Proceedings of the 3rd INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2008)
J. Li, D. Aleman, R. Sikora, eds.

NursingCareWare: Warehousing for Nursing Care Research and Knowledge Discovery

Ray Hylock
Management Sciences Department
University of Iowa
Iowa City, IA 52242
ray-hylock@uiowa.edu

W. Nick Street
Management Sciences Department
University of Iowa

Der-Fa Lu
College of Nursing
University of Iowa

Faiz Currim
Management Sciences Department
University of Iowa

Abstract

Widespread adoption of healthcare information systems has led to large amounts of patient care data with the potential for improvements in the cost and quality of patient care. However, extracting useful information from these large data sets using current systems is a cumbersome process and it is difficult to effectively utilize the data. This paper addresses is the creation of a multidimensional design to manage complex patient care data. We propose and implement a comprehensive warehouse model that includes nursing diagnoses, outcomes, and interventions. Unlike current systems, our design incorporates all the aforementioned aspects and accommodates their inherent multi-valued nature. The objective is to uncover useful patterns involving existing patient care plans, as well as symptom and disease associations across demographic groups. It is envisaged that this will lead to better treatment plans, and ultimately improve the quality of patient care and the cost of providing it. We describe various warehousing challenges that arise during the development of the warehouse and our plans for evaluation.

Keywords: Clinical data warehousing, data integration, data mining, nursing care plans

1 Introduction

There is a clear distinction between traditional relational databases designed for online transaction processing (OLTP), and data warehouses which are designed to support data analysis. A relational database is optimized for continuously inserting, updating, and deleting records. However, when performing complex queries and aggregations over a wide range of data sources, the time and physical resources required substantially increase and the ability to obtain results in a timely and consistent manner significantly diminishes. Data warehouse systems using Online Analytical Processing (OLAP) technology, on the other hand, are designed to overcome this limitation. In the medical domain, while everyday processing of healthcare information should be done with OLTP systems, analysis of the data needs to be done on a clinical data warehouse (CDW) platform intended for intensive and complex data computations.

Previously, we designed an OLTP system for a nursing dataset [7]. The data was generated at a large regional hospital in the Midwest where over 600 registered nurses recorded information about their care plans. In this paper, we build on that effort and describe the multidimensional design, extraction, transformation, and loading (ETL) process, and materialization strategy for the corresponding clinical data warehouse.

Our contribution includes a comprehensive design for analyzing nursing patient care treatment plans. This is motivated by the desire of clinical researchers to have access to all pertinent information (so nurses, doctors, administrators, and other end users can make decisions with a high degree of certainty based on a maximal number of facts). Related to this, we tackle the problem of managing multi-valued dimensions, and developing an upper bound on the number of materialized cuboids arising from the dimensionality. We also advocate the transition to standard North American Nursing Diagnoses Association (NANDA) [10], Nursing Interventions Classification (NIC) [3], and Nursing Outcomes Classification (NOC) [8] classification codes. As a consequence, our system does not suffer from limitations such as a restrictive subset of dimensions or partial storage of data within dimensions. By doing so, we increase the size of our data warehouse which leads to ETL as well as data reduction and compression challenges. However, we believe that the benefits of such a design (increased scope of data, ability to answer complex queries accurately, and improvements in the quality of patient care) justify the complexity. The implemented warehouse will allow visual browsing of the data via an OLAP server, as well as efficient correlation mining in the space of nursing diagnoses, interventions, and outcomes. In this way we hope to uncover frequent (or surprisingly frequent) interactions among the items, and compare the results with standardized nursing sources as a means of quality control. The addition of outcome scores would allow the comparison of different intervention strategies and a search for subpopulations with unusual response patterns.

2 Literature Review

A few CDW designs have appeared in the literature in recent years. For example, Berndt et al. [2] created a system called the CATCH data warehouse which collects, organizes, analyzes, prioritizes, and generates reports for the CATCH indicator system used in Florida. Their CDW, however, only stores the top 10 ICD's and Procedures along with age, race, and gender whereas we store all ICD's through the use of bridge tables. Similarly, Hristoviski et al. [6] produced a CDW for the Public Health Institute of the Republic of Slovenia using outpatient data. Like Berndt et al., their goal was to convert an existing system that was virtually impossible to extract answers to ad-hoc queries from and provide the means to do so. The focal point of the system is the patient, and ICD-10 is the primary means of evaluation, which is but one of the dimensions we incorporate. Pedersen and Jensen [11] built a CDW with patients as the central subject. They consider diagnoses and treatments (a subset of the dimensions we store) and focus solely on diabetes. They stress the importance of maintaining multi-valued dimensions and the need for strong data modeling features (an emphasis we agree with, an extend to consider all possible nursing care plans). To summarize, existing warehouses focus either on operational activities such as billing, or are used to track (occasionally, recommend) treatment plans in a limited way. Those in the latter category typically use a small number of dimensions or a finite amount of vital information (e.g., only 10 diagnoses per patient).

3 Data Warehouse Design

Our warehouse design is by no means limited to just our data set. The listed attributes and dimensions stem from the existing data elements as well as the coding schema based on the standardized Uniform Hospital Discharge Data Set (UHDDS). Although the current codes do not conform to NANDA, NIC, and NOC, we have modeled those standards and their inherent hierarchies into our design. We are currently in the process of converting this data set into those

standards (to allow for better normative analyses, as well as future integration of data across years and from multiple datasets). This system is based around the *Visit* event, but that can easily be altered to accommodate the central entity of the existing schema. Also, many elements, such as the patients and date dimensions, can be found in most OLTP systems, requiring only minor modifications to the multidimensional design below.

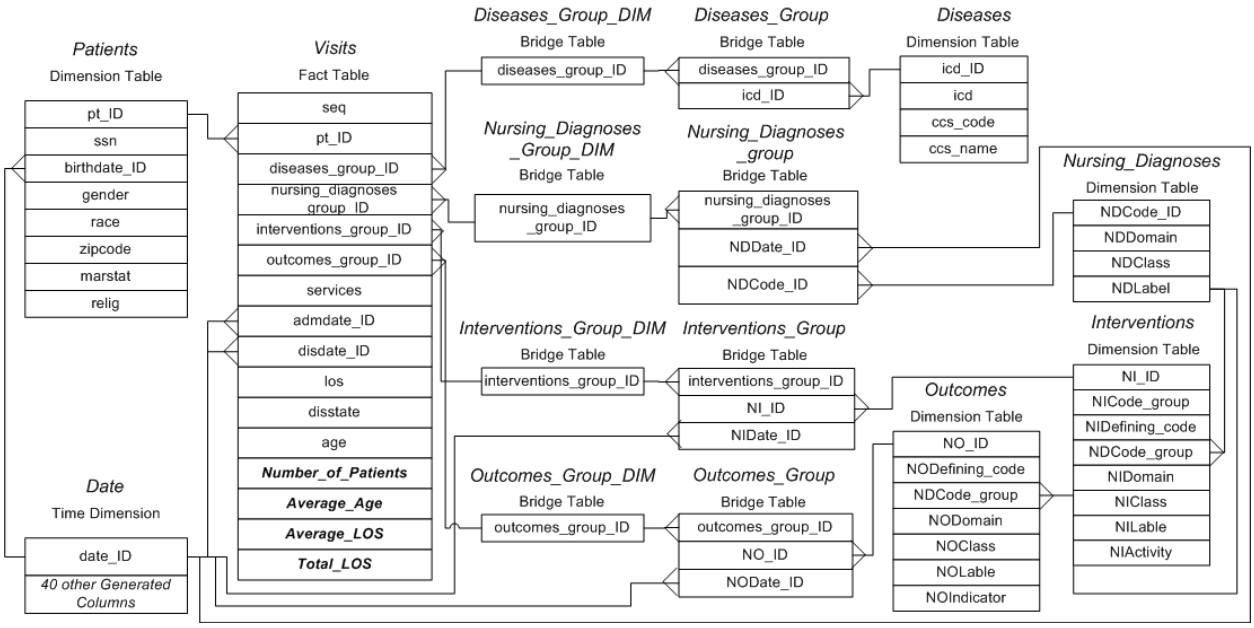

Figure 1: Full multidimensional design

3.1 Fact Table

The center of this model is the Visits fact table (Figure 1). Visits was chosen since the data was primarily connected with the visit event. Normally, the primary key for the fact table is the set of all foreign keys. However, in this instance those values do not guarantee uniqueness. Therefore, SEQ (a degenerate dimension and a unique identifier assigned to each visit) was added to ensure entity integrity. The attributes uniquely assigned to each visit are: services given, length of stay (los), discharge state (disstate), and age. The measures (bold and italicized) listed are simple examples and can be altered to suit the needs of the user.

3.2 Dimension and Bridge Tables

Figure 1 shows the dimensions implemented in the current system; a more comprehensive clinical dataset would result in additions to this framework. A brief description of each dimension, and its connection to surrounding dimensions, is presented below.

The patients dimension stores all of the patient data (e.g., gender, race, and zip code). Diseases, Nursing_Diagnoses, Interventions, and Outcomes store unique codes and hierarchical information, facilitating a "lookup table" effect. Also, Nursing_Diagnoses is connected to Interventions and Outcomes because they are related to a specific nursing diagnosis. We cannot directly connect Visits to these multi-valued dimensions, since, for example, a patient can have one to many nursing diagnoses per visit. Therefore we use bridge dimensions (labeled using "_Group") to preserve the one-to-many relationship between a single visit event and a multi-

valued dimension. These tables store the actual data per visit (the data unique to an individual visit, e.g., the date in which it was entered). The data is also grouped by all attributes except the identifier to decrease the size of the dimension. This means that if patient A and patient B have the same set of nursing diagnoses (with respect to a single visit), we store one set of records for both patients (see section 3.4). These bridge dimensions are connected to the Date dimension because each code is entered on a specific date. This will allow the user to roll-up and drill-down by the established hierarchy. Linking the bridge dimensions and the fact table are dimensions that hold distinct grouping identifiers, in order to maintain a one-to-one relationship between a specific visit and all connecting dimensions. These dimensions (labeled using "_Group_DIM) are necessary to ensure referential integrity since the primary key of each bridge is composite.

3.3 Concept Hierarchies

Figure 2 shows the six concept hierarchies used in our CDW to date. The hierarchies are composed of levels (circles) that can be used to roll-up and drill-down upon in order to adjust the level of granularity for a particular item. For example, Nursing Diagnoses has levels: Label (most specific) → Class → Domain (most general) forming its (only) hierarchy.

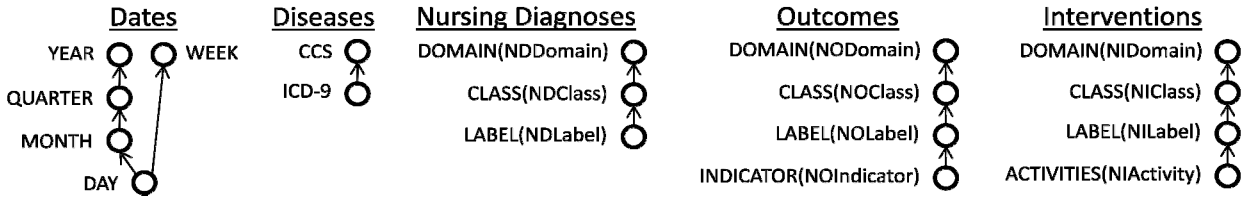

Figure 2: Concept hierarchies

3.4 ETL Challenges

The ETL process is a significant challenge for the four bridge table groups since we need to reduce the overall size (in number of records) of the tables. We extracted all of the data and selected only the distinct records to store while maintaining a list of visit events (based on SEQ values) associated with each record in order to update the Visits table's foreign keys. Since they follow the same process, only one dimension (diseases) will be outlined.

The Diseases dimension is populated by joining three external tables (corresponding to ICD codes, CCS codes, and the mapping between them) and inserting only the distinct results. To accomplish this, a temporary table is used to flatten out the diseases by SEQ. While a patient with 10 diseases will have 10 tuples in the relational view, in the temporary table, they will have 1 entry with 10 attributes. This allows us to order and then compare the records. For each distinct group of codes, a diseases_group_id is incrementally generated and a record is inserted into the Diseases_Group_DIM dimension and the visit(s) associated with the distinct group are updated with the generated ID. By grouping the data, we have achieved between a 0.037% and 41.98% reduction in the number of records stored in a bridge dimension without loss of information. The more attributes involved, the lower the reduction percentage. We have experimented with mini-dimensions to group similar factors (e.g., date) from all bridge dimensions, however, the physical storage space increases on average 12.6%. Although this might be a successful tactic in warehouses with one-to-one relationships, multi-valued dimensions require additional attributes for this to work, negating any potential gains. More advanced techniques, likes those employed

in RFID systems, can be used to further compress the data. Currently, only one data source [7] is being used so we have not yet addressed any data integration issues such as naming conventions, encoding discrepancies, and physical table structure differences.

3.5 Materialization

Materialization is vital to the success of any warehouse system. The proper selection of table joins and attribute aggregation can significantly improve the speed of the analytical process. The view selection problem (VSP), however, is NP-Complete [5]. There are many published approaches (e.g., data mining [1], bottom-up with top-down [12], and cost analysis [5]; see survey in [9]) to solve this problem. Most require knowledge of the queries. When determining the approach to be used, knowing the upper bound gives one an idea of the selection space one is working in. If the bound is extremely high, then approximation methods are used instead of direct optimization. Current equations in literature view all dimensions as independent and yield, for our model, 921,600 possible cuboids. This might lead one to believe that approximation is the only course of action. However, bridge group dimensions are in fact dependent upon one another. With this in mind, we propose equation (1) to calculate the number of possible cuboids:

$$\prod_{i=1}^{b}[(\prod_{j=1}^{b_i} L_j)_i + 1], \quad (1)$$

where b = the number of non-bridge dimensions + the number of bridge groups, b_i = the number of dimensions in bridge group i or 1 if i is not a bridge group, and L_j = the number of levels in bridge group i's jth table. Using equation (1), the total number of cuboids is 3,600; a more manageable upper bound. Since we do not have the luxury of knowing the queries in advance, we adopt a top down approach and materialize cuboids we believe will be used or aggregated from directly. The cuboids currently materialized are: each bridge group, each bridge group (lowest level) to Visits, Visits to Outcomes grouping (lowest level) with Nursing Diagnoses grouping, and Visits to Interventions grouping (lowest level) with Nursing Diagnoses grouping. We are currently composing a text mining algorithm to parse query logs and determine the tables, aggregation operators, and levels to materialize.

4 Evaluation

To evaluate the system, we will conduct an experiment using targeted end users (e.g., clinicians,, senior nurses, and doctors; all meeting minimum requirements) who will use the system and then complete a survey to measure their satisfaction level. Potential questions include: "Did the system return the level of detail you expected?", "Was the system easy to understand and use?", and "Were you satisfied with the response time (i.e., relative to the normal query development process)?" We anticipate both qualitative and quantitative variables. Preliminary feedback has been positive and the sense from end-users is they anticipate significant improvement in their ability to access data, and make better decisions to improve the safety, cost and quality of care.

5 Future Work

We have developed a prototype recommender system [4] that provides clinicians with a web-based tool to enter care plans, and provides interactive suggestions of commonly associated items. This tool must interface with the data warehouse and extract the information necessary to facilitate its objectives. Also, we will need to implement and evaluate the text-mining based materialization algorithm for accuracy and, to a lesser extent, speed. Furthermore, we will be

increasing the granularity of the data stored from days to hours to accommodate queries that consider patient metrics aggregated to the resolution of hours. Since Oracle 11g Warehouse Builder's finest detail is day, we will need to devise a strategy and accompanying table structure and code to produce the new Date/Time dimension. Finally, the proper indexing and partitioning strategies for the dimensions and materialized views will need to be determined in accordance to the queries asked of the system. When working with large tables, partitioning and indexing increases the efficiency and speed of queries needing to be evaluated.

6 Conclusions

Large volumes of patient care data is recorded by most hospitals. Existing systems are designed for either a specific research question (e.g., diabetes) or for operational purposes; most do not take into consideration all significant dimensions and store a limited amount of information within each. Extracting useful information from these systems is a time-consuming, multi-step process making it difficult to effectively utilize data. We propose a comprehensive nursing design (encompassing, e.g., patient demographics and visits, nursing diagnoses, characteristics, outcomes, interventions, ICD-9 and CCS codes) that manages multi-valued dimensions that were previously considered in a limited way. We also converge to using standardized coding schemes and associated hierarchies to allow both nurse practitioners and researchers to analyze patient care treatment plans. Our system will lead to better clinical decision support and quality control.

References

1. Aouiche, K., Jouve, P., and Darmont, J., 2006, "Clustering-based materialized view selection in data warehouses," Lecture Notes in Computer Science, 4152, 81-95.
2. Berndt, D., Hevner, A., and Stundnicki, J., 2003, "The Catch Data Warehouse: support for community health care decision-making," Decision Support Systems, 35(3), 367-384.
3. Dochterman, J., Bulechek, G., 2004, "Nursing Interventions Classification (NIC)," 4th ed. St. Louis, MO., Mosby.
4. Duan, L., Street, W.N., and Lu, D.-F., 2008, "A nursing care plan recommender system using a data mining approach," Proceedings of the 3rd INFORMS DM-HI Workshop, accepted.
5. Gupta, H. and Mumick, I., 1998, "Selection of views to materialize under a maintenance cost constraint," Lecture Notes in Computer Science, 1540, 453-470.
6. Hristoviski, D., Rogac, M., and Markota, M., 2000, "Using data warehousing and OLAP in public health care," Proceedings of the AMIA Symposium, 369-373.
7. Lu, D., Street W. N., Currim, F., Hylock, R., and Delaney, C., "A data modeling process for decomposing healthcare patient data sets," Online Journal of Nursing Informatics, accepted.
8. Moorhead, S., Johnson, M., Maas, M,, 2004, "Nursing outcomes classification (NOC)," 3rd ed. St. Louis, MO., Mosby.
9. Morfonios, K., Konakas, S., Ioannidis, Y., and Kotsis, N, 2007, "ROLAP implementations of the data cube," ACM Computing Surveys, 39(4), article 12.
10. NANDA, 2005, "Nursing Diagnoses, Definition & Classification," Philadelphia, PA., NANDA International.
11. Pedersen, T. and Jensen, C., 1998, "Research issues in clinical data warehousing," Proc. of the 10th International Conference on Scientific and Statistical Database Management, 43-52.
12. Xin, D., Han, J., Li, X., and Wah, B., 2003, "Star-Cubing: computing iceberg cubes by top-down and bottom-up integration," Proceedings of the 29th VLDB Conference, 29, 476-487.