ABSTRACT:

Web Mining is a process that extracts interesting and useful data from an array of information produced by Web servers. The healthcare with web mining and intelligent data analytics has major applications in the design of real-time healthcare systems. The healthcare system was developed to be compact, lightweight, low power consumption, low-cost and comfortable for use at any time. The self-health care for aged people can be improved by using big data analytics and web mining in maintaining their own health conditions without institutional care. This research paper study and analyze the big data analytics to find the solutions in health care issues considering various factors affecting the human beings. Several solution schemes that can solve the health care problems in the human beings are also studied and reviewed.

Keywords: Web Mining, Healthcare, Big Data Analytics and Intelligent Data Analytics

[1] INTRODUCTION

Big data is a collection of data sets. It is huge and complex to deal with using traditional data processing methods and algorithms. The challenges include acquisition, storage, searching, sharing, transferring, analysis and visualization. Big data exceeds existing storage, computational and communication capabilities of conventional methods or systems. Big Data are characterized by large volume, high velocity with respect to rate of generation and time period available for interpretation, and disparate variety [1].

Web analytics (WA) is defined as “the collection, measurement, analysis and reporting of Internet data for the purpose of understanding and optimizing Web usage”. WA refers to a tool that collects click stream data regarding the source of website traffic (e.g., e-mail, search engines, display ads, social links), navigation paths, and the behavior of users during their website visits and that presents the data in a meaningful format [2]. With fast development of
social network technologies some treatments related to health are found by using web
Analytics. All different kinds of information related to humans can be founded on the web
that can be seen as a universal data repository or called web of Data.

Health care systems cannot perform effectively without sufficient numbers of workers
who are skilled and motivated. Researchers have offered other suggestions for scaling-up a
public-health approach based on decentralized, integrated, equitable delivery of care that uses
simplified and standardized operational approaches. This can be provided through the use of
available technology. Many small and middle-income countries face various challenges in
health care delivery that includes weak health system, poor infrastructure, communal war,
corruption, and financial resources. In addition to this health workers are also distributed
unevenly.

Usage of Internet as a communication tool for health-related information is growing
rapidly. Health information-seeking behavior varies changes depending on the type of
information required with the reasons and the experience of searching in the web. Research
shows that women search for health information more than men. Online health consumers
tend to be more educated and online support groups dynamically exchange health
information. The credibility and trust of the web sources becomes more important for the
users and several research show that health related search has increased drastically to obtain
health and medical information. Furthermore it is found that more number of physicians
present printed Internet-sourced health information at visits to patients. The traditional
doctor–patient relationship is a challenging task. Data mining techniques such as
classification, clustering, association and regression are used to solve some issues occurring
in the medical referral process [6 , 7]. Machine learning technique will lead to more effective
Information Retrieval (IR) applications in some domains such as medical healthcare,
biomedicine, web and bioinformatics [8].

[2] EXISTING TECHNIQUES

Users must use the most effective strategies for the promotion, protection and
maintenance of health. This can be achieved using the web to guide about the benefits, risks
and costs of communication. Summaries of existing knowledge compiled currently in the
web can provide a powerful resource for practitioners and consumers. Such reviews also
enable researchers to focus their attention on identified gaps in knowledge. Several popular
health care issues in big data analytics that aim to handle different kinds of issues are listed.

(a) Health Buddy in medical care:

The Health Buddy is an “information appliance”—using Donald Norman’s criteria in
the Invisible Computer (MIT Press, 1998), a “digital tool created to answer a specific need”
that replaces a telephone or hard-to-use computer. It is appropriate for all patients to use,
especially those who either cannot or do not want to use a desktop Web browser, telephone,
or handheld computer.

At home, patients use a simple messaging appliance Health Buddy, which looks like an
oversize Etch a Sketch with four large buttons. The appliance has a high-contrast black-and-
white LCD screen. The Health Buddy plugs into a telephone outlet, it starts working
immediately when plugged in, and does not require a separate phone line or an on–off switch. It uses a toll-free telephone number to connect to Health Hero’s secure Web site [9].

When patients activate their Health Buddy, they answer the health questions that appear on the screen. Using buttons below the screen and a sliding scale, they describe their symptoms (for example, pain or dizziness) about their chronic condition (diabetes or hypertension). The patients’ nurses access these responses through a secure Web site. Because the system flags high-risk patients, the nurses can take action. If patients have a telephone and the Health Buddy plugged in at home, nurses can call them soon after the Health Buddy has finished its “call” to the provider’s Web site.

Clinicians access the patient data remotely through the Web site to look for unusual activity. This monitoring process lets doctors change a medicine dosage or decides when it’s time for an office visit. Healthcare industry experts predict that in the near future more of this type of implanted devices will become available. On the other hand, the Internet provides a ubiquitous virtual space in which drug companies can find eligible patients more quickly than using other communication media, such as printed advertisements. Also, the Internet can help collect data and monitor the results of clinical trials.

(b) Biomedical hybrid browsing/search engine:

Biomedical hybrid browsing/search engine that simplifies access to needed information inside big clinical text. It illustrates the interface of this system (MED READ FAST (MRF)) that was made possible by recent advances in Natural Language Parsing (NLP), Machine Learning, and Data Management. MRF automatically analyzes big texts and creates a structured summary based on the corpus domain and its main concepts. This index automatically constructed for thousands of unclassified electronic patient records. We can see from this index that the corpus discusses Creatinine, Pain and Blood Pressure without searching through it. User can pick highly relevant patient record concepts from the index, compared to what would have been returned by a standard keyword-search.

Naturally, one of the key problems here is automatic extraction of this structural index to access the corpus. In general it is a very hard challenge, because there are many different domains and data representation. Therefore, even the state-of-the-art information extraction systems, Semantic Web search-engines, and other solutions usually focus on a subset of objects, a specific domain, extraction patterns, specific corpus or another reasonable limitation to make fully automatic extraction feasible [10].

(c) Social network analysis and data mining in medical referral process:

Social Network Analysis (SNA) and DM techniques can be used to solve some issues occurring in the medical referral process. Three social networks of General Practitioner (GPs) and Specialists (SPs) are developed and then analyzed to discover doctors’ communities and hidden patterns that will be useful to solve the issues related to the medical referral process [11]. SNA could benefit from data mining techniques as a pre-processing step to prepare the social network data. The main data mining technique to be used is clustering which is the assignment of a set of observations into subsets (called clusters) so that all the observations in the same cluster are similar in some sense. SNA and mining have received a good attention by researchers in the health care domain.

This way, indeed, causes the following problems: (1) A delay in the referral process if the GPs are just new and spend a long time trying to consult with other GPs to find the right
specialist. (2) GPs do not refer patients to specialists with the right skills. (3) Patients are referred to the right specialists, but the waiting time is too long where there are other available SP with the right skills who can serve the patients in a shorter time. Contribute of the SNA are: They incorporate SNA measures with data mining techniques and apply them to a very important domain (health care domain) to help the society to have a better medical referral process. They use SNA measures to find hidden referral patterns such as future consultations between GPs. This enhancement will increase the density of the network at some point in the future. They used SNA measures to work as a disease discovery.

[3] RELATED WORK

The Big data analytics and web mining including machine learning techniques have been studied by many researchers. Although there are many research conducted in past for analyzing the real time health care systems. These mining techniques plays an important role for new trends in health care industry which in turn helpful for all organization associated with this field. This survey also explores challenges and future issues of big data analytics in healthcare.

Google Health is a personal health information centralization service by Google introduced in 2008, which can help people organize, monitor, track, and act on their health information. This service allows the users to enter their health records manually in the Google Health system or by logging into their accounts at partnered health service providers. This records merged as separate health records into a centralized Google Health profile. According to [12] the healthcare data from each consultation will be stored in a medical database. The medical database records information in a concise format with compressed detailed clinical coding that includes symptoms, prescriptions, diagnostic results, treatments and other medical information for the consultation. When a piece of particular medical information is retrieved for further or next health care consultation/reference, the compressed medical data must be extracted to an understandable format for GPs. Volunteered information can include health conditions, medications, allergies, and laboratory result. Google Health uses these data to give the user with a merged health record, information on conditions, and possible interactions between conditions, medicines and allergies.

KarthikKambatla et.al [13], have proposed an overview of the state-of-the-art and focus on emerging trends to highlight the software, hardware and application landscape of big-data analytics. Data repositories for such applications currently exceed exa bytes and are rapidly growing in size. Beyond their sheer magnitude, these data sets and associated applications’ considerations stance significant challenges for method and software development. Datasets are often distributed and its size and privacy considerations warrant distributed techniques. Data often resides on platforms with widely changing computational and network capabilities. Considerations of fault-tolerance, security, and access control are critical in many applications. Analysis tasks often have hard deadlines, and data quality is a major concern in yet other applications. For most developing applications, data-driven models and methods, capable of operating at scale, are as-yet unknown. Even when known methods can be measured, validation of results was a major issue. Characteristics of hardware platforms and the software stack fundamentally impact data analytics.
Srivathsan M et al. [14], have found that Big data was a modern age concept that was used to process large amount of data in various fields ranging from remote sensing, medical, customer service etc. The Medical Sphere was a tangential aspect to individual's life. Technological advancement in this field has reached a saturation. A break-through can be achieved by Prognotive computing. Prognotive Computing was related to big data analytics as the process may require the collection, processing and analysis of huge large volume of structured and unstructured bio medical data stemming from a varied range of experiments and surveys collected by hospitals, laboratories, pharmaceutical companies or even social media which is implemented by using available tools for Big Data.

M.M. Sufyan Beg et al. [15], have proposed personalized web search system to maintain the search profile of each user, on the basis of which the search results would be determined. This requires the integration of techniques for measuring search quality, learning from the user input and biased rank aggregation, etc. For the purpose of measuring web search quality, the “user satisfaction” was gauged by the sequence in which he picks up the results, the time they spends at those documents and whether or not he prints, saves, bookmarks, e-mails to someone or copies-and-pastes a portion of that document. For rank aggregation, they adopt and evaluate the classical fuzzy rank ordering techniques for web applications, and also a few novel techniques that outshine the existing techniques. A “user satisfaction” guided web search procedure is also put forward. Learning from the user feedback proceeds in such a way that there is a development in the ranking of the documents that are consistently preferred by the users.

Yueh-Min Huang et al. [16], have proposed web usage mining was widely applied in various areas, and dynamic recommendation on one web usage mining application. However, most of the current recommendation mechanisms need to generate all association rules before recommendations. This takes huge amount of time in offline computation, and cannot provide real-time recommendations for online users. They used a Navigational Pattern Tree structure for storing the web accessing information. Besides, the Navigational Pattern Tree helps incremental growth for immediately modeling web usage behavior. To provide real-time recommendations efficiently, they develop a Navigational Pattern mining (NP-miner) algorithm for discovering frequent sequential patterns on the Navigational Pattern Tree. According to historical patterns, the NP-miner scans relevant sub-trees of the Navigational Pattern Tree repeatedly for generating candidate recommendations.

Divya Tomar and Sonali Agarwal [17], have focused on usage of various Data Mining (DM) Techniques in health care domain and worked on appropriate way to find DM techniques for a specific problem. While applying classification, need to understand the unwanted attributes because these attributes slow down the processing task. They conducted that there is no single DM technique which give consistent results for all types of healthcare data. So, to give the better performance based on the problem any one of the following technique can be used which are hybrid / integrated techniques.

Steve Lawrence and C. Lee Giles [18], have proposed that World Wide Web has revolutionized the way people access information, and has created up new possibilities in areas such as digital libraries, general and scientific information dissemination and retrieval, commerce, education, entertainment, government, and health care. There are many avenues
for improvement of the Web retrievals; for example, in the areas of locating and organizing information. Current techniques for access to both scientific and general information on the Web provide much room for improvement; search engines is not providing comprehensive indices of the Web and have huge difficulty in accurately ranking the relevance of results. Scientific information on the Web too much disorganized. They discuss the effectiveness of Web search engines, including results giving that the major Web search engines cover only a less amount of the “publicly indexable Web.”

Michael Chau et.al [19], have proposed a machine-learning-based approach that combines Web content analysis and Web structure analysis. As the Web continues to grow, it has become increasingly difficult to search for relevant information using traditional search engines. Topic-specific search engines give an alternative way to support efficient information retrieval on the Web by providing more specific and customized searching in various domains. However, developers of topic-specific search engines require to address two issues: how to locate relevant documents (URLs) on the Web and how to filter out irrelevant documents from a set of documents composed from the Web. They represent each Web page by a set of content-based and link-based features, which can be given as the input for various machine learning algorithms. The approach was applied using both a feedforward/backpropagation neural network and a support vector machine.

Metaxas et.al [20] have proposed a comprehensive information processing, knowledge discovery and simulation platform for Big Data Healthcare. In addition, they present a related, well defined workflow that promotes model-guided personalized medicine. They start by identifying disease signatures and homogeneous patient groups, whilst modeling case-based patient similarity. Then they analyze correlations between variables and patient groups to deliver accurate and reusable predictive statistical simulation models. This framework provides significant advantages on both the clinician’s daily routine and targeted biomedical research.

Chih-Ming Chen [21], have proposed an intelligent systems with physiology signal monitoring for e-health care in big data is an emerging area of development, owing to the urgent needs of homecare for elderly people suffering chronic or unexpected diseases at home. Importantly, a physiology signal monitoring system can help medical staff to monitor and analyze physiology symptoms effectively, such that they can not only monitor the patients’ physiology states immediately, and also helps to reduce medical cost and avoid having to visit doctors in hospital. Therefore, use of system on chip (SOC) techniques to develop an embedded human pulse monitoring system with intelligent data analysis mechanism for disease detection and long-term health care. They also has a friendly web based interface for medical staff to observe immediate pulse signals for remote treatment. It provides aids long-distance medical treatment, exploring trends of potential chronic diseases, and urgent situations informing for unexpected diseases. Moreover, an intelligent data analysis scheme based on the modified cosine similarity measure to diagnose abnormal pulses for exploring potential chronic diseases.

**Review Objectives:**

The main objective of this literature survey is to provide an overview of online health information-seeking behavior by users from the perspective of both health consumer and
health professional. The major parameters that affect the review include a number of disciplines that include medicine, health promotion, social marketing, psychology and information technology. Several issues in documenting Internet accessibility, usage patterns and documenting, online health information consumer profiles, identifying online sources of health information, health professionals Internet use and ascertaining challenges for health professionals posed by Internet use are considered for literature review.

Table: 1 shows the statistics on how Internet has impacted the Health Care Industry
<table>
<thead>
<tr>
<th>SL No</th>
<th>Meaningful statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>More than 40% of consumers say that information found via social media affects the way they deal with their health.</td>
</tr>
<tr>
<td>2.</td>
<td>18 to 24 year olds are more than 2x as likely as 45 to 54 year olds to use social media for health-related discussions.</td>
</tr>
<tr>
<td>3.</td>
<td>90% of respondents from 18 to 24 years of age said they would trust medical information shared by others on their social media networks.</td>
</tr>
<tr>
<td>4.</td>
<td>31% of health care organizations have specific social media guidelines in writing.</td>
</tr>
<tr>
<td>5.</td>
<td>19% of smartphone owners have at least one health app on their phone. Exercise, diet, and weight apps are the most popular types.</td>
</tr>
<tr>
<td>6.</td>
<td>From a recent study, 54% of patients are very comfortable with their providers seeking advice from online communities to better treat their conditions.</td>
</tr>
<tr>
<td>7.</td>
<td>31% of health care professionals use social media for professional networking.</td>
</tr>
<tr>
<td>8.</td>
<td>41% of people said social media would affect their choice of a specific doctor, hospital, or medical facility.</td>
</tr>
<tr>
<td>9.</td>
<td>30% of adults are likely to share information about their health on social media sites with other patients, 47% with doctors, 43% with hospitals, 38% with a health insurance company and 32% with a drug company.</td>
</tr>
<tr>
<td>10.</td>
<td>26% of all hospitals in the US participate in social media.</td>
</tr>
<tr>
<td>11.</td>
<td>The most accessed online resources for health related information are: 56% searched WebMD, 31% on Wikipedia, 29% on health magazine websites, 17% used Facebook, 15% used YouTube, 13% used a blog or multiple blogs, 12% used patient communities, 6% used Twitter and 27% used none of the above.</td>
</tr>
<tr>
<td>12.</td>
<td>Parents are more likely to seek medical answers online, 22% use Facebook and 20% use YouTube. Of non-parents, 14% use Facebook and 12% use YouTube to search for health care related topics.</td>
</tr>
<tr>
<td>13.</td>
<td>60% of doctors say social media improves the quality of care delivered to patients.</td>
</tr>
<tr>
<td>14.</td>
<td>2/3 of doctors are use social media for professional purposes, often preferring an open forum as opposed to a physician-only online community.</td>
</tr>
<tr>
<td>15.</td>
<td>YouTube traffic to hospital sites has increased 119% year-over-year.</td>
</tr>
<tr>
<td>16.</td>
<td>International Telecommunications Union estimates that global penetration of mobile devices has reached 87% as of 2011.</td>
</tr>
<tr>
<td>17.</td>
<td>28% of health-related conversations on Facebook are supporting health-related causes, followed by 27% of people commenting about health experiences or updates.</td>
</tr>
<tr>
<td>18.</td>
<td>60% of social media users are the most likely to trust social media posts and activity by doctors over any other group.</td>
</tr>
<tr>
<td>19.</td>
<td>23% of drug companies have not addressed security and privacy in terms of social media.</td>
</tr>
</tbody>
</table>
20. The Mayo Clinic’s podcast listeners rose by 76,000 after the clinic started using social media

21. 60% of physicians most popular activities on social are following what colleagues are sharing and discussing

22. 49% of those polled expect to hear from their doctor when requesting an appointment or follow-up discussion via social media within a few hours

23. 40% of people polled said information found on social media affects how someone coped with a chronic condition, their view of diet and exercise and their selection of a physician

24. Of more than 1,500 hospitals nationwide who have an online presence, Facebook is most popular

* Source - getreferralmd.com/2013/09/healthcare-social-media-statistics

[4] CONCLUSION

In this paper a survey of the use of web in health care analytics is done. Various health related problem that mainly affect the human beings are also considered and according to these problems several solution schemes that can partly solve the health care problems is performed. An analysis of health care issues in the big data analytics, which may be a helpful to everyone those are using social networks and usage of web mining in health care have gain enough attention. The existence of problems in health care has made it necessary to find some effective solutions and protect the people from the various problems. Machine learning in big data acts as guidance to the health care related research works in this area. Several web mining approaches in health care domain can help protect the people from various diseases. Future work is to find ideas that can be further explored to find the solution for some aspects of health related problems like obesity, cholesterol and the application of machine learning techniques to improve the performance
REFERENCES


Author[s] brief Introduction

Priyanga P received her Masters Degree in Computer Science and Engineering in 2007 from Anna University, India. She is currently pursuing her PhD from Vishveshwarya Technology University, India. She is currently working as Assistant Professor, Department of CSE, KSIT, Bangalore. She has 9 years of work experience in teaching in India. She has published 8 research papers in National and International Journals and conferences. Her area of research includes Big Data analytics, Data Mining.

Dr Naveen NC received his PhD from SRM University, India. He is currently working as Professor & HoD, Department of ISE, Dayananda Sagar Engineering College, Bangalore. He has 20 years of work experience in teaching with Research, Administration and Industry. He has published more than 30 papers in national and international conferences and journals. He is guiding PhD students in areas of Big data Analytics, Image processing, Data mining.

Acknowledgments

The Authors would like to thank VGST (Vision Group on Science and Technology), Government of Karnataka, India for providing infrastructure facilities through the KFIST Level I project at KSIT, CSE R&D Department.

Corresponding Address-

#409, 2 B Main, 16th cross, JP Nagar, 4th Phase, Bangalore-560078. Phone No. +91-9980422788.