Medical Informatics Europe '99
P. Kokol et al. (Eds.)
IOS Press, 1999

Data warehousing as a basis for web-based documentation of data mining and analysis

Johan Karlsson[1], Patrik Eklund[1], Carl-Gunnar Hallgren[2], Jan-Gunnar Sjödin[3]
[1] *Umeå University, Department of Computing Science, SE-901 87 Umeå, Sweden*
[2] *Ersboda Health Centre, SE-906 25 Umeå, Sweden*
[3] *Umeå University, Department of Urology and Andrology, SE-901 85 Umeå, Sweden*

Abstract. In this paper we present a case study for data warehousing intended to support data mining and analysis. We also describe a prototype for data retrieval. Further we discuss some technical issues related to a particular choice of a patient record environment.

1. Introduction

In this paper we discuss the establishing of a data warehouse for the purpose of enabling high quality and fast documentation of data analysis. We will discuss the end-user perspective in a particular clinical environment. These developments demonstrate the benefits of retrieval tools in conjunction with support for generated web-based documentation.

Developments are done related to a regional record system used in clinics and health centres within the County Council of Västerbotten in Northern Sweden. This work is part of data warehousing and data mining developments for these record systems. Generation of documentation, involving aspects of guideline development, is discussed in [6].

Work on data warehousing as described in this paper has been inspired also by successful warehousing for patient administrative information in Västerbotten [15]. This system includes regular downloading of information and uses standard software for statistics and reporting. A comparable system was reported in [19] for the provision of various service statistics. A retrieval system oriented towards elaborated user-interface design together with selections from vocabularies can be found in [12].

Developments in this paper are partly based on a previous approach to data warehousing as described in [5], where a data retrieval system for MUMPS based record systems was developed for use within a laboratory environment. Search criteria given by users were converted to querying code, and resulting data files delivered to the end-user, thus minimising retrieval time from the end-user point of view. A similar approach was reported in [2], which describes graphical query generation based on object-oriented user views.

Note the distinction between our presentation of documentation and other approaches related to medical reporting which are based on concept representation, see e.g. [13, 7, 11]. Our approach is, however, in a natural and obvious way related to terminology as the record system includes parts of terminology. Concerning software and technological issues of terminologies, our influence comes from developments within Spriterm [14]. These approaches obviously are in no conflict, as the latter represents more of a top-down approach, and the warehousing bottom-up approach needs to be grounded to technological issues with terminologies in the forefront. In the bottom-up approach we always need to deal with technological platforms, and even more so, always have to anticipate involvement of neighbouring technologies, e.g. the inclusion of information residing in PACS/RIS systems.

A prototype including image retrieval in PACS installations is found in [17], and possibilities of using meta-patient-record approaches is described more generally in [18]. In the warehousing approach we also need to build upon standards, thus being dependent on record system middleware components as developed in [8, 16].

2. Case study: Lower urinary tract symptoms

The initial target group for the data warehouse is male patients with LUTS (Lower Urinary Tract Symptoms). These patients have symptoms such as a weak or interrupted flow of urine, difficulty urinating, frequent urination, pain or a burning sensation during urination, blood in the urine, urinary leakage etc. The symptoms are often vague and can have many causes. Patients may have conditions such as prostate cancer, enlarged prostate (Benign Prostatic Hyperplasia, BPH) or combinations of the two, neurological diseases and diabetes. Treatment of LUTS depends on the diagnosis and degree of symptoms. Thus from watchful waiting to major surgery.

The data that will be included in the data warehouse will be acquired from databases at the urology department of the University Hospital in Umeå. The patients that had an enlarged prostate and were operated upon, were more extensively tested and the data were subsequently more complete. For that reason, the initial data warehouse will only contain data on such patients. Discussions with the physicians at the department have resulted in a set of laboratory tests that was considered important for diagnosing these patients; PSA (Prostate Specific Anti-gen), IPSS (International Prostate Symptom Scoring) given by the patient before and after the operation, flow voided volume (urine flow, ml/s), remaining urine (amount of urine that remains in the bladder after urination, ml), blood in urine (yes/no) and IPSS difference (as measured after the operation).

3. Web-based documentation

In general, the process of producing web-based documentation contains modules for respectively retrieval of data, acquisition and representation of knowledge, and creation of the final web-based system. Knowledge acquisition in form of decision structure identification, as provided by the process of mining, enrichment and refinement of data into decision support, obviously invites to a vision of a automated documentation within which evidence-based documentation including corresponding executable DSSs are produced by end-users for end-users. Initial suggestions for a system architecture was suggested in [9].

Decision support usually means text-based guidelines. But since the documents that the proposed system will produce are web-based, executable elements can be added. We believe this will result in more dynamic guidelines that will give additional support to the end-users. Historically, expert systems have been developed during close co-operation between a domain expert and a computer engineer. It is our contention that various computational models, either of feedforward or inference type, can serve as underlying mechanisms in the expert systems. These models would of course only be applicable to certain diseases. They would be based on data acquired from a data mining module. Since the documentation is web-based, important aspects such as availability and easy maintenance are met.

4. Technical aspects

Data mining involves a wide range of technicalities, which in an integrated environment must be considered in relation to existing systems. For data warehousing, and in particular for the retrieval systems, development platforms must be carefully selected so as to provide an efficient environment for development and maintenance of systems, and at the same time not to disable integration capabilities.

In order to maintain platform independence, we use Java technology [1]. The Java platform was first announced in 1994, where one of its main advantages was the "write once, run everywhere" philosophy that allowed developers to construct programs that would run on any platform with a Java virtual machine (JVM) without recompilation. The aspect of platform independence has recently been made topical when the County Council of Västerbotten decided to change operating system to Windows NT. The task of updating existing medical applications to this new platform will demand considerable time and resources. Applications developed in Java would, if written correctly, be immediately available for use under the new operating system.

We also need a flexible way to access the data warehouse. JDBC (Java Database Connectivity) is an API that provides database access from Java applications or applets. It abstracts the necessary functions, thus allowing platform independent access to databases. If additional technology, such as JNDI (Java Naming and Directory Interface), is used, applications can retrieve information about the database, e.g., its name. This makes the applications even more dynamic.

Additional work will be necessary if the structure of the database is changed, i.e., if tables or fields are added or changed. The use of a repository, as suggested in [10], is a typical suggestion in order to minimise the work required in such cases.

The urology clinic at the University Hospital, Umeå, uses BMS Cross from IBM for managing their patient records. This system is based upon the DB/2 database engine. As most vendors, also IBM supports JDBC.

Some data in the records is written as free text, and thus some problems will arise when such data is extracted for inclusion in the data warehouse. However, the hospital is currently introducing new routines for journal entry. The content will be tagged to make the data more accessible and exchangeable. The record system complies to using terminologies, even if the user is not fully enforced to include structured data only.

5. A prototype system for data retrieval

The data warehouse must include efficient, flexible and user-friendly retrieval tools. Retrieval can be broken down in steps shown in Figure 1.

A description of a disease is acquired from a medical expert. This description will typically consist of the selected disease together with the laboratory tests that are important when diagnosing the disease. This information is translated to a SQL-statement by the code generator. A retrieval program is transferred over to the information pool (data warehouse) and a raw data set is returned.

Since physical input in itself does not always indicate all the dynamics of the measurements, it is necessary to organise and combine the data. One example where this is important is when symptom values vary over time.

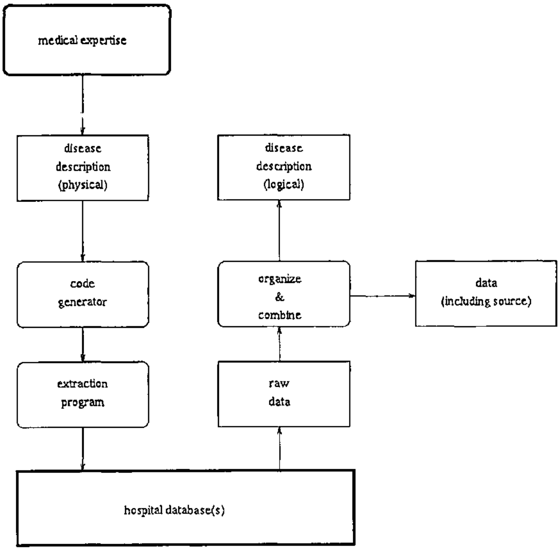
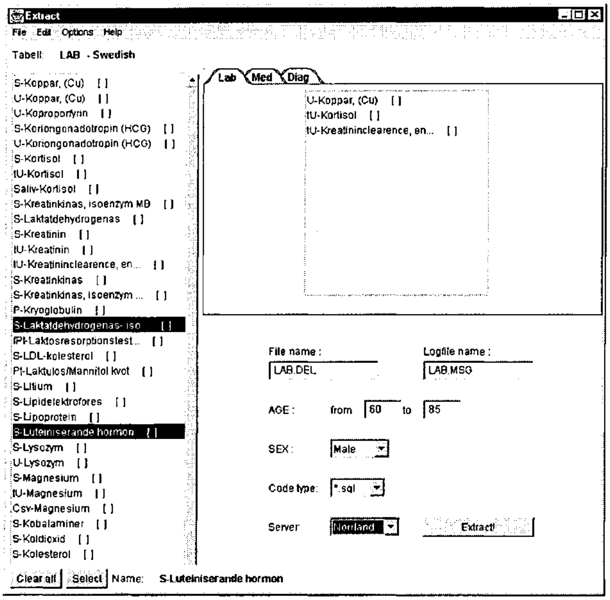


Figure 1

A prototype retrieval program is shown in the Figure 2.



Tabell: LAB - Swedish

S-Koppar, (Cu) [ ]
U-Koppar, (Cu) [ ]
U-Koproporfyrin [ ]
S-Koriongonadotropin (HCG) [ ]
U-Koriongonadotropin (HCG) [ ]
S-Kortisol [ ]
tU-Kortisol [ ]
Saliv-Kortisol [ ]
S-Kreatinkinas, isoenzym MB [ ]
S-Laktatdehydrogenas [ ]
S-Kreatinin [ ]
tU-Kreatinin [ ]
tU-Kreatininclearence, en... [ ]
S-Kreatinkinas [ ]
S-Kreatinkinas, isoenzym ... [ ]
P-Kryoglobulin [ ]
S-Laktatdehydrogenas- iso [ ]
fPt-Laktosresorptionstest... [ ]
S-LDL-kolesterol [ ]
Pt-Laktulos/Mannitol kvot [ ]
S-Litium [ ]
S-Lipidelektrofores [ ]
S-Lipoprotein [ ]
S-Luteiniserande hormon [ ]
S-Lysozym [ ]
U-Lysozym [ ]
S-Magnesium [ ]
tU-Magnesium [ ]
Csv-Magnesium [ ]
S-Kobalaminer [ ]
S-Koldioxid [ ]
S-Kolesterol [ ]

U-Koppar, (Cu) [ ]
tU-Kortisol [ ]
tU-Kreatininclearence, en... [ ]

File name : LAB.DEL
Logfile name : LAB.MSG
AGE : from 60 to 85
SEX : Male
Code type: *.sql
Server Norrland

Name: S-Luteiniserande hormon

Figure 2

This system enables a user to browse and select various laboratory results, medicines and diagnoses. The application produces an SQL request that is exported to the data warehouse.

The prototype, written in Java, was developed earlier to fit another patient record environment, and JDBC capability is now being added.

Acknowledgements

This work is funded by the Foundation for Knowledge and Competence Development (KK-foundation, Sweden, http://www.kks.se) and is part of IT in health and health care within the area of IT-R&D. Research and development is related to work within the Centre for Medical Informatics and Telematics in Norrland, CMIT (http://www.cs.umu.se/~medinfo/cmit), covering health care IT needs in northern Sweden.

References

[1] Arnold, K., Gosling, J.: The Java™ Programming Language, Addison Wesley (1998)

[2] Banhart, F., Klaeren, H.: A Graphical Query Generator for Clinical Research Databases. Methods Inf Med, 34 (1995) 328-339

[3] van Bemmel, J., Musen, M.A. (eds.): Medical Informatics. Springer-Verlag, Berlin Heidelberg New York (1997)

[4] Ceusters, W., Spyns, P., De Moor, G., Martin, W. (eds.): Syntactic-Semantic Tagging of Medical Texts: The Multi-TALE Project. IOS Press, Amsterdam Berlin Oxford Tokyo Washington, DC (1998)

[5] Eklund, P., Forsström, J., Holm, A., Nyström, M., Selén, G.: Rule Generation as an Alternative to Knowledge Acquisition. Fuzzy Sets and Systems, 66 (1994) 195-205

[6] Eklund, P., Karlsson, J.: Data mining and structuring of executable data analysis reports: Guideline development and implementation in a narrow sense. Submitted.

[7] GALEN project. Home page URL: http://www.cs.man.ac.uk/mig/galen. (1997)

[8] HANSA project. Home page URL: http://www2.echo.lu/telematics/health/hansa.html. (1998)

[9] Karlsson, J.: A Generic System for Developing Medical Decision Support. Graduation thesis for MSc degree, UMNAD 232.98, Umeå University, Department of Computing Science (1998)

[10] Niinimäki, J., Selén, G., Kailajärvi, M., Grönroos, P., Irjala, K., Forsström, J.J.: Medical Data Warehouse, an Investment for Better Medical Care. In: Brender, J., et al. (eds.): Medical Informatics Europe '96. IOS Press (1996) 766-770

[11] Moorman, P.: Towards Formal Medical Reporting. Thesis. Erasmus Universiteit Rotterdam, (1995)

[12] Poon, A.D., Fagan, L.M., Shortliffe, E.H.: The PEN-Ivory Proj-ect: Exploring User-interface Design for the Selection of Items from Large Controlled Vocabularies of Medicine. JAMIA, 3 (1996) 168-183

[13] Rector, A.L (ed.): Terminology and Concept Representation (Theme Issue). Artif Intell Med, 15 no 1 (1999)

[14] Spriterm 2.01. The Swedish Institute for Health Service Devel-opment (December 1998)

[15] Strukturanalys 2000 (projekt). Västerbottens läns landsting. (1998)

[16] SYNAPSES project. Home page URL: http://www.cs.tcd.ie/synapses/public/index.html. (1997)

[17] Taira, R.K., Johnson, D.B., Bhushan, V., Rivera, M., Wong, C., Huang, L., Aberle, D.R., Greaves, M., Goldin, J.G.: A Concept-based Retrieval System for Thoracic Radiology. J Digit Imaging, 9 (1996) 25-36

[18] Weser, A.J., Kleinholz, L., Söderström, P.O., Eklund, P., De Moor, G., Williams, H., Venters, G., Mahr, B.: In: Brender, J., et al. (eds.): Co-operative Health Information Networks "CHIN". Medical Informatics Europe '96. IOS Press (1996) 1068-1072

[19] Woelk, G.B., Moyo, I.M.: Development of a Computerized In-formation System in the Harare City Health Department. Methods Inf Med, 34 (1995) 297-301