Development of Specific Disease Data Warehouse for Developing Content from General Guide for Hypertension Screening, Referral and Follow Up

DR. TEH YING WAH, NG HOOI PENG, CHING SUE HOK
Department of Information Science, Faculty of Computer Science and Information Technology,
University Malaya
Lembah Pantai, 50603, Kuala Lumpur.
MALAYSIA
Email: tehyw@um.edu.my

Abstract: - This paper proposes a method of developing specific disease data warehouse such as hypertension data warehouse. Significant steps in developing the data warehouse will be described especially data extraction, transformation and loading. The purpose of developing this data warehouse is to help/assist the specialist, health care team, pharmacists to figure out the best and suitable strategies to be implemented during the screening, referral and follow up process. As the data may come from various data sources mostly different websites, the amount of time spent of this tasks is often underestimated. Issues on how we crawl the data from various data sources and store it into database will be discussed further.

Key-Words: Data warehouse, Hypertension, Screening, Referral, Follow up

1 Introduction

The main goal of this paper is to explore the steps in developing the specific disease data warehouse. Throughout this paper, we will focus on one specific disease which is Hypertension or can be refer as High Blood Pressure. Due to the awareness of public on health care issue; prevention, detection, evaluation and treatment of hypertension has become important. Therefore, strategies that are suitable and significant over the improvement of hypertension control are needed. In this paper, we will focus on the Hypertension Screening, Referral and Follow Up process. The process of building Hypertension data warehouse will be presented in this paper. The process is included extraction of the data, which are the steps or strategies involved in Screening, Referral and Follow up process. Transformation and filter the data and finally the loading of the data will be discussed.

2 Literature Review

Hypertension is a common public health problem. However, the prevention and treatment of hypertension is an important due to the serious consequences caused by hypertension. Science news highlighted that hypertension is uncontrolled and taking over the world and the risk of becoming hypertensive for a person in a developed country exceeds a “staggering” 90%. This disease is also expected to be dramatically increased in the coming years. It also stated that the biggest problem for controlling hypertension is compliance with treatment [1]. Without proper treatment, it will cause damage throughout your body. Hypertension can cause damage to your arteries and these damages blood vessels will lead to the disease in your tissues and organs. These affected tissues and organs are brain, heart, kidneys and eyes [2]. As mentioned that hypertension cause damage to the whole body’s tissues and organs, these damages can cause people to have heart attack, heart failure, stroke, kidney failure and renal failure [3]. It has been estimated that at least 50 million people in USA are affected by hypertension disease that rarely presents any symptom. Therefore, most of the people are unaware that that they have been affected by such dangerous disease that threatening their life. Hence, hypertension is also called as a “silent killer” [4].

Preventing these potentially life threatening diseases is significant for ensuring a healthy body only with your hypertension checked, controlled hypertension and proper treatments [3]. An effective strategy for controlling the hypertension is definitely needed. A

nationwide survey had been done and its reveal that 88% throughout 1,800 organizations had carried out the hypertension screening project. However, much of the people complained that this is a short term programs that may not achieve the target of lowering the blood pressure as the program does not provide an effective referral, follow up and the evaluation of program effectiveness [5]. Obviously, screening itself is inadequate in achieving the goal of decreasing the blood pressure. An appropriate referral and follow up are needed as well for an effective hypertension management. Therefore, improvement in the management of hypertension is required.

Practice guidelines on the management of hypertension are important as well. It provides strategies or a general guideline for prevention and treatment of hypertension. British Hypertension Society is concerned about the increasing rates of the uncontrolled blood pressure in United Kingdom and therefore they have recommended their guidelines for hypertension management. These guidelines are prepared for used and followed by general practitioners, practice nurses and also generalists in hospital practice. For blood pressure measurement, the guidelines will include a detailed guidance while measuring the blood pressure. These guidelines are also to ensure the accuracy of the measurements. As the blood pressure levels are different, it also had been classified according to its level [6]. The various types of hypertension will be treated by using different strategies.

To improve the hypertension treatment and control, pharmacists had been suggested to take responsibility for medication management and patient outcomes. However, an evaluation had been done on the practices of the pharmacists and results shown that most of the pharmacists did not perform up to the expected standard for hypertension management. A major change in pharmacy practice is suggested in order to improve the hypertension care and control [7]. These studies had revealed that the practices or guideline are so much important in improving the hypertension treatment and control. Moreover, future work is suggested to focus on identifying the gaps in practices as to reveal the insufficient of the current practices in improving the hypertension management [7]. Perhaps, analysis on the practices or guidelines with data mining might help to overcome the shortcomings. Since a well-designed generic guidelines are not suitable to apply on all individual patients, analysis should be done on all the practices guidelines to find the possible associative among the guidelines [8].

In Canada, clinicians are exposed to various hypertension management guidelines. Through studies, the potential barrier to the implementation of hypertension management guideline had been identified and they had summarized the shortcomings of the current hypertension guidelines as Table 1 [9]. These problems need to be overcome in order to improve the guidelines' impact in the future. More attentions are required from the guideline developers on the issues and shortcomings that need to be tackled. Performing mining technique on these various guidelines may reveal the significant patterns or relevancy among the guidelines.

Source of barrier	Examples of barriers
Guideline	- Discordance between guidelines produced by different organizations - Failure to address clinically relevant issues - Format that are not user friendly - Lack of local involvement - Lack of implementation strategy - Failure to incorporate patient-clinician values - Poor methodological quality
Clinician	- Lack of awareness - Lack of familiarity - Lack of agreement – with guidelines in general or with specific guidelines - Lack of motivation - Lack of self-efficacy - Lack of outcome expectancy
Environment/practice setting	- Lack of time - Lack of resources - Lack of incentives to change - Lack of opinion readers
Patient	- Patient preferences contrary to guideline - Questionable applicability of recommendation to the individual patient

Table 1 Barriers to the successful of Hypertension Managament Guidelines Implementation
Source: McAlister, F.A. et al. (2001)

Information can be found of the general guide for hypertension screening, referral and follow up through the search in web. This information is required in building data warehouse which will ease the specialist, health care team and pharmacists in analyzing the strategies as to deal with different categories of hypertension.

Building a data warehouse involve the extracting of operational data and entering it into the data warehouse [10]. However, extracting data from various data sources is a complex problem. In this case, data that we need will be from the web page we search through Google search engine. Web pages are programmed in different language or format such as

plain text, HTML, Pdf and etc. Therefore, web crawler nowadays has to suite the needs on extracting web pages in the different format. Other than extracting part, many inconsistency issues need to deal with while integrating the data.

Extracting the web page content and mine the useful information from web page content is considered as web content mining. Web content mining is related with data mining and text mining as data mining techniques can be implemented in web content mining and also web contents are texts. However, web content data are mainly semi-structured and unstructured, while the data mining are mainly deal with the structured data and also the text mining focus on the unstructured data [11]. We are currently interest on performing extraction on each web page to find those meaningful and relevant strategies and then apply the mining techniques on the extracted relevant data.

3 Data Warehouse Architecture

Recently, data warehousing technology are getting popular and implemented in medical field to discover the trends of medical and patient data. Due to the rapid growth in information technology, medical area also produces increasingly voluminous amounts of electronic data [12]. To deal with these large volumes of data, data warehousing technology is applicable. The data volume of a data warehouse is often claimed to be hundreds of gigabytes to terabytes in size [13].

One of the existing system that implement data warehousing technology is Health-Mining. Health-Mining provides disease management support service based on data mining and rule extraction. Once the collections of guidelines and processes are gathered, statistical groups and the result variance can be calculated. The results will help in identifying and extract the rules about the patients' care processes and also a new approach to guidelines [14]. Fig. 1 shows parts of the software architecture of the Health-Mining. The patients' data are collected using form management, tree input management and also HL7/ DICOM protocol to collect the electronic patient record data. This data are collected into a database and ETL processes are performed to move data from multiple sources, cleanse it and load it into the data warehouse for further analysis [14]. As compare to ours, we are collecting the data from various sources that are available in internet. ETL processes are performed on each web page. We are more to focus on the steps in collecting the data from large amounts of data sources.

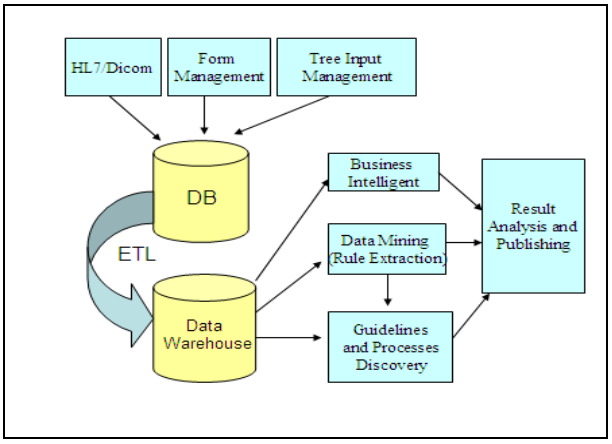

Fig. 1 Software architecture of Health-Mining
Source: Bei, A. et al. (2005)

4 Development of the Hypertension Data Warehouse

The main function of an ETL tool is to take data from many formats, transform and load it into database. However, the trend of an ETL tools are evolving as to fulfill the requirements. So, the vendors of ETL keep on adding the capabilities and functions of their ETL tools. Among the changes is to have larger volumes as to store more data. One of the reasons for increase data volume is due to the users who want to cull data from a wider variety of system. According to the report of evaluating ETL and data integration platforms, although most of the companies use ETL to extract data from relational databases, flat files, and legacy systems, a significant percentage shown that they want to extract data from application packages, such as SAP R/3 (39 %), XML files (15 %), Web-based data sources (15 %), and also EAI software (12%) [15]. Information is available and easily accessible through internet nowadays. Due to the advanced of internet technology and the growth of internet users, the percentage of extracting data from web-based data source will increase. Therefore, the ETL tools need to support the function of extracting the web-based data source. This paper will continue by focus on extracting the web-based data source.

4.1 Data source identification

The data that we grab will be the information on web page which are relevant to general guideline for hypertension screening, referral and follow up. To narrow down the scope, the data sources that we identified are search through the Google search engine. The current keywords that we used in searching the information are “management of hypertension”. From Fig. 2, the result from Google search is over three hundred thousand. We are dealing with a significant number of complex data sources which the available ETL tools may not support. Hence, we choose to develop ETL program.

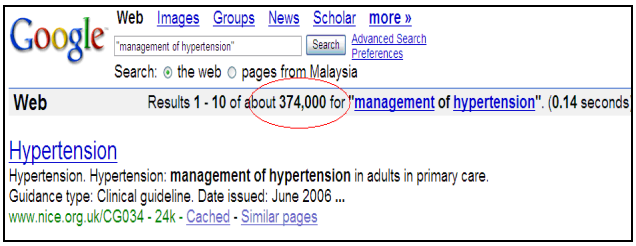

Fig. 2 Result of "management of hypertension"

4.2 Extraction

Two main extractions will be involved, hyperlink extraction; as for the data we needed is from various data source, we will crawl each hyperlink from search result of Google. The hyperlink will be stored as a list of links in database as a root data source. While the web data extraction part, with the concept of storing only the relevant data in database, each web page will be filtered to ensure high quality of extracted data.

4.3 Transformation

While extracting the web pages, we perform data mining technique as to find the relevant keywords; especially Screening, Referral and Follow up in these web page. Text mining is performed to search and extract useful information. The importance of web is rapidly increased as it provides richness information. Those informative web pages can be found through the powerful search engine we have nowadays. However, the searching result may not be what we expected and it does not make sense if the information or web pages do not contain information that you need or interest on it. The existing search engines rank the web pages mainly based on the keywords matching [16]. It may have the keywords, but there is nothing relevant content that you need. We also cannot determine the relevancy of the search result based on the ranking. Therefore, mining concept plays an important role and thus is implemented during transformation. After the web pages extraction, we need to filter and eliminate those unrelated web pages. In this paper, we concerned about the strategies related to screening, referral and follow up for hypertension. We need an effective technique in order to perform the task of mining and extract the relevant and useful information. Many information extraction techniques are available which include keyword based search, wrapper information extraction, web queries, user preferences and resources discovery [16].

Task that we need to perform includes keywords based search and wrapper information extraction. However, the keywords based search that mentioned earlier is not enough as the ranking is not based on the content information. Hence, we start the mining task on the result of keywords based search. The tasks we performed are summarised as below.

Given
- Web pages content
- Keywords: screening, referral, follow up

Find
- Sentences with relevant keywords
- Extract the relevant paragraph or sentences and ignore the non- relevant paragraph and sentences.
- Group the related sentences or paragraph according to the keywords

Fig. 3 shows the result after have performed the text mining. The relevant data will be grouped according screening, referral and follow up.

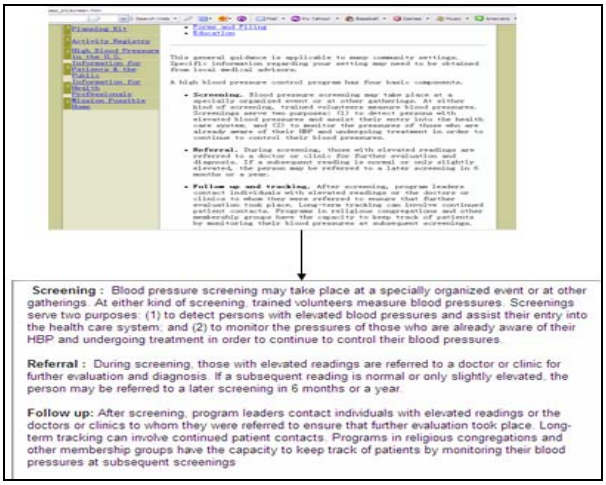

Fig. 3 Result of text mining

4.4 Loading

In data loading, we design the table in first normal form and partition it as to have better performance during SQL selection and updating. The relevant data will be inserted into table in paragraph forms. The workflow of the development is shown in Fig. 4.

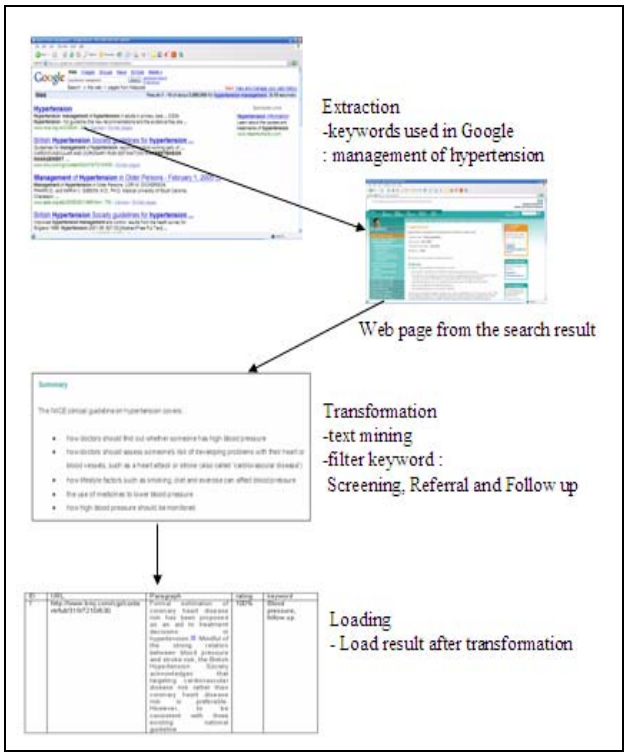

Fig. 4 Workflow of developing hypertension data warehouse

4.5 Architecture

In developing this kind of data warehouse, we should have these two main basic tables:

a. Table for root data source: Store all unique links that gathered from extraction process using power of search engine.
b. Table for loading result: Store all data from each unique root data source which is going through transformation: filtered and validated with text mining process.

Huge amount of root data source must take in consideration which application must design in multithreading and can run in multiple instances. One thread will perform the whole workflow from a root data source and lock the root data source using an indicator. This is to prevent the next thread from processing a root data source that has been processed.
During loading process, data duplication will be checked. To minimize overhead in processing of work flow, path or URL crawled will store in a separated partitioned table. This table is to keep track the URL that has been processed as shown in Fig. 5.

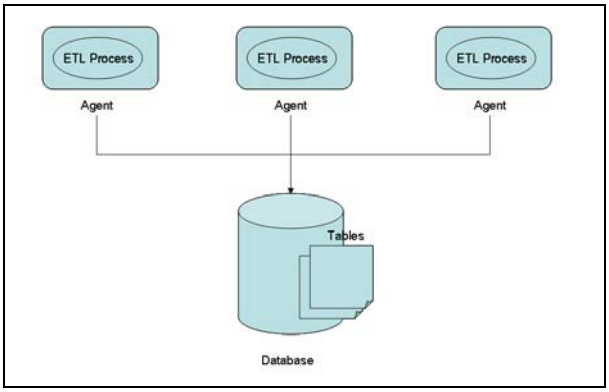

Fig. 5 Architecture

5 Results

The main result will be a hypertension disease data warehouse which consist all information of guidelines for hypertension screening, referral and follow up. Part of the result is shown in Table 2.

6 Conclusions

The goal of this piece is to propose a method of building a disease data warehouse and sharing the knowledge and problems that we had facing during the development. This data warehouse will be useful especially in health care environment. The data can be extracted using various kind of data mining technique and used in areas such as decision support, prediction, forecasting, and estimation [17]. In health care environment, doctors need this up-to-date information for diagnosis decision making.

ID	URL	Cache_ID	Short_Des	Relevant	Keyword	Date_Update
103	http://www.heartfoundation.org.au/document/NHF/hypertension_management_guide_2004.pdf	7t0wTyl1m9wJ	A guide to assessing and managing raised blood pressure in patients. Summary points\n_Raised blood pressure, particularly systolic blood pressure, is directly related to increased risk of cardiovascular events and death._Lifestyle modifications are first-line interventions for high blood pressure management even where drug therapy is instituted. _ High blood pressure should not be managed in isolation. All cardiovascular disease risk factors need to be addressed.	89	guide|blood pressure|follow up|diagnostic|treatment|monitoring	20070513 13:04:45
204	http://hp2010.nhlbihin.net/nhbpep_kit/screen.htm	6t0tkPl1l4vT	This general guidance is applicable to many community settings. Specific information regarding your setting may need to be obtained from local medical advisors.A high blood pressure control program has four basic components.Screening. Blood pressure screening may take place at a specially organized event or at other gatherings. At either kind of screening, trained volunteers measure blood pressures.	95	screen|referral|guide| blood pressure|follow up|process|education| general	20070513 13:54:30

Table 2 Result of Hypertension Disease Data Warehouse

References:

[1] The Lancet (2007, August 22). Hypertension: Uncontrolled and taking Over the World. Retrieved December 17, 2007, from http://www.sciencedaily.com/releases/2007/08/070818102318.htm

[2] U.S. News and World Report 23 May 2006. Retrieved November 11, 2007, from http://health.usnews.com/usnews/health/heart/hypertension/hyper.about.complications.htm

[3] "Hypertension: could your life be at risk? High blood pressure is a major risk factor for heart attack, stroke, and renal failure, but one out of four people don't even know they have it". Healthy Years, 2007.

[4] High Blood Pressure – A Silent Killer. 29 May 2003, MedicineNet.com. Retrieved November 25, 2007, from http://www.medicinenet.com/script/main/art.asp?articlekey=13118

[5] Gillum, R.F., Stason, W.B., Weinstein, M.C., Screening for hypertension: a rational approach, *J Community Health.* Vol.4, No.1, 1978, pp. 67–72.

[6] Williams, B., Poulter, N.R., Brown, M.J., Davis, M., McInnes, G.T.,Potter, J.F., Sever, P.S. and Thom. S.M., British Hypertension Society guidelines for hypertension management (BHS-IV): summary, *BMJ* Vol.328, 2004, pp. 634-640.

[7] Mc Lean, D.L., Bungard, T.J., Hui, C. and Tsuyuku, R.T. Community pharmacist practices in hypertension management, *CPJ/RPC*, Vol. 139. No. 5, 2006, pp 38-44.

[8] Svatek, V., RIHA, A., Peleska, J. and Rauch, J., Analysis of Guideline Compliance – A Data Mining Approach. In: *Symposium on Computerized Guidelines and Protocols (CGP-04), IOS Press*, 2004.

[9] McAlister, F.A., Campbell, N. R. C., Zarnke, K., Levine, M. and Graham, I.D., The management of hypertension in Canada: a review of current guidelines, their shortcomings and implications for the future, *CMAJ*, Vol.164, 2001, pp. 517-522.

[10] Inmon, W.H. *Building the Data Warehouse, 3rd Edition*, New York: John Wiley & Sons, 2002.

[11] Liu, B. May 15, 2005, Web Content Mining Retrieved November 5, 2007, from http://www.cs.uic.edu/~liub/WebContentMining.html

[12] Kambayashi, Y., Winiwarter, W. and Arikawa, M., *Data Warehousing and Knowledge Discovery: Third International*, 2001.

[13] Dana, B., B-to-b Companies learn to get more out o rich stores of data, *B to B*, Vol.86, No.17, 2001, p.19.

[14] Bei, A., Luca, S.De., Ruscitti, G. and Salamon, D., Health-Mining: a Disease Management Support Service based on Data Mining and Rule Extraction, *EMBS*, 2005, pp 5466-5470.

[15] Eckerson, W. and White, C., *Evaluating ETL and Data Integration Platforms*, The Data Warehouse Institute, Report Series, 101Communications LLC, 2003.

[16] Liu, B., Chin, C.W. and Ng, H.T. 2003, Mining Topic-Specific Concepts and Definitions on the Web, *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 251-260.

[17] Dan et al., Data mining for network intrusion detection: A comparison of alternative methods, *Decision Science*, Vol.32, No.4, 2001, pp. 635-660.