Data Warehouse Design For Knowledge Discovery From Healthcare Data

Aftab Ahmed, Kashif Zafar, Abdul Basit Siddiqui, *Umair Abdullah, *Member, IAENG*

Abstract— **Due to shifting of healthcare data in electronic form, huge volumes of data have been accumulated in organizations related to medical claim processing domain. This paper presents design of a data warehouse implemented as integral part of a rule based expert system. The system is being used for scrubbing of medical claims' data. Approximate star schema has been used for designing the proposed data warehouse. The data warehouse is being used by the production rule mining module and inference engine of the system. Rejected medical claims got corrected in operational database. Payments are received against corrected claims. Production rule mining module needs both versions of a claim, to learn new rules. Therefore data warehouse is required to store all states of medical claims. Moreover, efficiency of the rule based inference engine of the system has increased due to de-normalized format of data in the warehouse.**

Keywords; Data Minning, Data Warehousing, Rule Based Systems, knowledge engineering , production rule mining, and Medical Billing.

I. INTRODUCTION

Medical claim processing domain is very rich in terms of data, and knowledge. Most of the medical claim-processing organizations have shifted their data in electronic form. As a result huge volumes of medical claim related data have been accumulated in government and private organizations, thus attracting data mining researchers and scientists to apply their theories and concepts on the data and extract something useful from it [1]. A successful and applied data mining research work can provide considerable benefit to the society. A 'medical claim' also known as 'encounter' is generated when a patient visits a medical provider (doctor, surgeon, physician etc.) Medical claim contains information about the diagnosis and treatment of the patient along with demographic and insurance information. Medical claim is sent to insurance for reimbursement to provider. If any data is missing or is faulty, insurance rejects the claim. Rejected claim is sent to insurance again after making corrections and inserting missing values. Finally, insurances pay by processing the corrected state of the claim.

A rule based system expert system has been implemented at a USA based company working in medical billing domain [2]. System applies claim scrubbing rules on claims present in operational database. Simplified version of the architecture of the system is shown in Figure 1. Claims' data from different resources such as billing software, website of the company, HL7 (i.e. Health Level 7 format) files, synchronization server, is transferred to the operational database of the company. This data is refined by the rule based expert system which applies knowledge oriented data consistency checks. These checks are implemented in the form of production rules, which are edited and validated by domain experts using knowledge editor as shown in Figure 1. For success of the rule based expert system, quantity and quality of knowledge is vital. In order to increase quantity and improve the quality of production rules, a production rule mining module has been developed (also shown in Figure 1). The production rule mining module works on the data present in the data warehouse. Design of the data warehouse is the focus of this research work as indicated by magnifier glass in Figure 1.

Numerous clinical data warehouses have been reported so far [1], [3], developed for different purposes. Clinical data warehouse at the University of Virginia is a well-documented research oriented data warehouse [3]. It provides researchers, clinicians, administrators and other users with direct access to historic view of clinical and financial patient data. The data warehouse itself has 200 registered users engaged in clinical research, medical and quality management, and education. Clinical research consists of "hypothesis testing, practice pattern analysis, identification of candidates for clinical trials medical and quality management of outcomes analysis, physician profiling, risk assessment, cost containment" etc. [3]. Data warehouse presented in this paper is different from others as it focus operational environment instead of research or analysis.

Next section presents rational of why operational database or offline copy of operational database cannot serve the purpose of a data warehouse. It describes why a data warehouse is required for knowledge discovery from data. Section III presents design of the data warehouse implemented for storing medical billing data. Section IV describes utilization of the data warehouse and at the end after results and discussion, conclusion contains summary of findings.

Manuscript received March 18, 2013. This work has been supported by Higher Education Commission of Pakistan under '5000 – Indigenous PhD Fellowships Scheme'.

Dr. Aftab Ahmed is Director at Rawalpindi campus of Foundation University Islamabad, Pakistan. (email: aftab_ff@hotmail.com).

Dr. Kashif Zafar is associate professor at National University of Computer and Emerging Sciences, Islamabad, Pakistan. (email: kashif.zafar@nu.edu.pk)

Dr. Abdul Basit Siddiqui is Assistant Professor at Foundation University, Islamabad.

*Corresponding Author: Dr. Umair Abdullah (member IAENG, no: 114326) is Assistant Professor at Foundation University, Islamabad. Pakistan. (Phone: +92515151432; fax: +92515151433; e-mail: umair_pitafi @yahoo.com)

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. Why a Data Warehouse

Development and maintenance of a data warehouse needs a lot of resources and consumes money, therefore it is important to justify the need of a data warehouse.

Claims' data from multiple sources comes to the main operational database of the company. Operational database saves final version of every claim. If a claim gets rejected due to some error or data entry mistake, it is corrected and the corrected version is submitted to insurance again. Erroneous data of claims is updated to remove errors. Therefore, operational database cannot store that particular data of claim for which any claim get rejected from insurance. Production rule mining algorithm integrated in the rule based expert system needs both rejected and corrected versions of a claim. Therefore, in order to develop the production rule mining module, it was necessary to keep both rejected and corrected versions of a claim. In order to avoid any extra burden on operational database, a separate data repository was developed to store rejected and accepted versions of claims. We can term the data repository as data warehouse of the company.

Second reason for implementing the data warehouse is the efficiency of the Rule Based Engine (RBE) of the system (shown in Figure 1). Initially the RBE was working directly on operational database, which was causing slowness in other operations of the company and also its own speed was slow. Later the inference engine also shifted to the data warehouse. The de-normalized format of data have increased the efficiency of the RBE by allowing the processing of batch of claims, which was not possible with normalized data format of operational database.

III. Data Warehouse Design

In all companies and organizations, which are related to medical billing or medical claim processing, entities of medical billing process are almost same i.e. patient, provider, payer etc. Therefore data warehouse design presented in this paper can be utilized in any application related to medical claim processing.

Data is in normalized form in operational database and it is required to be in de-normalized form in data warehouse, for efficiency and ease of data fetching. Another major difference between operational database and data warehouse is use of primary key. In operational database primary key is logical, IDs are generated with some logic (like Practice code appended with serial number to form a patient ID). While in data warehouse primary key is physical, i.e. an auto generated serial number which increments when new record is inserted, even if the same record is already present. Physical primary key is used in all dimension tables and fact tables of the data warehouse.

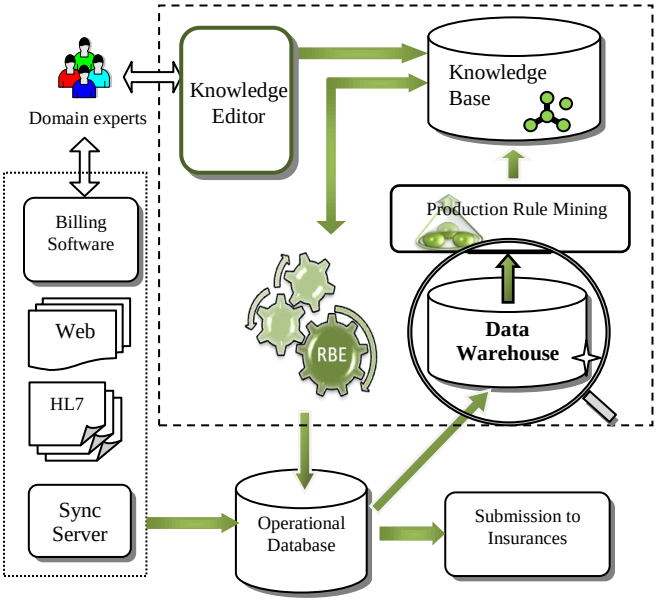

Figure 1: Architecture of the system, with focus on data warehouse

Data warehouse of the system is shown in Figure 2. It has fact tables in the middle, surrounded by dimension tables, thus forming an approximate star schema. Since payments are made by the insurance payers on the basis of current procedure terminology (CPT) therefore data warehouse of system is CPT centric. Fact table 'FctClaimProcedure' is the main table containing detailed information. A fact table 'FctClaimNotes' is included in data warehouse design to store notes about claims. Every claim has multiple note entries for storing comments by the domain users during different stages of claim processing. Therefore a lot of data, in natural language form, is associated with claims. This data is stored in 'FctClaimtNotes' fact table of data warehouse. Currently we are not mining this textual data but it can be utilized in future extension of this research work.

After CPT, next main and important entity is 'claim'. It is represented by 'DmClaim' dimension table. This table contains claim level information. Although this table stores claim level facts but it is dimension table of procedure level information stored in 'FctClaimProcedure', therefore it is defined as dimension table. All other tables related to claims and procedures are in de-normalized form, forming the dimension tables shown in Figure 2. Hence design of data warehouse is approximately star schema.

Entity relationship diagram of operational database is shown in Figure 3. Data is in normalized form i.e. focused to store one piece of data at one place.

In operational database insurance payers' information is in normalized form, it is split into three tables namely 'insurances', 'InsurancePayers' and 'InsuranceCategory'.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

In data warehouse all three insurance related tables have been de-normalized and saved as 'DmInsurance'. Further this dimension table is linked at procedure level i.e. 'FctClaimProcedure' instead of claim level, unlike claim-insurance table of production database. Similarly, in operational database procedure level data is split into two parts, first part contains procedures (i.e. CPTs with their related diagnosis (i.e. ICDs) and other part contains data related to payments, rejections or adjustments of the applied procedures (i.e. CPTs). Adjustment is that part of billed amount which is not reimbursed to the provider by the payer while rejection is that part of billed amount which is paid by the payer after correction of the claim data.

'Procedure' is being used as the granularity of medical billing data for charges and payments. Claim which is the main entity of medical billing data is represented as a dimension.

Claim dimension table along with 'fact charges' table is being used for mining claim level rules.

For example two CPTs used in a single claim, like Correct Coding Initiative Mutually Exclusive Edits etc. Measures which are visualized claim wise include 'daily rejected claims' and 'claim payment ratio'.

Data warehouse is approximately star schema or partially snow flack schema. It is a relational data warehouse with 16 dimension tables and 3 fact tables.

Design of the operational database shown in Figure 3, is aiming the operational efficiency of billing process by keeping consistent, integrated data. In operation environment 'insert' and 'update' are also main operations besides the 'select' operation, whereas in from data warehouse prospective main focus is on 'select' operation.

As shown in Figure 3, entities of the operational database have a lot of 'one-to-many' and 'many-to-many' relationships. For a 'select' command these relationships require many joins in order to fetch the desired information. 'Select' operations needs more processing time on operational database as compared to the proposed data warehouse. However, this efficiency of selection operation is achieved at the expense of saving redundant information in the data warehouse. Whereas operational database do not has any redundant data.

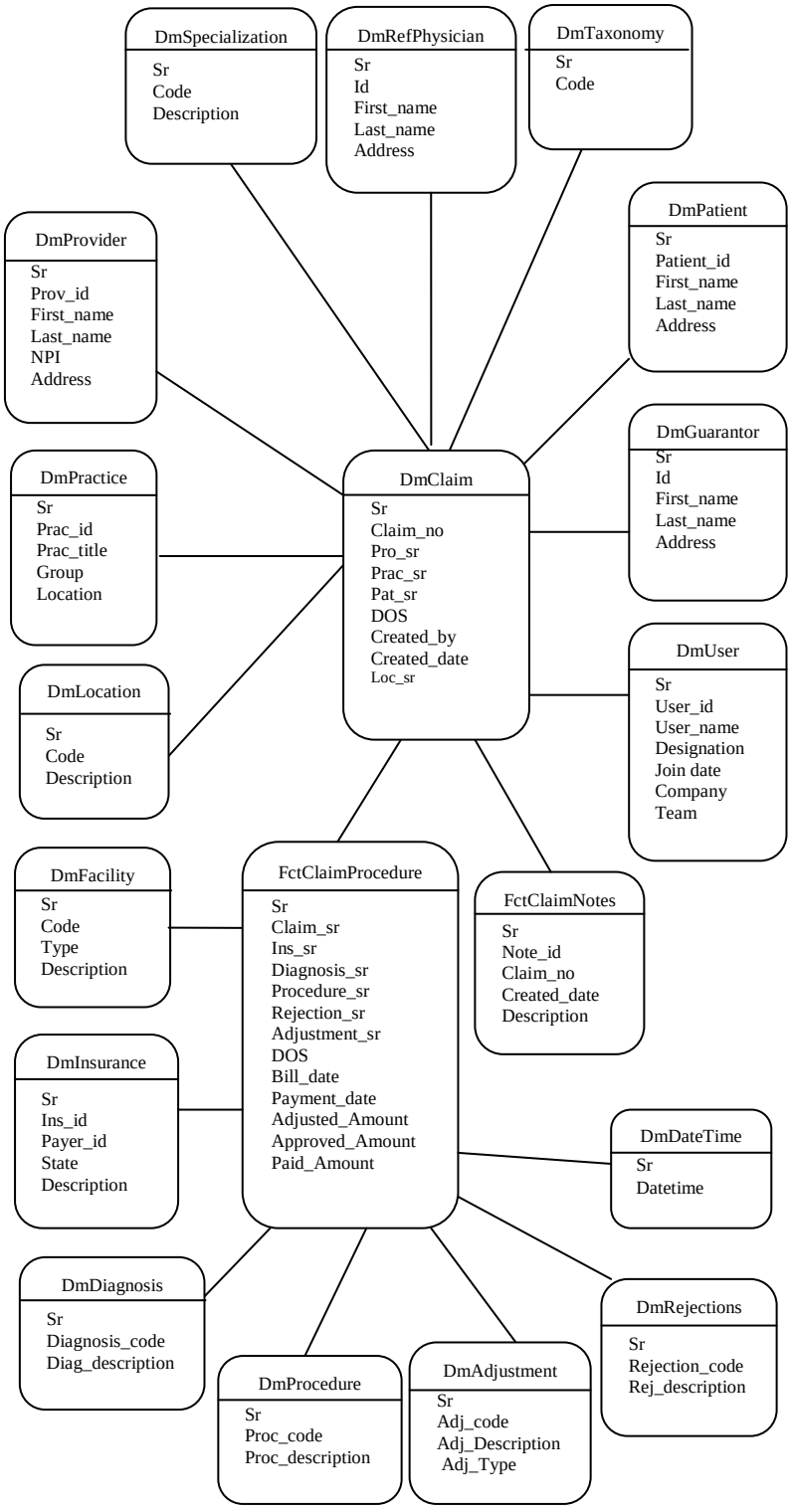

Figure 2: Data warehouse design (approximate star schema)

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

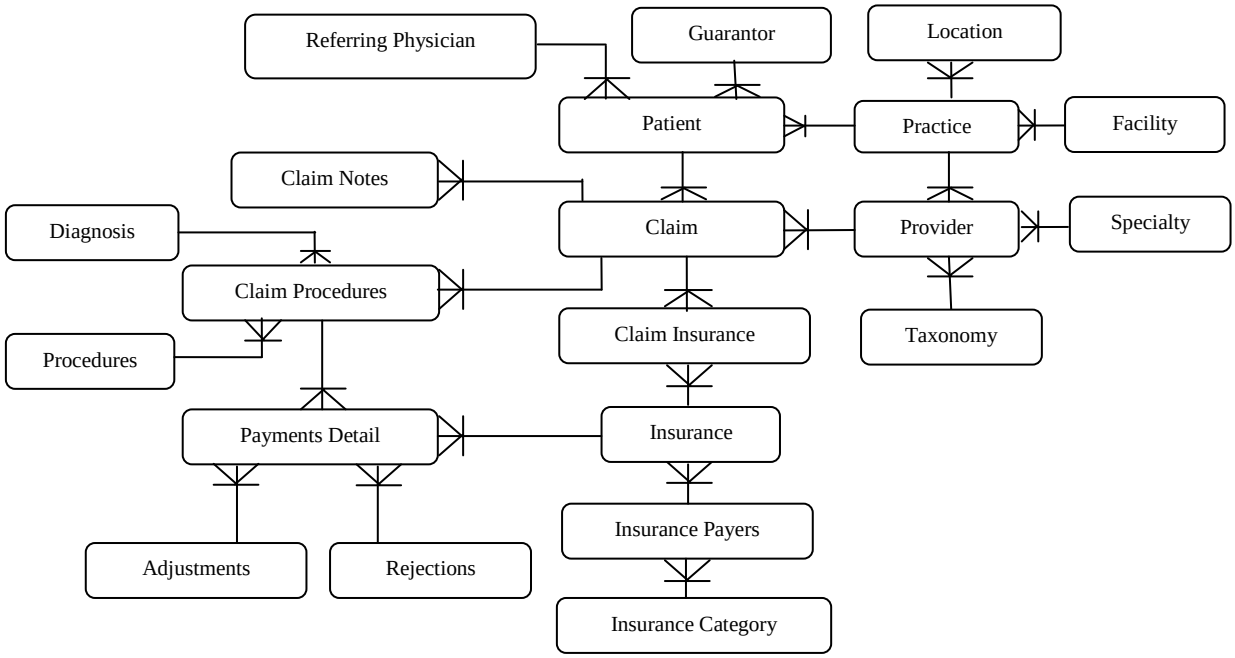

Figure 3: Operational database design

IV. Utilization of The Data Warehouse

Data warehouse presented above has been developed as one of the major components of the rule based expert systems. Over all architecture of the system, shown in Figure 1, is taken from Ahmed's work [2]. Data warehouse is being used for generating reports for analysis purposes. However, the two main utilizations of the data warehouse are as follow;

A. Rule Based Engine

Initial rule based engine using Structured Query Language (SQL) was developed and published in 2009 by Abdullah [4], [5]. An enhanced design was proposed in 2010 by Sawar [6]. SQL query is used to implement condition part of a production rule. If condition is true i.e. rule query give some results (one or more records) then action part of the rule is executed, which is a storedprocedure call, suffixed after the condition of 'where' clause of rule query. In conventional rule based systems, logical variables get their values by matching with working memory elements. For this engine whole database is serving as working memory. Logical variables get their values by executing a SQL queries. Each logical variable has a SQL query associated with it, which returns one or more values for that logical variable after execution.

All these execution of queries are carried on operational database till 2012. However, after approximately two years of operation and RBE started taking time and started influencing speed of operational database. Different measures have been taken in order to increase the efficiency of rule based system, like shifting major portion of knowledge base on submission server, and optimization of rule based engine. Besides other changes in design of the rule based engine the major change, which increased the performance, was to design it with respect to the data warehouse. New enhanced rule based engine work on the de-normalized data and processing time has reduced from hours to minutes.

B. Production Rule Mining Module

Most of the research in the area of rule mining focus on the mining of association rules [7], [8], [9]. There are only few examples of mining of production rules [10]. Quantity and quality of knowledge is important for the success of the system. Although knowledge editor has been provided to domain experts for easy feeding of knowledge to the system [11] therefore a production rule mining algorithm has been implemented to enhance and speed up the knowledge acquisition process. Domain experts now just need to validate the mined knowledge.

Production rule mining algorithm embedded in the mining module identifies what corrective action has been performed on a rejected claim in order to correct it. In other words mining module analyzes the difference between faulty version and corrected version of same claim. A 'record couple' is collection of two records, last one in which claim got paid. And second last one in which claim got rejected. Difference within a record couple is actually the corrective action performed on a claim (to eliminate the rejection). If same action is performed in all record couples of a rejection, then we have 100 % confidence for that action against the rejection. Same attribute values in all record couples of a rejection with same action identify 'when' that action needs to be performed.

With the support of production rule mining algorithm, knowledge engineers have been able to insert more and accurate knowledge, which increased the effectiveness of Rule Based System for scrubbing medical claims before submission to insurance payers. Extracted rules are in the form of SQL queries directly useable by rule based engine.

The production rule mining algorithm works on a data warehouse. It does not have any direct influence on the performance of operational database. Some of the rules mined by the algorithm and validated by domain experts are shown in Table 1 given below;

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
A SAMPLE OF FIVE RULES MINED BY THE PRODUCTION RULE MINING ALGORITHM

SR	Rule Condition		Rule Action
1	<Procedure code> = 54160 AND <gender> = 'Male'	➔	set gender = 'Female'
2	<Procedure code> = S0612 AND <gender> = 'Female'	➔	set <gender> = 'Male'
3	<Accident Date> > <DOS>	➔	set <Accident Date> = <DOS>
4	Length (<Policy Number>) != 8 AND <Payer ID> = 200204	➔	Block the claim
5	Format (<Policy number>) != '\N\N\N\N\N\N\N\N\n\n\n\x\x' AND <Payer ID> = 200114	➔	Block the claim

Where <Procedure code> is one of the procedure codes applied to the patient and mentioned in the claim (being processed by RBS).

<gender> = Gender of patient to who claim belongs to.

<Accident Date> = Accident date of the patient

<Payer ID> = payer id of the insurance being billed for the claim.

<Policy number> = policy number of the patient for the insurance being billed.

<DOS> = Date of service of the claim.

V. RESULTS AND DISCUSSION

The production rule mining module works on the data warehouse. Therefore it does not have any direct influence on the performance of operational database. Initially, RBE processes medical claims present in operational database, therefore due to considerable increase in size of the knowledge base, RBE started taking time and started influencing speed of operational database. Different measures were taken in order to increase the efficiency of the rule based system, like shifting major portion of knowledge base on submission server, and optimization of rule based engine.

Efficiency of the medical billing process increased after the RBE on data warehouse. Claim rejection rate was 5.60% prior to the implementation of the system i.e. 5.6 medical claims got rejected out of 100 claims submitted to insurance. Claim rejection rate reduced to 3.04% after the system is operational, thus showing 45.61% reduction in claim rejection rate.

Many healthcare organizations face audit and penalties for making medical mistakes. There are hundreds of cases judgments and settlements since 1986 False Claims Act against fraud, have totaled over $13 billion [12]. On June 06, 2011, twelve California hospitals face fines totaling $650,000 after medical mistakes that may have contributed to three patient deaths [13]. Data mining driven rule based system can eliminate avoidable medication errors, by applying data consistency checks at runtime. An RBS can easily apply a check to block those claims which have given CPTs for patients which are not new (have already received services during past three years).

VI. CONCLUSION

Most data warehouses are for reporting and research purposes. Data warehouse developed during this research work is for mining of useful production rules. Data warehouse is also used by the rule based engine for scrubbing of medical claims. An approximate star schema is used for designing the data warehouse. Most data mining algorithms do mining of either classification rules or association rules, whereas mining module embedded in the system is concerned with mining of production rules. Main entity used in the mining algorithm is 'record couple'. A 'record couple' is collection of two states of same claim, second last (faulty) one when insurance rejected the claim and last (corrected) one when insurance paid the claim. Production rules mined from production rule mining modules are used after validation from domain experts; by a rule based system already in operation at a USA based medical billing company. Design of the operational database is for operational efficiency of billing process. In operation environment 'insert' and 'update' are also main operations besides the 'select' operation, whereas in from data warehouse prospective main focus is on 'select' operation. 'Select' operations needs more processing time on operational database as compared to the proposed data warehouse. However, this efficiency of selection operation is achieved at the expense of saving redundant information in the data warehouse.

ACKNOWLEDGMENT

Many thanks to the USA based healthcare IT Company, MaxRemindHealth (www.maxremindhealth.com) for providing excellent research environment.

REFERENCES

[1] C. S. Ledbetter, M. W. Morgan, "Toward best practice: leveraging the electronic patient record as a clinical data warehouse" Journal Of Healthcare Information Management Summer 2001, Pages 119-131

[2] A. Ahmed, U. Abdullah, M. J. Sawar, "Software Architecture of a Rule Based Learning Apprentice System" The 2010 International Conference of Computational Intelligence and Intelligent Systems, at World Congress on Engineering, London, U.K., 30 June - 2 July 2010. Pp52-56

[3] J. S. Einbinder, K. Scully, R. D. Pates, J. R. Schubart, R. E. Reynolds, T. A. Spraggins, R. M. Krumholz "Web-Accessible Patient Data Warehouse at the University of Virginia", AMIA Symposium 1999, Page 1216

[4] U. Abdullah, M. J. Sawar, A. Ahmed, "Design of a rule based system using Structured Query Language" in Proceedings of 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC09), Chengdu, China. 2009. pp. 223 – 228. Digital Object Identifier : 10.1109/DASC.2009.78

[5] U. Abdullah, M. J. Sawar, A. Ahmed, "Comparative Study of Medical Claim Scrubber And A Rule Based System" in Proceedings of IEEE 2009 International Conference on Information Engineering and Computer Science (ICIECS 2009), Wuhan, China. 2009. 1 – 4. Digital Object Identifier : 10.1109/ICIECS.2009.5363668

[6] M. J. Sawar, U. Abdullah, A. Ahmed, “Enhanced Design of a Rule Based System using Structured Query Language”, The 2010 International Conference of Computational Intelligence and Intelligent Systems, World Congress on Engineering, London, U.K., 30 June - 2 July 2010. pp 67-71

[7] N. Lavra'c, “Data Mining and Decision Support: A note on the issues of their integration and their relation to Expert Systems” PKDD'01 workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning: Positions, Developments and Future Directions, p 1-8. 2001.

[8] U. Abdullah, J. Ahmad, and A. Ahmed, “Analysis of effectiveness of apriori algorithm in medical billing data mining” in Proceedings of 4th IEEE International Conference on Emerging Technologies, Rawalpindi, Pakistan. 2008. Pp 327-331, DOI: 10.1109/ICET.2008.4777523.

[9] R. Agrawal, and R. Srikant, “Fast algorithms for mining association rules” Proceedings of 20th Intl. Conference of Very Large Data Bases, VLDB, pages 487–499. Morgan Kaufmann, 1994. pp. 12–15

[10] E. N. Hanson, J. Widom, “An overview of production rules in database systems.” Knowledge Engineering Review. Vol 8, issue 2, 121-143. 1993.

[11] U. Abdullah, A. Ahmed, M. J. Sawar, “Knowledge Representation and knowledge editor of a medical claim processing system Journal of Basic and Applied Scientific Reseach, 2(2), 2012.

[12] K. M. Cheung, California Hospitals Fined $650K for Public Health Violations. 06 June 2011, [Accessed: 16 June, 2012], from www.fiercehealthcare.com: http://www.fiercehealthcare.com/story/california-hospitals-fined-650k-public-health-violations/2011-06-06

[13] K. Cheung-Larivee, "California hospitals fined $650K for public health violations", http://www.fiercehealthcare.com/story/california-hospitals-fined-650k-public-health-violations/2011-06-06 [accessed: April 15, 2013]

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)