SHIFTING FROM LEGACY SYSTEMS TO A DATA MART AND COMPUTER ASSISTED INFORMATION RESOURCES NAVIGATION FRAMEWORK

Nikitas N. Karanikolas
Areteion University Hospital, University of Athens, 76 Vas. Sophias, Athens, Greece
Email: nnk@aretaieio.uoa.gr

Christos Skourlas
Dept. of Informatics, Technological – Educational Institute of Athens, Athens, Greece
Email: cskourlas@hol.gr

Keywords: Automatic Classification, Data Mart, Document Management, Free Text Retrieval

Abstract: Computer Assisted Information Resources Navigation (CAIRN) was specified, in the past, as a framework that allows the end-users to import and store full text and multimedia documents and then retrieve information using Natural Language or field based queries. Our CAIRN system is a general tool that has focused on medical information covering the needs of physicians. Today, concepts related to Data Mining and Data Marts have to be incorporated into such a framework. In this paper a CAIRN-DAMM (Computer Assisted Medical Information Resources Navigation & Diagnosis Aid Based On Data Marts & Data Mining) environment is proposed and discussed. This integrated environment offers: document management, multimedia documents retrieval, a Diagnosis–aid subsystem and a Data Mart subsystem that permits the integration of legacy system's data. The diagnosis is based on the International Classification of Diseases and Diagnoses, 9th revision (ICD-9). The document collection stored in the CAIRN-DAMM system consists of data imported from the Hospital Information System (HIS), laboratory tests extracted from the Laboratory Information System (LIS), patient discharge letters, ultrasound, CT and MRI images, statistical information, bibliography, etc. There are also methods permitting us to propose, evaluate and organize in a systematic way uncontrolled terms and to propose relationships between these terms and ICD-9 codes. Finally, our experience from the use of the tool for creating a Data Mart at the ARETEION University Hospital is presented. Experimental results and a number of interesting observations are also discussed.

1 INTRODUCTION

Computer Assisted Medical Information Resources Navigation – CAIRN was specified, in the past, as a framework that allows physicians to store full text and multimedia medical information and retrieve it using queries in Natural language. Today, concepts related to Document Management Systems, Data Warehousing, Data Mining, Data Marts, etc. offer new possibilities and have to be incorporated into such a framework forming the CAIRN-DAMM (Computer Assisted Medical Information Resources Navigation & Diagnosis Aid Based On Data Marts & Data Mining) environment.

1.1 The outline of the real problem

In this subsection, mention of a real problem and a short discussion will be given to clarify things and uncover the various aspects of the problem. ARETEION University Hospital (AR.UN.HO) of Athens is the Hospital of the University of Athens, in Greece. It was founded in 1898 and has three clinical departments: Surgical, Obstetrics and Gynecology and Nephrology.

In 1995, "HELIOS", a Hospital Information System (HIS) was purchased and installed at the ARETEION University Hospital (AR.UN.HO) of Athens. In the same time, "MediLab Lims", a Laboratory Information System (LIS) was also

installed and separately operated. New versions of the systems were ordered and operated, in 1999, to ensure the interconnection between them. The evaluation of the operation of the two Information Systems has revealed some difficulties in using them for nursing personnel, laboratory staff, etc. Apart from this, new end users' requirements have been specified which are not be covered by the two systems e.g. physicians have the duty to select the correct, for each patient, ICD-9 diagnosis, looking at thousands of possible ones. The existing systems do not support them, at all.

Patient discharge letters can be thought as narratives that describe in more detail the ICD-9 diagnosis that the physicians are looking for. Hence, methods have to be investigated to automatically match a patient discharge letter against the ICD-9 list and retrieve the most likely ICD-9 diagnosis.

There are also hospital units that are obliged to operate and maintain local sources of information apart from the central system database. Hence, data exported from the central HIS and other local resources must be "cleaned", stored and organized into new database(s) (Data Mart(s)) permitting user friendly information retrieval, extraction of summary information and specific information upload to local systems of the hospital units.

1.2 A technical discussion of the problem – Method of attack

The conclusions from the last five years of operating the Hospital Information System at the ARETEION Hospital could be summarized as follows:

- Divisions of a Hospital have often invested large amounts of money for distinct, operational, "legacy" systems covering their needs. These systems can not be replaced easily and their integration for supporting Decisions is a difficult task.
- Doctors have only partially accepted the use of the existing HIS. There are doctors, that exclusively use the system but also there are doctors that prefer to use general purpose Office Automation software for writing diagnosis notes, discharge letters, etc.
- There is a rigid trend for storing, retrieving and handling medical images from various sources e.g. MRI, CT, Ultrasound, etc.

Based on the above conclusions we decided to operate the existing HIS for all their users but we also proposed the use of the CAIRN system for the doctors that usually work with general purpose software (mainly word processor software) and for the store and retrieval of multimedia information.

The advantage of this approach is that the doctors continue to use their favourite software and only during the storing of their documents they face the simple interface of the CAIRN system. Hence, they work using their way and the medical files they produce are correlated and eventually form an electronic patient record. Moreover, we decided the introduction of a text data mining (Hearst, 1999) module in order to permit the automatic or semi-automatic selection of ICD-9 diagnosis codes, in case of a patient discharge letter.

The integration of the legacy system (HIS) and CAIRN is based on the attitude / estimation that Subject-oriented, integrated, non-volatile, time-variant data store (Data warehousing) can provide a single source of data for all decision support activities and play a significant role for solving the every day problems. The concept of Data Mart is also proposed as a volatile, "limited", "tailored to the purpose at hand", special purpose "warehouse". Our system is named CAIRN-DAMM and combines possibilities for Document Management, Free/Full Text Retrieval and Data Mining techniques and also uses Data Mart/Warehouse methodologies to integrate itself with the existing "legacy HIS".

1.3 Previous work

Lenz, Blaster and Kuhn (1999) present the advantages and tradeoffs of HIS integration. Lenz and Kuhn (2001) summarize the state of the art in web technology and compare it with the needs of Hospital information systems' integration.

Verma and Harper (2001) presented the results of a "classical" approach, which is being taken for designing and implementing a data mart.

Ricciardi et al (2001) discussed the benefits to individual physicians from low-cost clinical documentation and reporting tools in contrast to large-scale data warehousing projects.

According to Nada Lavrac (1999) there is a need to couple medical information systems with methods for efficient computer-assisted analysis.

In the next sections, the CAIRN-DAMM environment is presented in a step-by-step manner.

2 DATA MART – DATA WAREHOUSE CONCEPTS

In this section some terminology is given and an outline of the Data Mart, Data Warehouse principles adopted in the CAIRN-DAMM environment are briefly presented.

2.1 Data sources

Data sources that are related to the various phases of operating the target system must be specified. An information handling process has to focus on:

- The common, "corporate data" stored in various database files. These types of data are mainly extracted from the central HIS and other hospital units based local systems. Therefore, various formats have to be supported.
- Data of the new target database (CAIRN-DAMM Data Mart). This database has to be "built up" around specific topic(s) (subject(s)) and covers specified needs. For example, patient discharge letters and Radiology reports can be inserted into the target database as RTF, DOC or plain text documents. Final diagnosis, admission date and discharge date are represented in simple fields. Medical images are represented as TIFF documents.
- Other data sources (external or not). For example there is a need to incorporate a complete list of ICD.

2.2 Data Definition and Manipulation Activities

2.2.1 Data Definition (DD)

It includes data types support, input procedures and data organization and store (storing structures) for:

- Data in text form (plain text, doc and RTF files),
- (Medical) Images (e.g. TIFF, JPEG),
- Structured (PDF, HTML, XML) Documents.

2.2.2 Data Cleaning (DC)

Data exported from the Hospital Information System (HIS) database(s) or other data inputs must be "cleaned". It means:

- Selection of the appropriate information, which is applicable to the specific topics of interest.
- Isolation of useful information.
- The essence of the DC process is the fact that the user feedback can cause successive repetitions. Therefore, there is a need for cleaning data stored in the (main) HIS database(s).
- Format conversion is desirable.

2.2.3 Back Flashing (BF)

The experience from the establishment and use of a data mart can drive to the re-organization of the main HIS database(s). Hence, we need batch procedures for storing back to DB(s) "cleaned" data.

2.2.4 Data Reformatting (DR)

After DC activities, there is a need for storing "cleaned" data into the data mart. It means that (batch) procedures must be conducted to support the loading process. In cases where the data fields in the target system are ordinary fields, the loading process can be a simple, ODBC based, exchange of data. In other cases, where the target documents (e.g. DOC files) must be created from a number of text rows of a table, OLE Automation seems to be a good solution and is implemented in CAIRN-DAMM.

3 AN OVERVIEW OF THE CAIRN ENVIRONMENT

CAIRN, the previous version of our system, is a medical information retrieval that allow physicians and students to store full text and multimedia medical information from any resource, organize and retrieve it. The most important feature of CAIRN (Karanikolas and Skourlas, 2000) is its capability to assist the user, physician, student, etc. in selecting documents against a submitted query in Natural Language.

The design of CAIRN system was focused on the concepts of document management, document retrieval, medical imaging and integration with other popular, general purpose, software. The architecture of CAIRN, is described in figure 1.

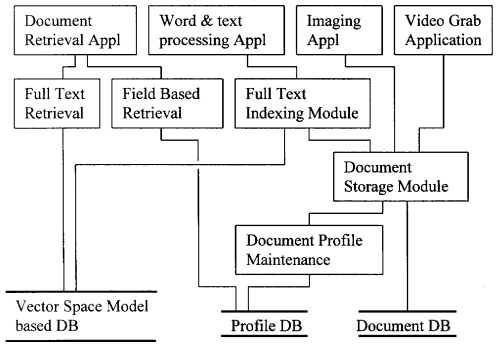

Figure 1: The architecture of CAIRN.

3.1 Document Management

It was decided that documents stored in CAIRN's database will not simply form a flat collection of unrelated documents. It is obvious that various documents can usually represent a patient's episode.

A patient can have more than one episode in different periods of time. Thus, there is a need to group all these related documents in a hierarchical tree structure. CAIRN implements a tree structure of categories where each level has a different number of fields that characterize the category. The fields that characterize the first hierarchy level can be demographic elements of patients while the fields that characterize the second hierarchy level can be elements that describe a specific episode (e.g. admission date, discharge date, final diagnosis and so on). The leaves (documents) in this hierarchy can also be characterized by a number of fields. This set of fields is called document profile or simply profile.

Document Management and retrieval can be conducting by giving values / constraints for the various fields of the document. An interesting feature of CAIRN is the support of both *multiple-value (multi-row)* and *single-value* fields. Data types supported by the CAIRN system include *integer* and *real* numbers, *dates*, *strings*, *lookup* fields and *multiple attribute* (or *multi-column*) fields. The combination of a multiple attribute field that accepts multiple values derives the ability to use two-dimensional tables in the position of a single field.

It is also possible to store structured documents (e.g. PDF and XML documents), documents downloaded from the Internet (HTML documents), etc.

3.2 Document Retrieval

Document retrieval can be based on fields' values of the documents. Full text retrieval (Karanikolas and Mantzaris 1992, Kowalski 1997, Salton 1983) can also be conducted or both. Figure 2 presents the use of the "Search/Retrieve" module of CAIRN.

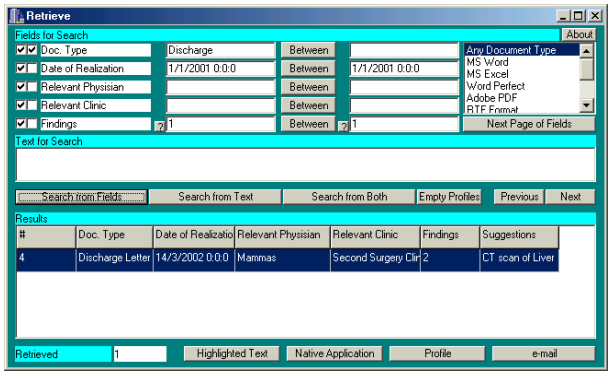

Figure 2: The "Search/Retrieve" module of CAIRN.

Full Text Retrieval can be conducted by typing simple natural language queries describing the topics of interest. The multilingual character of CAIRN permits the indexing and retrieval of multilingual text(s) and also supports multilingual queries.

The retrieval of documents can also be realized in a hierarchical manner where the user browses through the different levels of a tree structure and selects an instance of a category in order to retrieve all ancestor documents to this instance.

3.3 Imaging module

CAIRN's «Imaging module» permits the interconnection with any "TWAIN source". Formats supported include TIFF (simple and multi-page), JPEG, etc. It is possible to store digital images (digitized, scanned, imported from digital camera). CT Scan slides (digitized with a film scanner) can also be handled. Another usage of CAIRN' s "Imaging Module" could be for digitizing of a reference letter, signed by a General Practitioner, and given to the patient for a potential appointment with hospital doctors. This module of CAIRN permits the interconnection of documents with images from any TWAIN source. Using the "Save To DM" or "Save As" buttons the end users can store the multi page TIFF documents, or JPEG images, etc., into CAIRN database or into ordinary operating system files.

3.4 Integration with other popular general purpose software

The users of the system can also work with their favorite application programs (e.g. MS Word, Excel etc.) and then to store their documents into the database. The ability to store documents from office automation software packages directly to CAIRN permitted us to extend the Electronic Patient Record (EPR) with information that could not be embodied in an ordinary Hospital Information System (HIS). For example, reports from the pathologoanatomical laboratory, written in MS-Word, that could not be stored into HIS, now, can be directly embodied into CAIRN.

The interconnection of CAIRN with e-mail client is also supported. Such an interconnection offers the possibility of e-mail messages and documents' (attachments) exchange between the different units of the hospital(s).

4 DOCUMENT CLASSIFICATION BASED ON UNCONTROLLED TERMS

The diagnosis classification of patient discharge letters could be done using various techniques. The method described in this section could also be applied in other, not medical, domains. It is based Knowledge Discovery (Data mining) methods.

All patient discharge letters share a common structure (form). It is written as a plain text (in the worst case) or is written (exclusively) using some uncontrolled terms. Uncontrolled terms, for the librarians, are not a form of plain text but phrases like keywords. The difference is that keywords have to (and are checked to) belong to a specific authority list (or in a thesaurus). Hence, a list of uncontrolled terms has to be specified and then, using various methods an authority list of keywords can be formed and maintained.

Recent characteristics of CAIRN (Karanikolas and Skourlas, 2002) offer possibilities for extracting and proposing, even from the plain text, uncontrolled terms (Amini 2001, Chuang et al 2000, Helma et al 2000). These terms were used as the source for a first method to attack. Hence, some simple statistics were extracted related to the terms used by specific doctors, who wrote and signed such letters. Results were given to the doctors and discussed and were eventually integrated in a list for all the doctors in the Second Surgery Clinic of ARETEION. The rationale of using this method was our attempt to organize in a more systematic way all these uncontrolled terms, eliminate spelling errors, and incorporate a subsystem of handling keywords in the target CAIRN-DAMM system.

Moreover, we decided to measure the existence of phrases (uncontrolled terms) in patient discharge letters having a common ICD-9 final diagnosis. In table 1, some of the uncontrolled terms, extracted from discharge letters, and their relationship with ICD-9 code(s) are presented.

Table 1- Uncontrolled term <*ΕΞΕΡΓΑΣΙΑ ΗΠΑΤΟΣ> and its relationship with ICD-9 codes.

ICD-9 code	ICD-9 diagnosis	%
155	Malignant neoplasm of liver and intrahepatic bile ducts	10
155.0	Malignant neoplasm of Liver, primary	20
155.2	Malignant neoplasm of Liver, not specified as primary or secondary	50
157.8	Malignant neoplasm of pancreas, Other specified sites of pancreas	10
197.7	Secondary malignant neoplasm of Liver, specified as secondary	10

Having an uncontrolled terms authority list, we automatically created a vector for each document (from a selected training set) where each item of the vector represented the existence or not of the corresponding uncontrolled term in the document. These labelled data permitted us to apply a Text Data Mining (TDM) algorithm and to produce a set of classification rules (Hearst, 1999). Each classification rule is based on the existence of some uncontrolled terms in order to propose some ICD-9 diagnosis.

The following figure (Figure 3) describes the architecture of Document Classification subsystem of CAIRN-DAMM.

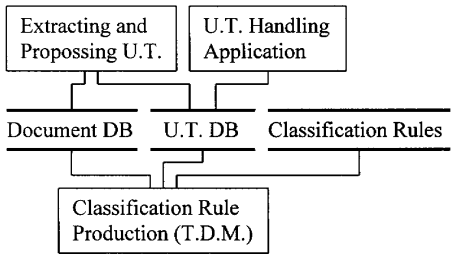

Figure 3: The architecture of Document Classification

5 CAIRN-DAMM ARCHITECTURE

The CAIRN-DAMM environment (and the related tools) was designed with the following targets:

- to store full text, medical images, structured documents, etc., from various information sources
- to support the well-known types of documents that are created using general purpose application software,
- to permit field based retrieval, free text retrieval and hierarchical retrieval,
- to permit data and documents extraction using various formats or types of documents,
- to propose salient uncontrolled terms (phrase patterns) and construct authority lists,
- to support automatic / semi-automatic classification based on the existence of uncontrolled terms in text documents and
- to provide a single source for patient data used for every day practice and problem solution but also for decision support activities.

Most (except the last) targets were satisfied by the previous CAIRN system in conjunction with the new Document Classification subsystem. In order to

satisfy the last target, we adopted the Data Mart, Data Warehouse concepts presented in the second section. The new subsystem, that adds a Data Mart aspect to CAIRN-DAMM, is described in the following figure (Figure 4).

There are two applications in this subsystem. Data and Rule Definition Application permits the users to define data sources and data cleaning and reformatting rules. The Loading Application does the actual loading of cleaned and reformatted data from legacy HIS to CAIRN-DAMM database.

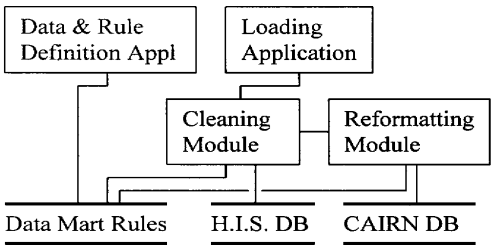

Figure 4: Data Mart Architecture.

6 CONCLUSIONS

The CAIRN-DAMM environment, based on various concepts, is presented. A description is given to a Diagnosis - Aid module and the methods incorporated for supporting the diagnosis. Some description is also given to methods that permit the loading of CAIRN-DAMM with non-volatile data from legacy HIS. Therefore CAIRN-DAMM can be used as a single source for data retrieval and decision support activities.

Apart from the, above mentioned, advantages (text retrieval, automatic diagnosis classification, multimedia documents, broader use and acceptance of the system) implied by the use of CAIRN_DAMM, we work to improve the new system in order to offer:

1. a solution to the problems implied by the usage of the legacy system, e.g. Clinical departments have difficulties in using the complex and text only based electronic patient record of the legacy system,
2. a gradually substitution / replacement of the present restricted (only text based) medical file handling with the multimedia patient data handling offered by the CAIRN system,
3. a possibility of using, in the future, a new flexible tailor made special purpose software more oriented to the needs of the Personnel and Financial departments of the Hospital.

REFERENCES

Amini, M.R. and Gallinari P., 2001. Automatic Text Summarization Using Unsupervised and Semi-supervised Learning. In *PKDD'2001, 5th European Conference on Principles of Data Mining and Knowledge Discovery*.

Chuang, W.T. and Yang, J., 2000. Extracting Sentence Segments for Text Summarization: A Machine Learning 99Approach. In *SIGIR'2000, 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Hearst, M.A., 1999. Untangling Text Data Mining. In *ACL'99, 37th Annual Meeting of the Association for Computational Linguistics.*

Helma, C., Gottmann, E. and Kramer, S., 2000. Knowledge Discovery and Data Mining in Toxicology. *Statistical Methods in Medical Research 9.*

Karanikolas, N.N. and Mantzaris, S.L., 1992. Innovative directions in information retrieval. In *HERMIS'92, Hellenic Research on Mathematics and Informatics, Athens.*

Karanikolas, N.N. and Skourlas, C., 2000. Computer assisted Information resources navigation. *Medical Informatics and the Internet in Medicine 25.*

Karanikolas, N.N. and Skourlas, C., 2002. Automatic Diagnosis Classification of Patient Discharge Letters. In *MIE'2002, 23rd International Congress of the European Federation for Medical Informatics, Budapest.*

Kowalski, G., 1997. *Information Retrieval Systems. Theory and Implementation*, Kluwer Academic Publishers. USA, 1st edition.

Lenz, R., Blaser, R. and Kuhn, K.A., 1999. Hospital information systems: changes and obstacles on the way to integration. *Studies in Health Technology and Informatics 68*.

Lenz, R. and Kuhn, K.A., 2001. Intranet meets hospital information systems: the solution to the integration problem? *Methods of Information in Medicine 40.*

Lavrac, N., 1999. Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine 16*.

Ricciardi, T.N., Masari, F.E. and Middleton, B., 2001. Clinical benchmarking enabled by the digital health record. *Medinfo 10.*

Salton, G., 1983. *Introduction to Modern Information Retrieval*, McGraw-Hill. USA, 1st edition.

Verma, R. and Harper, J., 2001. Life cycle of a data warehousing project in healthcare. *Journal of Healthcare Information management 15.*