A NewFramework for Analyzing and Mining Medical Data with aProposed Data Warehouse Model

Kadhim B. S. AlJanabi
Department of Computer Science
Faculty of Computer Science and Mathematics
University of Kufa
Kadhim.aljanabi@uokufa.edu.iq

Abstract

The work in this paper presents a proposed solution for preprocessing, analyzing, mining and data warehouse model for personal medical data collected from different hospitals and clinics. The proposed solution contains different phases and steps, including Extraction, Transforming and Loading (ETL) and data preprocessing focuses on converting the logged data into categories suitable for analysis and mining process, a star warehouse model was implemented that fulfills the required processing techniques, data are represented by multi-dimensional cubes for efficient and better data representation, and finally link analysis was applied on the data.

The proposed framework is simple and straight forward for implementation. Personal medical data from different sources mostly in Excel files were converted into clean, complete and consistent data by different preprocessing techniques. Logged data were converted into high quality, reliable and suitable for analysis and mining process. Star warehouse schema was implemented since it is very suitable for such type of data and mining techniques. 19900 patients records were collected and used in this work. Excel and WEKA software were used for the analysis and mining processes.

Keywords: Data Warehouse, Data Cube, Preprocessing, Star Model, Link Analysis

الملخص:

تقدم ورقة البحث هذه حلا مقترحا لمعالجة، تحليل وتنقيب البيانات إضافة إلى نموذج مقترح لبناء مستودع للبيانات الطبية الشخصية والتي تم تجميعها من مصادر متعددة كالمستشفيات والعيادات وغيرها. يشتمل الحل المقترح على مراحل وخطوات متعددة تتمثل باستخلاص، تحويل، تحميل ومعالجة البيانات والتي تشتمل بدورها على تحويل البيانات الخام إلى فئات تتلائم ومتطلبات عملية التحليل والتنقيب. كما وتم بناء مقترح مستودع للبيانات من النموذج ألنجمي يلبي متطلبات التقنيات المستخدمة في معالجة البيانات. اضافة الى اقتراح نموذج متعدد الأبعاد لغرض تمثيل البيانات والذي يسهم في رفع كفاءة الخوارزميات المستخدمة والمطبقة على البيانات، كما وتم اقتراح نموذج التحليل الترابطي لإيجاد العلاقات بين مختلف صفات البيانات الطبية الشخصية والذي يعتبر من تقنيات تنقيب البيانات الكفوءة كونها تكتشف ترابط الصفات مع بعضها البعض.

تتسم هيكلية العمل المقترحة بسهولة البناء والتمثيل والكفاءة. حيث تم تجميع البيانات الطبية الشخصية من مصادر مختلفة كانت اغلبها بصيغة ملفات اكسل والتي تمت معالجتها باستخدام تقنيات معالجة مختلفة تتلائم ومتطلبات التحليل. حيث تم الأخذ بالاعتبار جودة البيانات والمعولية وملائمة البيانات لطرق التحليل والتنقيب المختلفة. وقد تم اقتراح نموذج مستودع البيانات النجمي لتمثيل البيانات كونه الأكثر ملائمة لهكذا نوع من البيانات.وقد تم تجميع واستخدام ١٩٩٠٠ سجل من سجلات المرضى من مستشفيات مختلفة وتم استخدام نظام اكسل ونظام WEKA لمعالجة وتحليل البيانات.

1. Introduction

A data warehouse**(DW)** is a repository of information collected from multiple sources, stored under a unifiedschema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.

To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity in marketing field and disease, diagnosis, sex, age and other attributes in health field). The data are stored to provide information from a historical perspective[1,2].

Data Mining(DM) refers to the process or method that extracts or "mines" interesting knowledge or patterns from large amounts of data. Knowledge Discovery (where DM is part) consists of an iterative sequence of the following steps: Cleaning, Integration,Selection,Transformation, Mining, Pattern Evaluation and Knowledge Presentation [1,3,4,5,6,7,8,9,10].

Mining process can be classified into Classification, Association, Clustering, Prediction and Link Analysis [1]. Mining process and KDD have different applications in many fields including marketing, banking, crime analysis, employment, medicine, health insurance and many others[3,4,6,8].

Krzysztof J. CiosamdG. William Moore specified that medical data mining has special characteristics to be considered when analyzing such data[9].

Jules J Berman in the paper entitled "Confidentiality issues for medical data miners" specifies some important issues to be considered when working with medical data [10].

Joseph L. Breault, Colin R. Goodall and Peter J. Fos in their paper "Data mining a diabetic data warehouse" suggested suitable warehouse model for diabetic data [11].

J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond suggested suitable warehouse model for different clinical data[12].

2. Problem Statement

Building suitable data warehouse as a repository for medical data using a design model such as star model and applying the proper mining techniques to extract the required knowledge from these data represent a big challenge for the researchers. The work in this paper helps in presenting a new framework for solving the obstaclesfacing researchers in collecting, storing, presenting and mining medical data.

3. Proposed Model

The proposed framework architecture consists of four phases as shown in figure1.

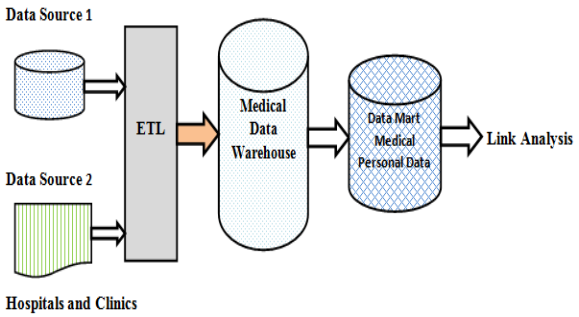


Figure (1). The Proposed Framework Architecture.

It was important to mention that most of the data were implemented using Excel software which is not very well fitted for such systems, and hence huge amount of work was required to capture and process the data. The proposed architecture is divided into four basic phases including [2,3,13,14,15,16]:

a. Data collection, Extraction, Transforming and Loading into the warehouse (ETL).
b. Data Preprocessing (categories, summarization, etc.) both ETL stage and data warehouse.
c. Data Cube and data mart Implementation.
d. Data Link Analysis.

3.1. Data Collection

The total amount of the collected data is (19900) records that represent patient's personal information that were collected from different hospitals and private clinics. The personal data represent patient's sex, age, education, marital status and disease type. Table (III) shows sample of 20 records from the data.

3.2. Data Preprocessing

The following preprocessing techniques were applied on the raw data:

1. Removing some patients records having some attribute's missing values [1,2].
2. Removing of some of the records with noisy data and those having inconsistency.
3. Converting numerical age into five different categories as shown in table I.

Table I. Patient's Age Categories

Category	1	2	3	4	5
Age	Age ≤ 18	18 <Age ≤ 35	35 < Age ≤ 55	55 < Age ≤ 70	Age >70

4. Reducing the different types of diseases into 10 categories as shown in table II.

Table II. Patient's Disease Categories

Category	1	2	3	4	5	6	7	8	9	10
Disease	Infectious	Tumors & blood	Endocrinology and Nutrition	nervous system	Eye and ear	Circulatory and digestive	Skin	Pregnancy, childbirth and urinary	Deformities and disabilities	Injuries, poisonings and other

Table III Sample of the Preprocessed Data.

	Sex	Education	Marital Status	Age	Disease
1	2	1	2	2	3
2	2	1	1	1	10
3	2	1	2	2	10
4	2	1	2	3	6
5	2	1	2	2	6
6	2	1	2	2	8
7	2	1	2	3	6
8	2	1	2	3	6
9	2	4	1	1	8
10	2	1	2	2	10
11	2	1	2	3	8
12	2	1	2	4	10
13	2	1	1	1	8
14	2	1	2	2	2
15	2	1	1	1	8
16	1	4	1	1	1
17	2	1	2	3	3
18	2	1	2	2	6
19	1	4	1	1	10
20	2	1	2	5	8

3.3. Proposed Warehouse Model

In case of medical data set with personal data, the most suitable warehouse model is the star model since the data in the different dimensions are almost normalized [1,13,14]. The star model is shown in figure 2.

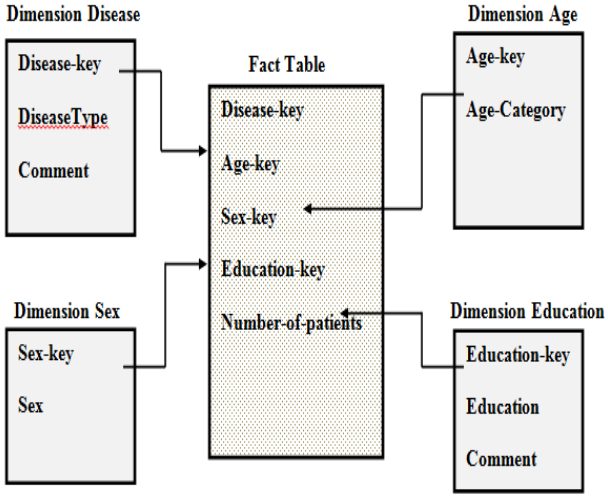


Figure (2). Star Schema for Medical Data Warehouse.

4. Data Analysis

WEKA and Excel tools were used to find out different distribution patterns of the logged data according to the personal medical data attributes including disease type, patient's age sex, education and marital status as shown in tables IV and V and in figure 3.

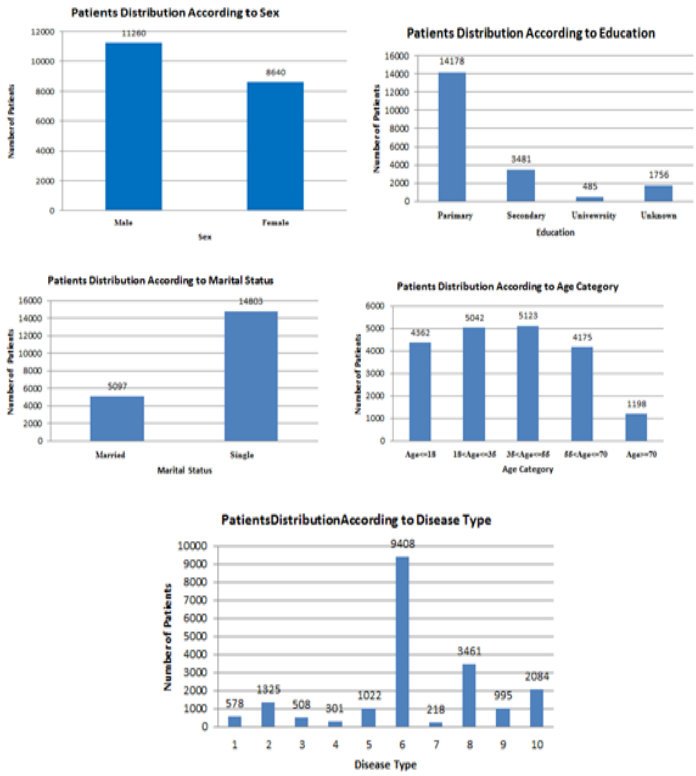


Figure (3). Medical Logged Data Distribution for Different Features.

Table IV. Medical Logged Data Distribution.

Category	Sex	Education	Status	Age	Disease
1	11260	14178	5097	4362	578
2	8640	3481	14803	5042	1325
3		485		5123	508
4		1756		4175	301
5				1198	1022
6					9408
7					218
8					3461
9					995
10					2084
Total	19900	19900	19900	19900	19900

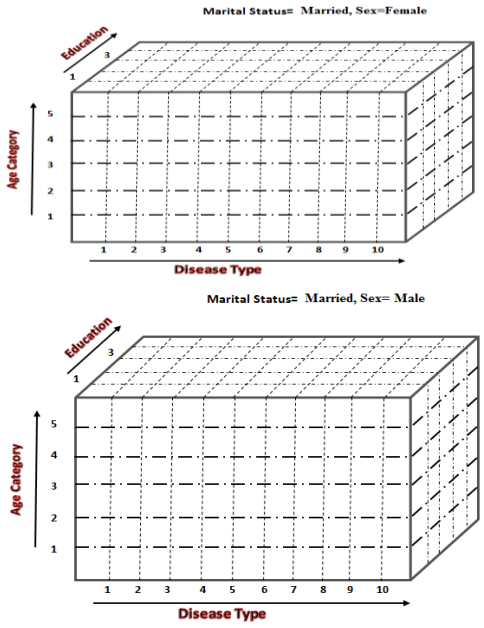
Table V. Patient's Logged Data Distributions.

Features		Disease									
		1	2	3	4	5	6	7	8	9	10
Sex	Male	177	627	234	154	515	5327	83	2253	531	1362
	Female	401	698	274	147	507	4081	135	1209	463	722
Education	Primary	431	1084	416	225	724	6920	161	2359	537	1320
	Secondary	91	178	71	41	128	1674	34	746	141	377
	University	12	23	15	6	6	295	5	69	7	47
	Unknown	44	40	6	29	165	518	18	287	310	340
Marital Status	Married	150	201	40	85	289	1978	2	801	592	900
	Single	428	1124	468	216	733	7429	217	2660	403	1184
Age Category	Age<=18	135	163	26	73	265	1664	49	670	538	778
	18<Age<=35	204	341	92	98	107	2266	56	925	242	711
	35<Age<=55	153	390	198	66	185	2555	59	975	124	419
	55<Age<=70	72	339	155	48	374	2248	46	681	70	141
	Age>70	14	92	37	16	91	676	7	210	21	35

5. Medical Data Cube

Since relational data model represents the data in two dimensional tables where rows represent data records and columns represent the required fields or features, this presentation is not very suitable for huge amount of data directed for analysis rather

than daily transactions. For this reason, an optimum data cube design that covers all the logged data with five different attributes is suggested as shown in figure 4.

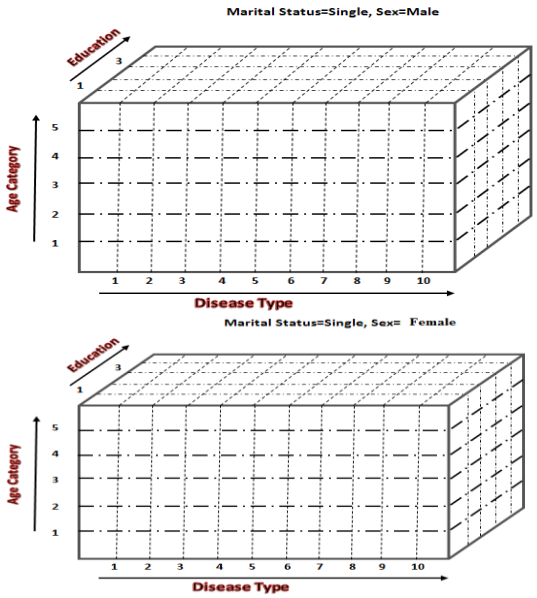


This model enables different calculations and analysis to be applied on the logged data [1,7]. The size and dimensions of the data cubes are calculated depending on the number of distinct values in each attribute as shown in table VI.

Since sex and marital status have the lowest number of distinct values, then disease, education and age represent the cube dimensions and the other attributes represent the number of cubes as shown in figure 4. Since the attributes sex and marital status has two distinct values for each then the number of cubes is 2x2=4 cubes.

Table VI. Patient's Attributes with Number of Distinct Values.

Attribute	Number of distinct values
Disease	10
Sex	2
Education	4
Marital Status	2
Age	5



Figure(4). Multi-Dimensional Data Cubes for Patient's Medical data with Five attributes.

6. Link Analysis

Link analysis represents one of the most powerful technique used in analyzing and mining medical data since it gives the correlation and relationships between the targeted attribute (disease) and other personal data attributes (age, sex, education, marital status, job, address and other attributes). And hence, this technique is used in the proposed work to get an idea between the different logged data attributes. In order to apply different link analysis techniques on the personal medical data, it is important to convert the data given in table V into normalized data using min-max normalization technique given in equation (1) [1].

$$Dinew = \frac{Di - Dmin}{Dmax - Dmin} \qquad \ldots (1)$$

Where **D**inew is the new value of Di, **D**min and **D**max are the minimum and the maximum values in the data set. The results of normalization process are shown in table VI, where the normalized data are between 0 and 10.

Figure 5 shows link analysis diagram for disease, age sex and education.

Table VII. Min-Max Normalized Data

Features		Disease									
		1	2	3	4	5	6	7	8	9	10
Sex	Male	0.18	1.04	0.29	0.14	0.82	10.00	0.00	4.14	0.85	2.44
	Female	0.67	1.43	0.35	0.03	0.94	10.00	0.00	2.72	0.83	1.49
Education	Primary	0.40	1.37	0.38	0.09	0.83	10.00	0.00	3.25	0.56	1.71
	Secondary	0.35	0.88	0.23	0.04	0.57	10.00	0.00	4.34	0.65	2.09
	University	0.24	0.62	0.34	0.03	0.03	10.00	0.00	2.21	0.07	1.45
	Unknown	0.74	0.66	0.00	0.45	3.11	10.00	0.23	5.49	5.94	6.52
Marital Status	Married	0.75	1.01	0.19	0.42	1.45	10.00	0.00	4.04	2.99	4.54
	Single	0.29	1.26	0.35	0.00	0.72	10.00	0.00	3.39	0.26	1.34
Age Category	Age<=18	0.67	0.84	0.00	0.29	1.46	10.00	0.14	3.93	3.13	4.59
	18<Age<=35	0.67	1.29	0.16	0.19	0.23	10.00	0.00	3.93	0.84	2.96
	35<Age<=55	0.38	1.33	0.56	0.03	0.50	10.00	0.00	3.67	0.26	1.44
	55<Age<=70	0.12	1.33	0.50	0.01	1.49	10.00	0.00	2.88	0.11	0.43
	Age>70	0.10	1.27	0.45	0.13	1.26	10.00	0.00	3.03	0.21	0.42

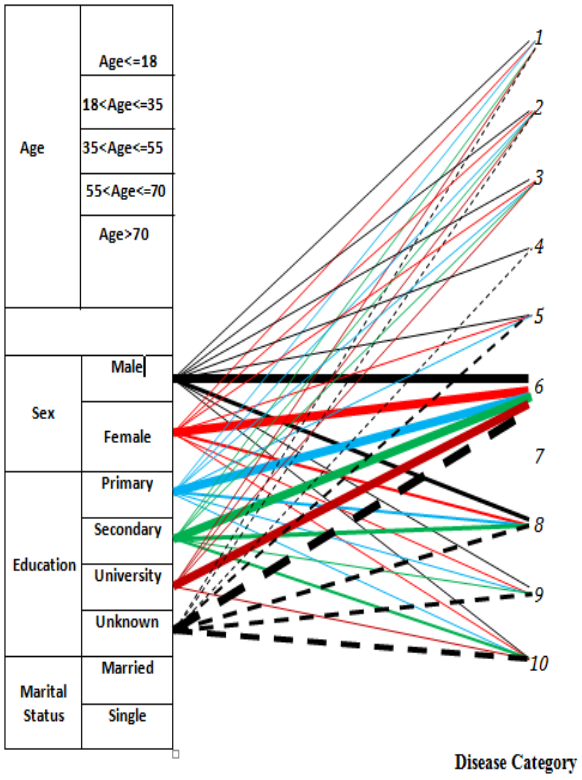


Figure 5.Link Analysis Between Disease and the Attributes Sex and Education.

7. Conclusion

The proposed framework represents a suitable environment to collect, preprocess, analyze and mining personal patient's medical data with a very suitable warehouse model which can be expressed as easy to design and implement. Data cubes are also presented as an efficient way to represent such data since these cubes can easily and efficiently be used in both roll up and drill down calculations.

Comparing the real data with the preprocessed data, it is important to mention that efficient mining algorithms are well estimated since both data and attribute reductions and category hierarchies are used in the proposed work. The complexity of most mining techniques is highly related to three attributes; number of records N, number of attributes M and number of distinct values in each attribute. It is clear from tables IV through VII and the data cube of figure 4 these three attributes were highly reduced which in turns improve the time complexity of the applied algorithms (e.g. age categories).

Personal patient's data were collected from different sources mostly in excel format which were converted into WEKA data format. 19900 patient records were collected. Results in tables I through V show the excellent performance of the proposed framework since the results give good understanding of the link between different attributes including sex, age, education and marital status and the disease type.

Data in the given tables are very suitable for link analysis which leads to the results shown in figure 5. The results show high link between these attributes and disease 6 (Circulatory and digestive).The results obtained showed that the proposed framework is straight forward, simple to be implemented and fulfill all the analysis and mining requirements. And can easily be used in other mining techniques such as classification, clustering and association.

8. References

[1]Jiawei Han and MichelineKamber ,"Data Mining: Concepts and Techniques",3rd Edition , Morgan Kaufmann, 2012.

[2] M. H. Dunham, "Data Mining: Introductory and Advanced Topics",
Prentice Hall, 2002.

[3] M. Durairaj, V. Ranjani, "Data Mining Applications In Healthcare Sector: A Study", NTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGYRESEARCH VOLUME 2, ISSUE 10, OCTOBER 2013 ISSN 2277-8616 29

[4] DivyaTomar and Sonaligarwal," A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266, ISSN: 2233-7849

[5]Tianyi Wu, Yuguo Chen and JiaweiHan, "Association Mining in LargeDatabases: A Re-Examination of ItsMeasures", in Proc. 2007 Int. Conf. onPrinciples and Practice of KnowledgeDiscovery in Databases (PKDD'07),Warsaw, Poland, Sept. 2007.

[6] Dr. Bushra M. Hussan ," Data Mining based Prediction of Medical data Using K-means algorithm" , Basrah Journal of Science (A), Vol.30(1),46-56 2012.

[7] Krzysztof J. Cios ,WitoldPedrycz, Roman W. Swiniarski and Lukasz A.
Kurgan," Data Mining :A Knowledge DiscoveryApproach",Congress,2007,
ISBN-13: 978-0-387-33333-5.

[8] Kadhim B. SwadiAljanabi ," An Improved Algorithm for Data Preprocessing in Mining Crime Data Set", Journal of Kufa for Mathematics and Computer , Vol. 1 , No .4 , Nov ., 2011, pp.81- 87.

[9] Krzysztof J. Cios, G. William Moore, "Uniqueness of medical data mining", Journal of Artificial Intelligence in Medicine, volume 26, Issues 1-2, 2002 Pages 1–24.

[10]Jules J Berman, "Confidentiality issues for medical data miners", Journal of Artificial Intelligence in Medicine, Volume 26, Issues 1-2, 2002 Pages 25–36.

[11]Joseph L. Breault, Colin R. Goodall, Peter J. Fos, "Data mining a diabetic data warehouse", Journal of Artificial Intelligence in Medicine, Volume 26, Issues 1-2, 2002 Pages 37–54.

[12]J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond, "Medical data mining: knowledge discovery in a clinical data warehouse", Proc AMIA Annul Fall Symposium. 1997 : Pages 101–105.

[13]S. Tsumoto, "Problems with mining medical data", Computer Software and Applications conference, 2000. COMPSAC 2000, IEEEXplore, The 24th Annual International conference, 2000 ,Pages 467 - 468 .

[14] A. Kusiak, K.H. Kernstine, J.A. Kern, K.A. McLaughlin, and T.L. Tseng, "Data Mining: Medical and Engineering Case Studies", Proceedings of the Industrial Engineering Research 2000 Conference, Cleveland, Ohio, May 21-23, 2000,
pages 1-7.

[15] Susan Jensen, "Mining Medical Data for Predictive and Sequential patterns: PKDD 2001", Proceedings of the 5th European Conference 2001.

[16]M. Ilayaraja, "Mining medical data to identify frequent diseases using Apriori

algorithm", International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), 21-22 Feb. 2013, Pages 194 – 199.