure Parses .

[10] E. A. Mendonca, J. Haas, L. Shagina, E. Larson, and C. Friedman. Extracting information on pneumonia in infants using natural lanuage processing of radiology reports. Journal of Biomedical Informatics, 38:314-321, 2005.

[11] C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson. A general natural language text processor for clinical radiology. Journal of the American Medical Informatics Association, 1(2):161-174, March April 1994.

[12] Nai-Chung Yang, Wei-Han Chang, Chung-Ming Kuo *, Tsia-Hsing Li A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval Department of Information Engineering, I-Shou University, Tahsu, 840 Kaohsiung, Taiwan Received 2 October 2006; accepted 22 May 200 Available online 9 June 2007.
.