Ratna Madhuri Maddipatla al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 2, Issue 2, June 2015, pp. 169-174

Data Mining - Healthcare Data

Ratna Madhuri Maddipatla[1], Nagamani Maddipatla[2]

[1]Graduate Student, UNC Charlotte, North Carolina,
[2]Research Scholar, JNTUH, Hyderabad, India

***ABSTRACT*: In the current scenario, where most decisions are data driven, health care is one such sector which can phenomenally effect lives, create awareness and educate people through various mining user opinions available in form of reviews on web. In this age of Internet, where one does a thorough research prior any decision taken; people no more rely completely on physicians and pharmacists. This paper is an attempt to throw light and deep dive on the insights available in mining health care data available on web and understand how Health Care Analytics could indirectly enhancelives digitally by providing intellectual solutions through data mining and data visualization along with the reviews present in the form of numbers and text for various drugs and medical conditions.**

***Keywords*: Data mining, Data Visualization, Tableau, Web scraping**

I. INTRODUCTION

Data Mining is a vast discipline where discovery of useful information in the form of hidden patterns and trends happens from large raw data sets through various methods involving Machine Learning, Statistics, Computational Capabilities, Database Systems and Data Visualization. Data Visualization proved to be one of the smart and articulate ways of looking at the story behind the numbers and text in the data. The journey of data mining majorly deals with slicing and dicing of data, understands the relationships between various factors involved and the impact a certain factor could or would have on an event which can be built as a predictive model and used for future predictions. Analysis and Visualization act as two pillars for Data Mining. While adding structure to unstructured data has always posed challenges, analyzing the semi structured data and visualizing the patterns would prove useful as well. Data Mining is a concept which can be adopted by various domains and business verticals to delve deep into the nuances and enable them to take better decisions around business strategies. The domain could range right from oil industries, retail, supply chain to banking and insurance. In recent years usage of data mining techniques for business and data analysis increased many folds and many universities are offering new specialization courses in this science. Data mining is being used extensively by various organizations and many have started to collect data in a structured format for future prediction purposes.

Data mining with the term data in it definitely speaks for its size and frequency of data collection and generation hence the term big data was coined to cater to the space and complexity issues. In addition to these characteristics, data should be collected from different sources and in multiple formats. When so many complexities are involved, data storage itself will become a big challenge. To overcome this various data warehousing techniques are being used. Analysis of data to find new patterns is also termed as data analytics which include many data mining techniques.

Healthcare is one of the domains where data size is increasing tremendously with a great speed due to the various issues faced by various people with the changing lifestyle, needs and habits. Usage of data mining applications in this area would immensely benefit all the stake holders of the healthcare. For example, same data with various probing points can be used by insurance companies to investigate fraud and hospitals to maintain better customer relationship and make decisions during healthcare schemes design for the doctors to identify best practices etc. Data generated in bulk quantity by daily transactions will be too differential and complex and cannot be analyzed by conventional procedures. Latest data mining techniques can be used to convert this complicated data into more useful information needed for making effective and timely decisions. With the advancements in internet technologies and public health services, we are able to surf different healthcare related information where various medical conditions, drugs used, patient details and their comments are available in the form of feedback. In this paper we made an attempt to scrap, collect, mine and visualize this kind of data and bring out few analyses which will be useful for interested general public. Such analyses and visualization can also be used by various healthcare websites along with the data that is available.

II. DATA COLLECTION

In recent times, there are many websites which provide information and experiences for various drugs based on the symptoms provided.It will help patients a lot to cross check the medicines they are using andattainpsychological satisfaction about the usage of drug. This sort of open information available on websites not only increases the power of networking but also creates an awareness which is required in the process of decision making. Extraction of information from data on the websites to unearth various facts about the data by different analytical methods is the most trending aspect at the moment.

In the process of unearthing, various technologies need to be used since the data present on the source website could be unstructured to various degrees. In which case, the only option available to obtain this data to local storage for analysis is development of proprietary application development to meet the needs.

Extraction of this data is termed as web scarping as data is being extracted from websites. It is also alternately called as screen

scraping or web harvesting etc. By this process we are automating the process of data extraction from websites which can be performed with in no time. Web scraping application will interact with specified website in the manner similar to any web browser. Based on the specified web accessing parameters particular website will be reached and required data will be extracted and stored to local storage system.

Most of the development platforms like Java, .NET provide APIs to support web scraping process.Also various web parsing packages are available on open source platforms such as R and Python with powerful functionalities which allow the analyst to invest more time into looking and slicing the insights rather than mere procuring the data. Although, most analysts are provided with the required data for analysis, there are exceptional cases where one is posed the challenge to achieve the goal from scratch which is data collection. The art of decision making begins right at the data collection stage where one needs to decide on various metrics and variables that would or would not be required for the analysis later on.

Approach .Net:

To collect data from the above mentioned website a desktop application was developed using .NET C# with suitable user interface. It is compatible with Visual studio 8 and higher versions. Using this applications data relating to various medical conditions was extracted from the website and stored as a CSV file for further analysis.

Main window of the application is as shown below.

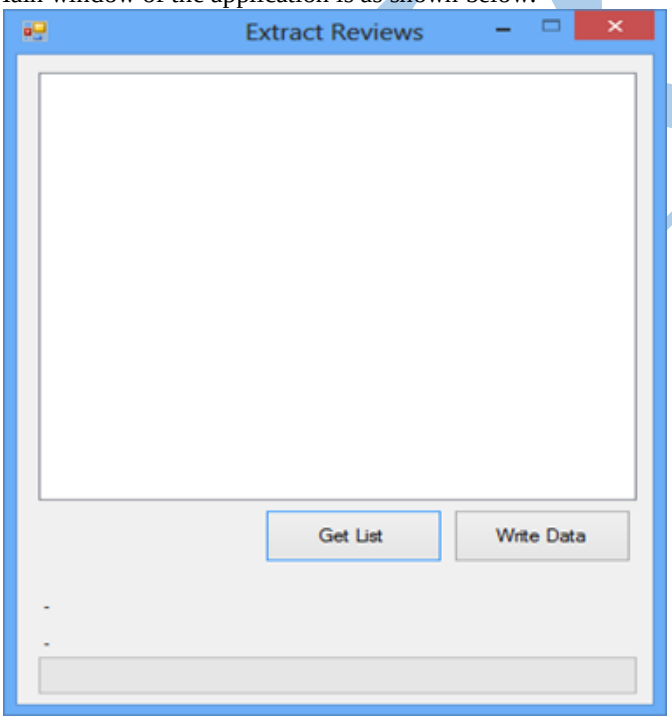

Fig 2.1Web scraping application user interface. 'Get List' is for displaying medical conditions such as Hypertension, Asthma etc. and 'Write Data' for extracting the data and storing in local storage.

On selection of 'Get List' button, all the medical conditions stored in the website are read and listed in the list box provided.

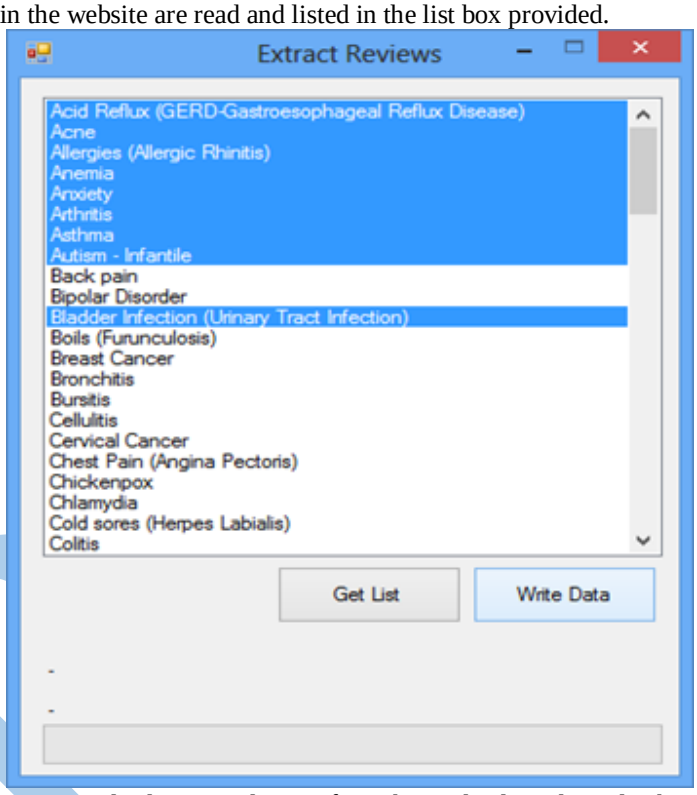

Fig 2.2Multiple rows selection from the medical condition list box for web scrapping.

Once "Write Data" button is selected, all the data pertaining to the selected medical conditions will be written stored in "c:\Temp\user_reviews.xls" file which is of CSV format. . Medical conditions for which reviews are not available will not be displayed in the list box. Single of multiple medical conditions can be selected from the list box provided (shown below). Though the application at present works only for medical conditions available through web scraping, it can be easily extended for other aspects like top drugs and all drugs based web scraping.

After converting to XLS format the data looks like below.

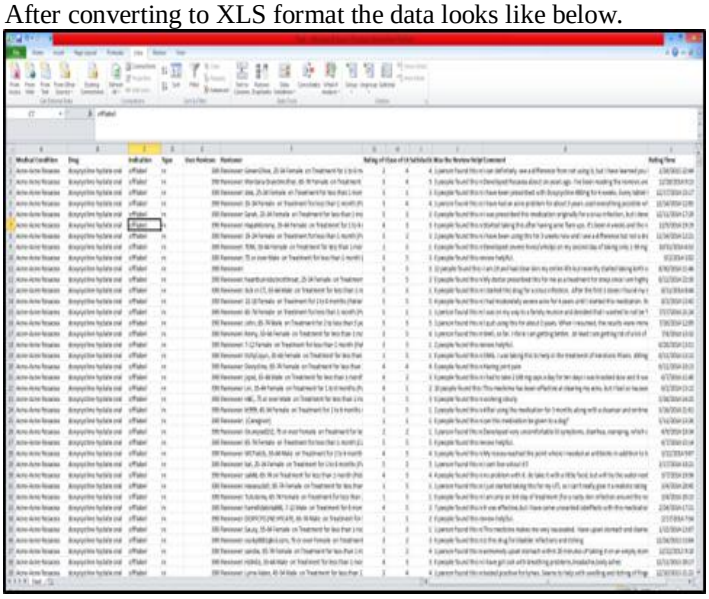

Fig 2.3Web scraped data after conversion from CSV to Excel format. It contains 9 different fields.
This CSV file will become the standard source for further analysis.

Approach R:

Web scraping is the method of pulling data that is present on various websites on the internet. In the current context, web scraping can be done using R which is an open source is using various packages such as "XML", "gdata" and the functions such as "htmlTreeParse()" and "xpathSApply()".These packages can be downloaded from the CRAN website and R Studio was used for this particular task.
This is one of the most efficient ways of scraping in terms of time consumption as R is highly efficient in terms of computation. rvest() package can also be used to achieve this objective which has functionalities similar to "beautifulsoup" in python. R is also catching up in terms of parsing packages. One could parse the data present on the webpage via various methods such as class, id, xpath etc. The method or technique that would be used would vary based on the requirement and automated accordingly for continuous data pull from immediate pages which are usually directed by the "next " button present on the top or bottom part of the webpage. This is one of the most interesting challenges that an analyst faces as the degree of programming expertise expected generally varies.

III. DATA VISUALIZATION AND FINDINGS

The visualization has been performed on Tableau which is a leading platform catering to various business solutions by allowing the users to cut through the data at multiple levels and visualize them with ease with various shapes, sizes and colors. Undoubtedly it is one of the leading visualization platforms across globe. It is a highly user friendly interface to deep dive into the hidden patterns and bridge the gap between raw data and useful information. Following are the various fields available in the collected data.

- Medical condition addressed by reviewer
- Drug used by the reviewer
- Profile of the reviewer including age and gender
- Feedback of the reviewer
- Ease of effectiveness rating
- Satisfaction rating
- No. of people who found the review helpful
- Time stamp of the review

Not all metrics listed above would be used on the analysis. Hence, it is important to decide which analysis is to be performed and how it will be helpful for other end users. Ideally one needs to cater to the objectives of a cause for any analysis to be useful unless it is research. Website authorities could perform this analysis and share the insights with people in the form an online dashboard. It would keep their customer traffic engaged as well as direct them in the right path and create awareness among them.
Below Fig 3.1 is a Box Plot visualization of drugs which were found useful by both males and females and the number of people who found it useful. One could quickly see the most popular drug which tops the chart and also the number of people in male and female categories who found it helpful.
The distribution and position of the median in the boxplot are explaining the % of male and female reviewers on the website. Although Lisinopril oral tops the charts, we can see that there are two different drugs with maximum reviewers in male and female categories.
One would always want to match their symptoms and profile with the closest experience shared to better their decision and suit them better. These websites allow them to do that and providing access to such user friendly visualizations on the webpage would enable the users think analytically and also take an informed decision.

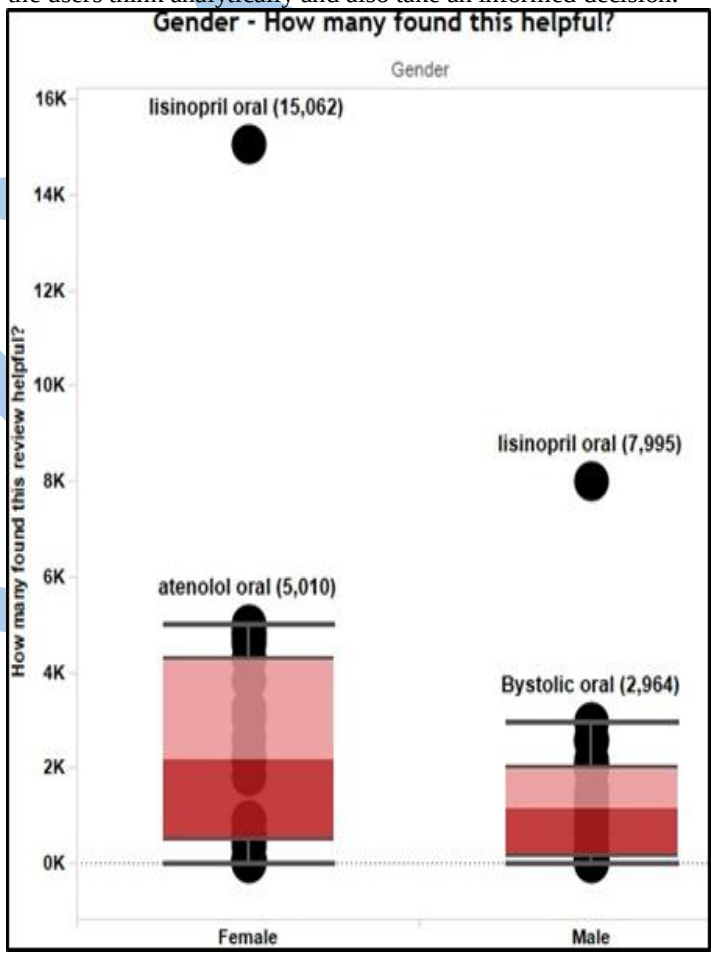

Fig 3.1 Box Plot distribution showing the number of people who found the drugs useful in the data

Below Fig 3.2 is the visualization which depicts the ease of use rating by both males and females and the rating associated with each of these drugs. The color legend indicates the corresponding color for both males and females. The width of the central bar for each drug indicates the number of people who provided their reviews which represents the traffic which clearly varies from drug to drug. This information could be used by various pharmaceutical manufacturing firms to understand the prominence of their drugs vs. competitor's drugs. The size legend named as Number of Records provides a reference to the scale of the size depicted.

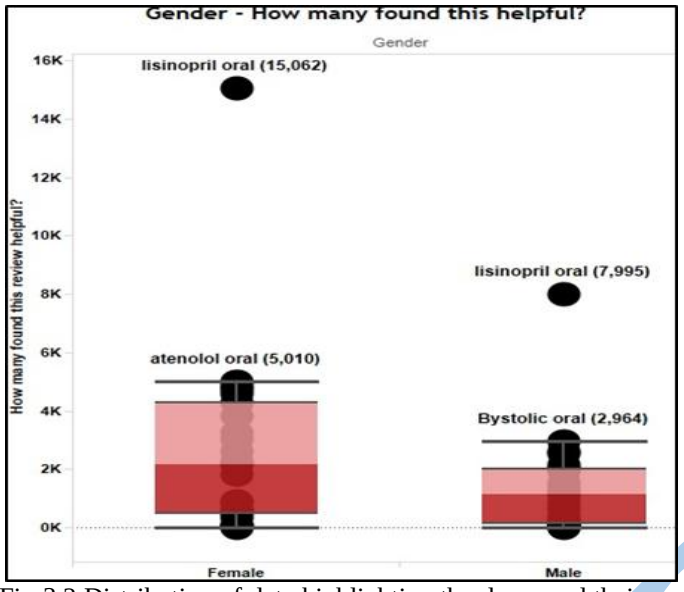

Fig 3.2 Distribution of data highlighting the drugs and their ease of rating in male and female reviewer groups

The visualization below Fig 3.3 talks about the satisfaction levels in the reviewers again at a gender level. This helps us understand the top drugs which successfully might generate decent sales in the market due to the satisfaction levels associated with them. Also, once visualized based on the variance in these variables, these factors can be considered as independent variables during statistical analysis through non parametric methods such as decision tree or random forests and parametric methods such as linear and logistic regression.Such insights could be used to understand their impact on the sales of the drug present in the market for significant amount of time.

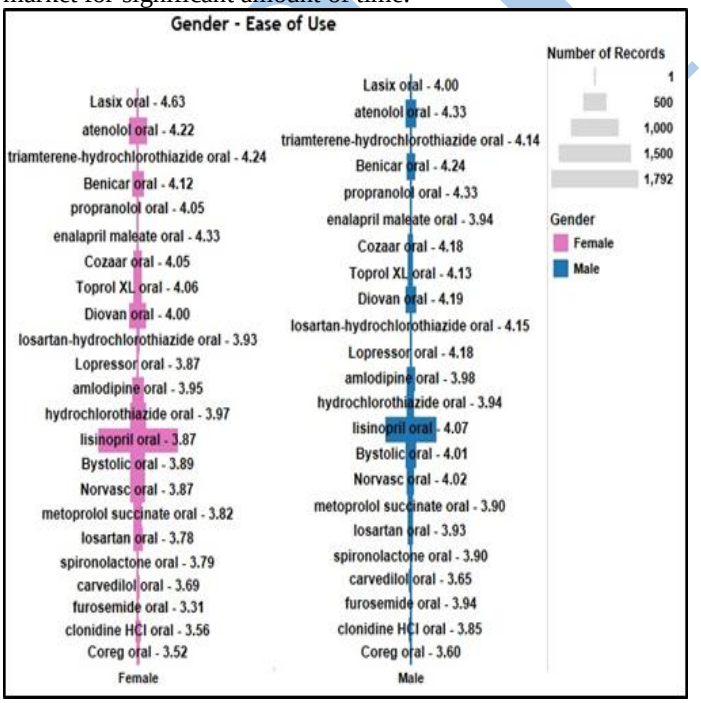

Fig 3.3Distribution of data highlighting the drugs and their satisfaction rating in male and female reviewer groups

It is indeed important to understand the times in a year during which the reviews have been posted and the gender of the reviewer which can be useful for the relevant users.

Below Fig 3.4 is a heat map visualization which helps in understanding the number of reviews at a Year, Quarter, Month and Gender level. The color legend at the top of the chart is used to associate the number of reviews in each of the blocks. Based on the patterns and trends observed during the exploratory data analysis, one could also choose to perform a time series analysis as a next crucial step in the analysis.

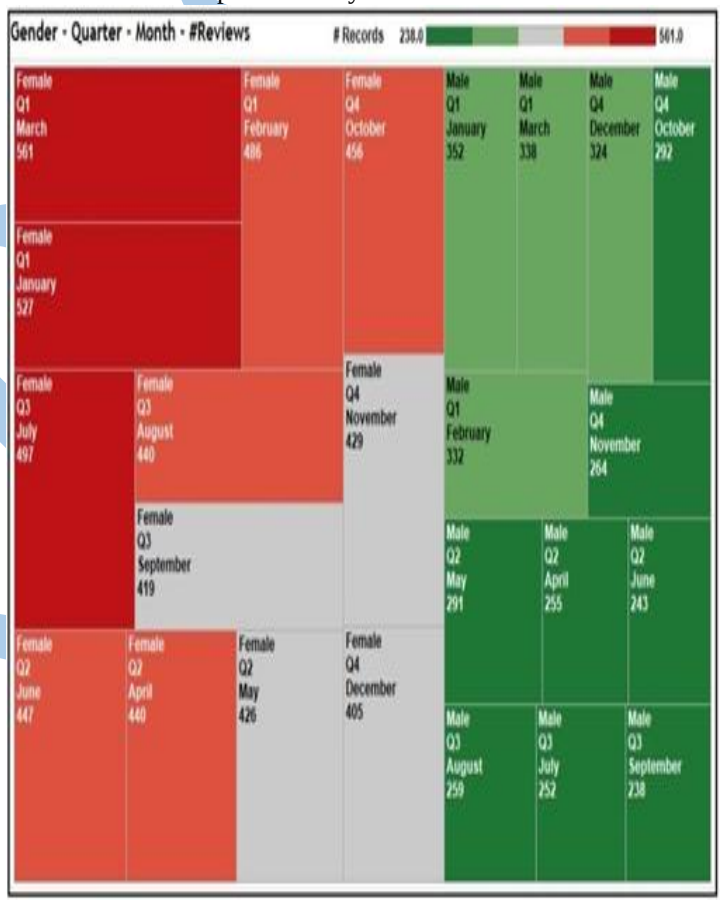

Fig 3.4 Heat Map elucidating the number of reviewers at an year, quarter, month and gender level

With the increasing data in the form of text, text mining has been on top of the data mining charts in the recent past years. Text Mining has various applications and below Fig 3.5 is the word-cloud visualization for the medical condition Hypertension which gives the percentage of people who found the relevant drug helpful based on the comment of the reviewer. This is a consequential/derived metric which speaks about the people who chose the drug based on the comment of the reviewer.

In a word-cloud, the size of the word indicates its frequency in the document and the associated percentage should add more insight in this word-cloud. Word Association is also one of the text mining techniques which can be performed as a next step where the % of association of a word with various other significant words present in the word cloud is listed. This way one could

identify the n-grams involved in the corpus of data. This percentage of association should help us build a story around the information hidden in the raw data.

It is one of the smart ways of drawing information from huge corpus of data. Text Mining is one of the pillars under semantic analysis in Cognitive Computing which is the next big word after big data.

Fig 3.5 Word cloud highlighting the drugs which were found the most useful and the percentage of people who found it useful associated with the drug

IV. APPLICATIONS

In the current scenario where there are medical conditions such as Uterine Fibroids in various females for which the cause has not been discovered and is a top research area, analysis of the user reviews in terms of the symptoms and precautions would mitigate the panic in the public. Most females' prior menopause and post child birth happen to generate these fibroids at various locationsaround the uterus which are later removed through surgeries. Hence, data analysis based on the knowledge prevailing in the patients is one way to regulate the panic and increase the awareness among people.

Besides addressing the medical conditions which have multiple drugs, it is a good initiative to spread awareness on the upcoming medical conditions.

Hence, Data Mining would increase the importance of data collection and identify what the raw data has to say and connect the dots to build a story around the text and numbers. Prediction of hospital readmission rates in United States in order to reduce the penalty has a lot of scope in data mining. The solution for various problems can be obtained by delving deep into its data itself.

V. CONCLUSION

Where there is will, there is a way. Analogically where there is data there is information. Along with various other domains, Health Care sector is no more a mere person to person interactive based industry. Health Care got digital in its own way and various components of this sector which are patients, physicians, nurses and administrators are pacing up technologically in order to benefit from each other's data. Various websites such as WebMD are a platform for healthcare reviews for various medical conditions and drugs. The various ways in which this data can be sliced and diced to bring the information out is commendable with the current and vast scope of data mining and data visualization. How different people react to a drug helps people match their profile and circumstances to take the right decision smartly. This would also encourage various others to share their experience since experiences transform into data which further transform into insights and then decision taken by various others.

The analysis in the form of web scraping and data visualization elucidated in the previous sections are not only used by various websites but also multiple pharmaceutical companies tie up with these websites to obtain the information and perform in-house analytics to make necessary strategic changes in case need be. Hence, information can be consumed by various groups in various ways which ultimately lead to serve a common goal which happens to be betterment of the society and eradication of disease and medical conditions.

VI. FUTURE WORK

Health care sector is one of the top sectors where data is being tremendously multiplied with time and needs a structured and planned storage of the data and continuous mining of insights to renew their policies and cater to the needs of public health in a much better manner. Various health care websites and organizations need to tap the technology to mine and visualize the data and bridge the gap between various groups of patients, physicians, pharmacies and hospitals. We need to look at a bigger picture to ask the right questions and solve the right problem. The art of showing the value hidden in patient and physician level data to the health care sector needs to pace up and take a better shape to build a better society.

VII. REFERENCES

[1]. http://cran.r-project.org/ - Reference for various packages used for data scraping
[2]. http://www.tableau.com/ - Application used for visualizations
[3]. http://www.webmd.com/ - Data source for the analysis
[4]. http://www.siam.org/meetings/sdm13/sun.pdf
[5]. http://www.journalofbigdata.com/content/pdf/2196-1115-1-2.pdf
[6]. https://www.himss.org/files/HIMSSorg/content/files/jhim/19-2/datamining.pdf
[7]. http://www.sciencedirect.com/science/article/pii/S1319157812000390
[8]. http://www.hissjournal.com/content/pdf/2047-2501-2-

3.pdf
[9]. [http://airccse.org/journal/ijasa/papers/2214asa01.pdf
[10]. [http://www.informaticseducation.org/Healthcare_Data_Analytics.pdf
[11]. http//www.informaticseducation.org/Healthcare_Data_Analytics.pdf
[12]. http://www.ijcsit.com/docs/Volume%205/vol5issue04/ijcsit20140504224.pdf
[13]. http://www.ijcsit.com/docs/Volume%205/vol5issue06/ijcsit2014050661.pdf
[14]. http://www.siam.org

