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

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

Abstract: - This paper will be proposed 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 part. The purpose of developing this data warehouse is to help the specialist 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 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 issues; 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 included is 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.

2 Literature Review

Hypertension is a common public health problem. However, the prevention and treatment of hypertension is important due to the serious consequences caused by hypertension. Therefore, improvement in the management of hypertension is required. Practice guidelines on the management of hypertension are important as it provides strategies for prevention and treatment of hypertension. The various types of hypertension will be treated by using different strategies. Numerous information can be found of the general guide for hypertension screening, referral and follow up through the search in the web. This information is required in building data warehouse which will ease the specialist in analyzing the strategies as to deal with different type of hypertension.

Building a data warehouse involve the extracting of operational data and entering it into the data warehouse [1]. 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 languages or formats such as plain text, HTML, Pdf and etc. Therefore, web crawler nowadays has to suite the needs on extracting web pages in different formats. Other than extracting part, many inconsistency issues need to deal with while integrating the data.

3 Major steps in developing a data warehouse

Generally, building a data warehouse consists of five following steps [2]:

1. Extracting the transactional data from data sources into staging area.
2. Transform the transactional data
3. Load the transform data into a dimensional database.
4. Building pre-calculated summary value to speed up report generation
5. Build or purchase a front-end reporting tool.

These steps can be summarized as how to collect the data and how to use the data. In this paper, we will focus on steps of how to collect the data. These steps are further described in the following section.

3.1 Data source identification

Before a data warehouse is being developed, you need to identify the data that you wish to store into the data warehouse and also where to collect the data.

3.2 Build customize ETL tool

Each data warehouse has different requirements. Therefore, a customizable ETL tool is a better solution in order to fulfill the requirements. Many ETL tools are available nowadays as to ease the data warehouse developers in developing their data warehouse. However, survey on ETL tools had been done and yet there are still a large number of organizations building their data warehouse without using an ETL tool, but writing their own, mostly very complex SQL statement which is often difficult to maintain [3]. Their reason given was the high cost of many ETL tools and the abundance of programmers on their staff to justify their decision [4].

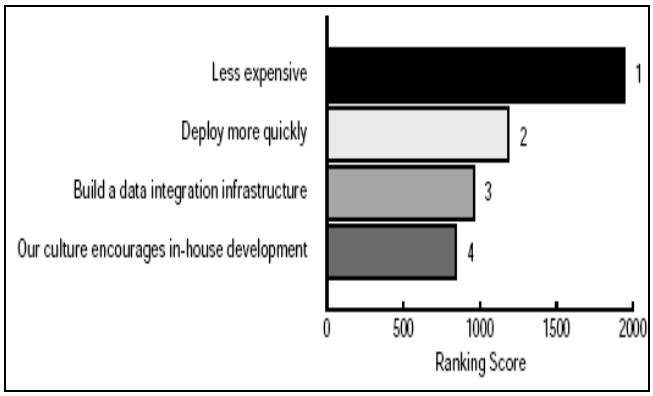

Figure 1 Reason for coding ETL program – Ranking [4]

Survey had been carried out and the result is shown in report that some organizations believe it is cheaper and quicker to code ETL programs than use a vendor ETL tool as refer to the Figure 1 above [4]. Much of them agree that custom software save them lots of money and time as they write their own code and everything is self documenting. Moreover, the cost to maintain their custom ETL code is less than the annual maintenance fees and training costs that other department is paying to use a vendor ETL tool [4]. Therefore, in this paper, we proposed an approach of developing disease data warehouse as it can be referenced by others who are interested in building custom ETL.

3.3 Extraction

This can be the most time consuming part where you need to grab the data from various data sources and store it into the staging database. Much of the time and efforts are needed in writing a custom program to transfer the data from sources into staging database. As a result, during extraction, we need to determine which database system will be used for the staging area and also figure out what are the necessary data that are needed before grab it. The declining in cost of hardware and storage has overcome the issues on avoiding the data duplication and also their worries on lack of storage as storing the excessive or unnecessary data. However, there is probably no reason to store the unnecessary data which had been identified not being useful in decision making process. Therefore, there is a necessary for extracting only the relevant data before bringing into data warehouse [5].

3.4 Transformation

After extracting the data from various data sources, transformation is needed to ensure the data consistency. In order to transform the data into data warehouse properly, you need to figure out a way of mapping the external data sources fields to the data warehouse fields. Transformation can be performed during data extraction or while loading the data into data warehouse. This integration can be a complex issue when the number of data sources getting bigger.

3.5 Loading

Once the extracting process, transform and cleansing has been done, the data are loaded into the data warehouse. The loading of data can be categorize into two types; the loading of data that is currently contained in the operational database and the loading of the updates to the data warehouse from the changes that have occurred in the operational database. As to guarantee the freshness of data, data warehouse needs to be refreshed to update its data. Many issues need to be considered especially during loading the updates to the data warehouse. While updating the data warehouse, we need to ensure that no data are loosed and also to ensure a minimum overhead over the scanning existing file process.

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 increasing data volume is due to the users who want to cull data from a wider variety of systems. According to the report of evaluating ETL and data integration platforms [4], although most of the companies use ETL to extract data from relational databases, flat files, and legacy systems, a significant percentage shown that they wan 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%) [4]. This can be referring to the Figure 2 below.

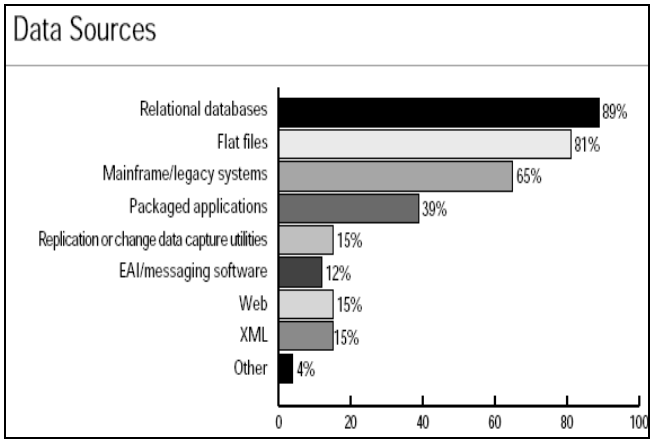

Figure 2 Types of data source that ETL program process [4]

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, ETL tools need to support the function of extracting the web-based data source. This paper will continue by focusing on extracting the web-based data source.

4.1 Data source identification

The data that we grab will be the information on web pages which are relevant to general guideline for hypertension screening, referral and follow up. To narrow down the scope, the data sources that we identified through the Google search engine. The current keywords that we used in searching the information are “management of hypertension”.

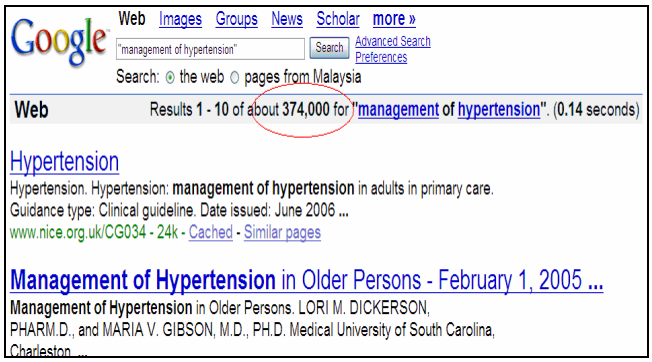

Figure 3 Result of "management of hypertension"

From Figure 3, the results from Google search are 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.

4.2 Extraction

Two main extractions will be involved, hyperlink extraction; as for the data we needed is from various data sources, we will crawl each hyperlink from search result of Google. For 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.

4.2.1 Hyperlink extraction

Each hyperlink from Google search result is crawled and stored as a list of links in database as a root data source.

4.2.2 Web data extraction

From the list of links, we identified out the relevant data sources. For relevant data sources that have deep resources, web crawler will be used to crawl the deep web resources.

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. These relevant keywords will be filtered based on the ranking. The tasks we performed are summarized as below.

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

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

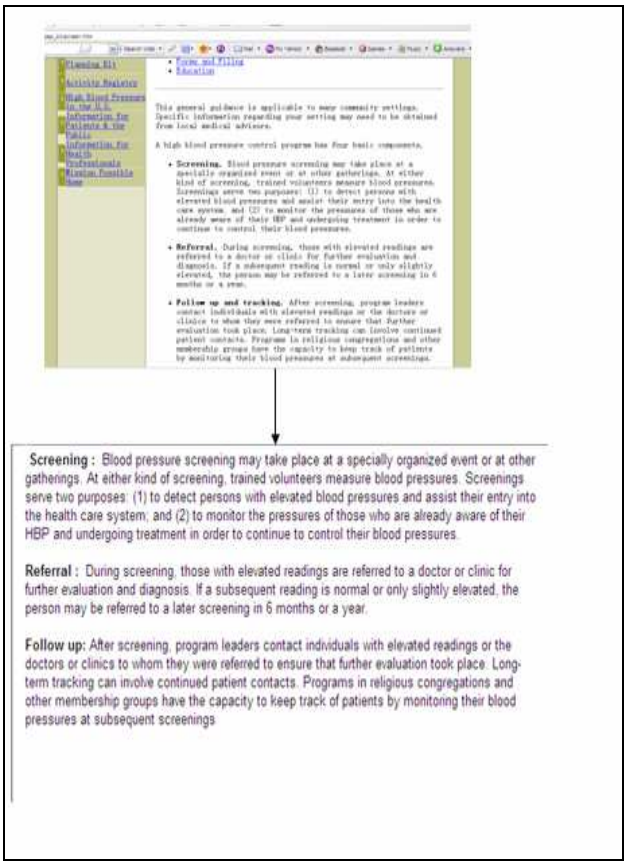

Figure 4 Result of text mining

Figure 4 shows the result after perform the text mining. The relevant data will be grouped according screening, referral and follow up.

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 Figure 5.

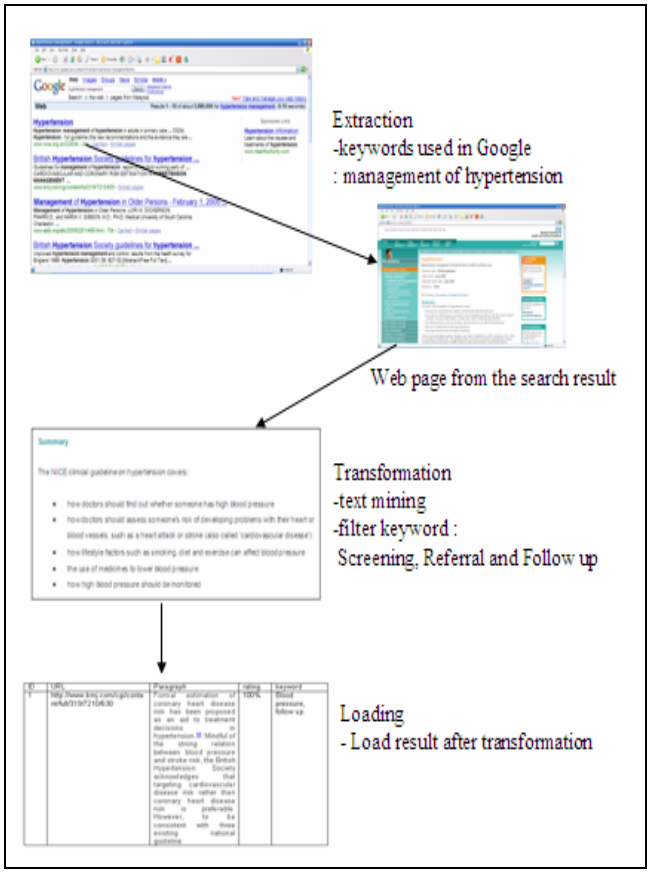

Figure 5 Workflow of developing hypertension data warehouse

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 after 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 had been processed.

During loading process, data duplication will be checked. To minimize overhead in processing of work flow, path or URL crawled will stored in separated partitioned table. This table is to keep track the URL that had been processed.

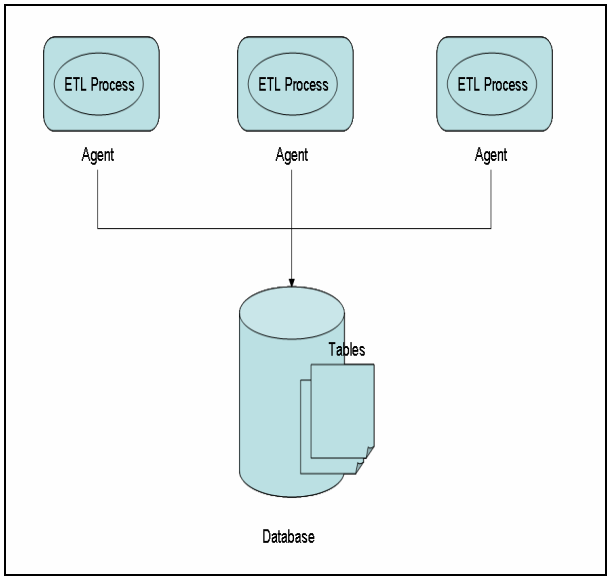

Figure 6 Architecture

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 1.

6 Conclusion

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 [6]. 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	6t0tkPl114vT	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 1 Result of Hypertension Disease Data Warehouse

References:

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

[2] B. Pavliashvili. Steps Involved in Building Data Warehouse, http://www.informit.com/articles/article.aspx?p=24902&rl=1 11 Jan 2002. (Accessed 16 Jun 2007)

[3] ETL Tools Survey 2007 – Third Edition http://www.etltool.com/index.htm (Accessed 2 October 2007)

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

[5] E.G. Mallach, *Decision Support and Data Warehouse Systems*, United States: McGraw-Hill, 2000.

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