Journal of Engineering and Applied Sciences 12 (20): 5261-5273, 2017
ISSN: 1816-949X

Challenge of Text Mining in Clinical Decision Support System: Review

Ala Aldeen Ismail Abuazab, Harihodin Bin Selamat and Rasimah Bt Che Mohd Yusoff
Faculty of Informatics, Advanced Informatics School, Universiti Teknologi, Johor, Malaysia

Abstract: This study provides an overview of work done by researchers in areas related to Semantic Text Mining (STM). This is particularly important since STM could be considered a generic technique which is applied to signal detection in the work presented and similarly, signal detection could be considered a generic problem which is solved using STM technique in the work presented. A closer look at the evolution and challenges to be addressed provides a good starting point in arriving at the solution framework. In the literature review also cover of NLP and predication diseases for general diseases not particular disease. The detail of Clinical Decision Support System (CDSS) and the research studies in the field converting unstructured medical data which doctor notes into structured medical data (NLP) and the research on prediction diseases framework regarding the aid of CDSS.

Key words: Text mining, medical data, decision support system, structured, particularly important, NLP

INTRODUCTION

Text mining is used for several areas of integrated techniques such as data mining, web mining, statistics, Information Retrieval (IR) and Natural Language Processing (NLP). It gives us how to detect new information from the data (document) based on text which is unclear and hard to find. That is it means a text-processing technique which encompasses information retrieval, information extraction, information systematization and information analysis.

Researchers started publishing results on their research in this area as early as 1970 (Cleverdon, 1970). While the results were encouraging, handling large volume of data in their native human readable electronic document form was beyond the capabilities of available compute, storage and communication technologies till few decades ago due to considerations related to cost and size of electronic memory and processor. Accordingly, researchers and developers made the catalogue or indices searchable online while still accessing the document offline in physical form. This resulted in the forming of Machine-Readable Cataloging (MARC) by the Library of Congress in the 1960s (Avram, 1968). These new areas of research are extending structured data mining capabilities to information at large including documents containing arbitrary data and exploit semantic relationships as well as metadata associated with them. Domain independent applications have also been tried as reported by Gatterbauer *et al*. (2007) and Sarawagi (2008). Group of like minded communities coin new words or use existing ones with new meanings and ontologies based on those have come to be known as folksonomies. It looks like while ontology is driven top-down and controlled approach, folksonomy is a bottom-up and user driven approach. Emerging social networking or web 2.0 makes use of folksonomies as reported by Jaschke *et al*. (2008) and Hotho *et al*. (2006). The main purpose of text mining is to convert text to the analysis-enabled data through the application NLP and analysis techniques. The ability to extend mining, search and retrieval to documents including rich media documents in different formats provide exciting possibilities but not without equally exciting challenges.

Semantic text mining: Semantic Text Mining (STM) is a technique that has grown immensely which has the capacity to handle vast amount of textual data (Srivastava and Sahami, 2009). STM offers methods to process narrative text as they are thereby offering a method of either eliminating or reducing the ETL process to a large extent. However, it has been more popular with web where the documents are annotated. STM for non-annotated text is a challenge which if solved would provide automated signal detection capability for regulatory and monitoring agencies. The Text Mining based Hadoop platform decides a patient's disease, predict his disease and provides the more accurate information about diseases by converting the unstructured data among the patient's collected information to structured data. The Apache Hadoop project develops open-source software for reliable,

Corresponding Author: Ala Aldeen Ismail Abuazab, Faculty of Informatics, Advanced Informatics School, Universiti Teknologi, Johor, Malaysia

scalable and distributed computing. It consists of HDFS (Hadoop Distributed File System), Hbase and Hadoop MapReduce which can analyze big data. It is open source framework that writes and implements an application program for processing big data. Therefore, in spite of known advantages of electronic data processing capabilities, creating large scale mining applications with online document access could not proliferate. This resulted in the emergence of the Database Management Systems (DBMS). DBMS provided an efficient way of storing annotated/labelled subset of key elements out of the documents. DBMS has evolved over time and is currently capable of handling large volumes of data and provides mining capabilities, helped by the rapid innovations in compute, storage and communication technologies. In spite of the maturity of DBMS and its value to organized and structured data, the efforts needed to make it function necessitates converting human readable documents to a format that could be processed by computers. This involves strict data capture process or if already captured in other formats, data conversion process. Generally, parts of the information are defined in data fields of specific data types. Only those data stored in the fields and any derivative information by manipulating them are available for further access. In a nutshell, only select facts and figures remain out of the source document. While this may suffice for some applications, the question and interest remains as to why is the entire document not available for analysis. Researchers are addressing this question by developing techniques for semantic web (Berners-Lee *et al*., 2001) including web 2.0 and Information Retrieval (IR) (Manning *et al*., 2008).

Semantic web: Semantic web has witnessed tremendous activities in this decade. A number of standards, specifications and languages have come up in its support. Semantic web as it is accepted and understood is based on an article published by Berners Lee *et al.* (2001) and Shadbolt *et al.* (2006). Sheth *et al.* (2015) and Berner *et al.* (2005) categorize semantics into implicit, formal and soft to explain various capabilities by each type of semantics by Sheth (2005). Different languages for supporting semantic web (Bailey *et al*., 2005) have come into existence like XML, RDF, OWL and SWRL Horrocks *et al.* (2004). These languages provide the flexibility of choosing between expressiveness and computability which are inversely proportional. Semantic web applications have been broadly categorized into two major areas; web usage mining and web content mining as discussed in various studies viz. (Kolari and Joshi, 2004; Han and Chang, 2002). Domain specific ontologies have been used to achieve better results like the ones for gene ontology (GOH) and legal (Breuker and Hoekstra, 2004). These domain specific ontologies are essential in taking the semantic web to the next level but ontologies themselves change as with any natural language. Further these ontologies could be with few levels, for example, the Medical Dictionary for Regulatory Activities with just five levels (MedDRA, 2009) or many levels like the gene ontology.

Personalizing user search experience, many of the solutions based on query or usage mining studies, has drawn a lot of attention with various techniques reported in articles including those by Sieg *et al.* (2007), Jansen and Spink (2006), Pierrakos *et al.* (2003), Srivastava *et al.* (2000), Mobasher (2007). However, when it comes to web content mining, the multidimensional and deep links have posed challenges and the progress has been slow. Only few prototype applications have been reported like by Sheth (2005) and Meza *et al.* (2008) where an application for detecting conflicts of interest using various RDF and other semantic languages is reported, and by Zaremba *et al.* (2006) a virtual travel agency using web services execution and modelling is reported. For more details on the evolution, current state and standardization of semantic web one can refer by W3C. While all these efforts are resulting in access to huge decentralized data resources, the issues that need to be resolved include querying them with confidence and trust, handling inconsistencies, inter and intra domain meanings leading to ambiguities and other natural language related issues remain to be addressed. These limitations, researchers have used available capabilities of semantic web for different applications. In essence, semantic web including web 2.0 and social networking rely on explicit semantic in the form metadata being defined and available along with their content. This explains the reasons for the limited number of applications using web content mining as the contents tend to be without sufficient semantics. For example, description of a news item in the form of narrative text is at best related to its title and may be date and source of the news. This does not provide enough information for the semantic web to successfully exploit the contents which is essentially lack of annotation. Alternate approaches using ontologies like the one by GOH have been tried where feasible with limited success. Organization document repositories are likely to have vast amount of narrative text with very little annotation or relation to ontologies.

Text mining and information retrieval: Text Mining (TM) handles text as a “bag-of-words” and analyzes them without considering associated semantics. Adding context or semantics to text data enables better processing and results in higher quality results. Hence, STM is more appropriate for developing solutions. STM enables accessing analyzing and processing vast amount of unstructured textual data to uncover hidden patterns taking semantics into consideration. Such semantics could be based on linguistic properties, domain ontology, etc. Realizing the potential of STM a number of traditional data mining solutions are trying to expand their coverage to unstructured data by using tools to support ETL process as could be seen from the Gartner report (Gilbert and Friedman, 2006). STM is designed to handle data as they are in their native representation which makes it attractive for a number of applications. Intuitively it could be seen that converting the majority unstructured data to structured data as existing applications are based on the structured data, instead of other way around is not optimal.

Further, structured data could be treated as unstructured data in a way while the reverse may not be true in most of the cases. Hence any application supporting unstructured data would include structured data handling as well. This makes STM an attractive solution option. Current STM solutions are capable of handling annotated data as well as data following a specific ontology. Any advance that could be made in STM significantly adds to the knowledge management capabilities that exist today as most of the existing data happens to be unstructured. As already indicated unstructured data could be as high as 90% of all data. In this research, a framework for STM using textual regulatory document repositories for signal detection is presented. The vast unstructured data available in electronic format has resulted in researchers making progress in rapid and automated information retrieval system over the last about 50 years. If the stored information pertains only to text as against other types of data like images and videos such retrieval systems could be called TM systems. The focus here is only on TM systems. IR is a major contributor to TM field. An overview of TM using IR in its early stages is presented (Rijsbergen, 1979). The researcher provides a good explanation of what an IR system means for automated systems as well as takes through various stages from storage to retrieval. Since, then the domain of IR has progressed. Association for Computing Machinery's (ACM) Special Interest Group on Information Retrieval (SIGIR) has been tracking the progress of IR since 1971 (SIGIR, 2009) with conferences, publications and forums. The main challenge for IR with respect to annotated or structured Data Retrieval (DR) could be summarized as shown in Table 1.

Table 1: Key differences between IR and DR

Variables	Information Retrieval (IR)	Data Retrieval (DR)
Search criteria matching	Partial or best match	Complete match
Inference	Employs induction	Employs deduction
Model	Probabilistic	Deterministic
Classification	Multiple classes	Specific class
Query language	Natural	Artificial
Expected results	Incomplete	Complete
Results score	Relevancy score	Matching score
Effect of errors in query	Insensitive	Sensitive

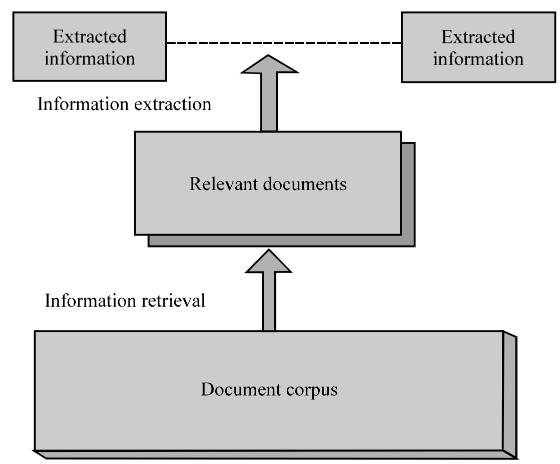

Fig. 1: Relation between IR and IE

It may be noted that the parameters chosen to highlight the differences between IR and DR are for discussion purposes only and some of these distinctions may not exist in what is observed in applications. Few key parameters used in IR include:

- Precision; refers to the fraction of retrieved documents that are relevant to the query/search with respect to the total documents retrieved
- Recall; refers to the fraction of retrieved documents that are relevant to the query/search with respect to the total number of relevant documents present in the document corpus
- Relevance score; a score to indicate the relevancy of the retrieved document against the query. Generally, a function using term frequency (tf) and inverse document frequency (idf) is used for arriving at the score
- Term frequency; gtimes a given term occurs in the document
- Inverse document frequency; generally refers to the ratio of the total number of documents to the number of documents containing the specific term on logarithmic scale

Information extraction: IR systems provide relevant documents. However, the documents could be large and the specific information being searched may be difficult to locate. This resulted in the evolution of IE.

The relation between IR and IE is shown in Fig. 1. The ability to extract useful information from relevant documents retrieved using IR has encouraged use of available meta data as well as structured within documents. This lends itself suitable for signal detection, which is the challenge being addressed in this thesis. A detailed survey of about two hundred articles related to IE with techniques and applications has been done (Sarawagi, 2008). It may be observed from the above discussions that adding IE-IR systems would enhance TM systems to STM systems. In the next section a brief overview of signal detection is provided.

Currently they include manual chart review by domain experts, statistical text mining, natural language processing, and machine. Each of these methods has their own particular strengths and weaknesses.

Document annotation: In document annotation, subject matter experts review the documents and code strings of text that represent the targeted concepts and the relationships between the concepts. In the medical domain we do "chart review" to annotate the concepts in progress notes. Doing manual chart review for annotation has been used extensively and when appropriate rigor is applied, the information extracted is very reliable and is often used as the "gold standard" to evaluate machine methods. However, it is a very expensive and time-consuming effort. Clinicians are recruited and paid full salaries to read each and every progress note. There are also concerns that over time mental fatigue will cause human error in the coding. To minimize such issues, multiple annotators can review the same documents but this further raising the cost and time commitment.

Regular expressions: Regular expressions are effective when the structure of the text and the terms of interest are consistent but these are essentially one-off methods that must be tailored to the extraction task. Regular expressions are bits of code containing characters that represent patterns in the way a concept is found in the text. A combination of wild cards, character classes and specific or literal text characters are used in a specified order. They can be compiled and used in a computer programs so they do have the advantage of speed. They are usually hand coded by a human being and can involve a complex and time consuming development effort, because in order to be inclusive we must know a priori all of the possible patterns that represent the concept in the text.

Natural language processing: Natural Language Processing (NLP) has been effective at identifying concepts in written text but requires that the text must be written in complete grammatically correct sentences. In NLP, the text is tokenized and parsed for sentence boundaries and parts of speech are identified. The noun phrases or verb phrases are then used to look up concepts in a domain specific controlled vocabulary or ontology to find one or more concepts represented in the text. A substantial amount of the text in medical progress notes is found in template form. The text parts of templates are anything but complete or grammatically correct so there are problems with part of speech tagging, negation and word sense disambiguation.

Medical progress notes are categorized by type according to their purpose. Currently there are hundreds of note types in use in the VA. Within the note there is a coarse order in which information is recorded according to the SOAP (Subjective, Objective, Assessment, Plan) format. Each section of SOAP can be further divided into sections based on content and structural features. The detailed structure of the text of each note type can vary greatly, limiting the effectiveness of approaches like rules or regular expressions.

While some work has been done for information extraction from semi-structured data, most of that research focuses on extracting data from web pages or from research articles and is not easily adapted to the medical domain (Smyth and Goodman, 1992; Uthurusamy, 1996; Frawley *et al*., 1992).

Methods of information extraction as applied to medical progress notes could be improved if we had better methods of identifying the semi-structured data in the notes. New machine-assisted methods need to be developed to identify the semi-structured information in medical progress notes. If these automated methods were developed and found to be effective, then advances in diagnosis and treatment based on text analysis could come significantly sooner and at less cost.

MATERIALS AND METHODS

researchers text mining evaluation methods: There are various bases for comparison of summarization system performance, e.g., summary to source, system to hum an-produced summary and system to system. In general, methods for evaluating text summarization approaches can be broadly classified into two categories; extrinsic evaluation (function evaluation) and intrinsic evaluation (performance evaluation). In extrinsic evaluation, the quality of a summary is judged based on how it affects the completion of some other task, e.g., determining the relevance of documents to topics and humans answering questions based on reading the summaries (Maybury, 1999). In intrinsic evaluation, humans judge the quality of the summarization directly based on analysis of the summary, e.g., the fluency of the summary, the coverage of "key" facts or similarity to an "ideal" summary. An ideal summary is hard to establish.

The human summary can be supplied by the researcher of the study by a judge asked to construct an abstract or by a judge asked to extract sentences. There can be a large number of generic and use focused summaries that can summarize a given document. The lack of uniqueness of the ideal summary is a major problem for the intrinsic evaluation. Research has shown that when a judge is required to extract sentences from a full-source text to construct an ideal summary which sentences are selected depends in part on what instructions are given to the judge (Marcu, 1999). Another study indicated that different abstractors may produce very different summaries and that an abstractor given the same document after 8 weeks may produce a substantially different summary (Rath *et al*., 1961). Another problem with trying to create an ideal summary is that there is evidence of low agreement among humans as to which sentences are good summary sentences. Salton *et al*. (1997) indicated in their research that subjects showed <50% overlap in their extracts when asked to extract at least five paragraphs from each of 50 articles from an encyclopedia. In addition, different compression rates make a considerable difference in what gets extracted.

Morris *et al*. (1992) experimented on an extrinsic evaluation task of question answering. The author picked four Graduate Management Admission Test (GMAT) reading comprehension exercises. The researcher measured how m any of the answers the subjects got correct under different conditions including a full-text condition (where the subjects were shown the original text) an extract condition (where the subjects were shown automatically generated generic extracts from the passages) and abstract condition (where the subjects were shown a generic hum an abstract of 25% compression created by a professional abstractor) and a control of no text (where the subjects picked the answer without reading any text). Their results showed that the performances of the extracts and abstracts were comparable to the fulltext. This suggests that summaries can be effective in certain tasks as substitutes for full text.

There are a number of the fundamental problems that exist in designing evaluations of text summarization/abstracting. Previous evaluations have provided some insights on how to conduct an evaluation (Matsuo and Ishizuka, 2004):

- In extrinsic evaluations, the task should adequately model real-world situations and information needs
- In intrinsic evaluations, clear instructions should be given to ensure a certain level of consistency
- A control experiment is useful to provide statistically sound results
- Adequate metrics of summarization accuracy or efficiency should be developed to measure the performance

Challenges in medical data discovery: The application of Data Mining Discovery and machine learning techniques to medical and health data is challenging and intriguing (Ronald *et al*., 1997; Belkin and Niyogi, 2004). The data sets usually are very large, complex, heterogeneous and hierarchical and vary in quality in spite of which there exists a huge knowledge base that demands a forceful alliance between Data Miners and healthcare professionals if any useful and previously unknown information (Adriaans and Zantinge, 1996; Ronald and Anand 1996; Kuperman *et al*., 2007; Matheus *et al*., 1993) is to be discovered and extracted. Medical data mining typically analyzes data that is generated by some experimentation or collected as part of a clinical routine or simply in the course of study of medicine. A common goal of the data mining is the detection of some kind of correlation, e.g., between genetic features and phenotypes or between medical treatment and reaction of patients. The analysis presents the experiences and issues encountered by my research and others in applying the data mining techniques to medical and clinical data. In the process it brings forth a number of generally experienced issues, the thing to keep in mind and overall factors to consider during the mining process. Electronic health records facilitate capturing transaction data and reporting that support the health industry. The uniqueness of medical data when captured during laboratory and clinical process includes issues of data availability and composite representation models make the data mining task challenging. Data preprocessing and transformation are required even before mining and discovery can be applied. Sometimes the characteristics of the data may not be optimal for mining or analytic processing. The challenge here is to convert the data into appropriate form before any leaning or mining can begin. Before, we begin any automatic learning model to extract useful data the data must be presented in the form acceptable to the learning method. Some of the issues are highlighted above for example some algorithms did not allow more than 30 distinct values for a dependent or to be explored variable. One of the techniques that have been widely accepted is to "flatten" the table. In the flattened model each row represents a case for training and or testing and each column represents the values for a variable across the various cases. Despite its simple and commonly used in the analysis of data, it is not the typical format used while capturing the data.

Poor data quality, encoding and inconsistent representation: There are a number of issues that must be addressed before any data mining can occur. The available datasets range from convenient, accurate, indexed and well managed to those that are incomplete, inconsistent, potentially inaccurate and extremely large. Unlike other domains like financial, demographic and geographic areas medical data is diverse, complex and to a non healthcare professional hard to interpret. Therefore it demands a forceful collaboration between the domain specialist and the data miner. Inconsistencies due to data entry errors are a common problem. Inconsistencies due to data representation can exist if more than one model for expressing a specific meaning exits (e.g., For a positive Thrombosis, one application may enter Yes/No and other may enter (+/-) and some others may be free text) additionally the data type does not always reflect the true data type. For example a column with numerical data type can represent a nominal or ordinal variable encoded with numbers instead of a continuous variable. This plays an important role during statistical analysis (mean and variance). Regional and geographical locations can play also part in capturing consistent data. For example a date format in the US is generally denoted as mm/dd/yy while in some other countries it may be denoted as dd.mm.yyyy. These differences may be subtle but pose additional preparation and transformation.

Storage structure: The database structure of raw medical/clinical data is usually structured in a way to facilitate online transaction systems which in turn needs optimization for patient based transactions with indexes and structures that cater to single patient transaction. The database structures work effectively with transactions involving the data of individual patients but are not effective with trans-population queries. For data mining and statistical analysis data must be populated that can ultimately be rendered as a flattened table.

Poor mathematical characterization of data: Business, financial, scientific data can be easily modeled, transformed and applied formulas in contrast to medical data whose underlying structure is poorly classified in mathematical terms. Medical data consists of images and free hand data with few constraints on vocabulary or image. In comparison business and financial data have formal structures into which we can classify and organize data that may be modeled by linear regression, neural networks and naive bayes versus medical attributes such as bloating, inflammation and swelling. However, with advances in technology and faster computers along with advanced tools of data mining some of the issues may no longer be relevant.

Poor integration: The fragmented and distributed nature of health data between hospitals, insurance companies and government departments poses substantial challenge for data integration and therefore data mining in terms of the confidence that can be placed in the result and the semantics of a derived rule. For example data may not include the patient's diagnosis for each episode and where the data is used for research purposes the diagnosis sometimes has to be inferred from pathology tests carried out or from prescribed medications. This pushes current data mining for medical/clinical data to their limits and it is this aspect that promises to provide a practical insight into some of the possible future directions for knowledge discovery systems more generally. Use of a common data dictionary and agreeing upon common standards is being seen as an important technique in standardizing and integrating data from heterogeneous systems. The emergence of XML as a data standard is gaining wider acceptance and hence making integration fairly easy. But for all medicate data to be converted to XML form has a long way to go.

Number of variables: For certain algorithms where the computational complexity is not linear, the time required may become infeasible as the number of variables grow. However if the number of variables for each patient can be >1000 which makes many algorithms impractical, some may go into an endless discovery and the time taken can increase exponentially.

Missing data: Clinical data elements often are not collected for all data required for analysis or discovery. Some data elements are not collected due to omission, not relevant, excess risk or inapplicability in a specific clinical context. For some model learning methods like logistic regression a complete set of data elements may be required. Even when methods that accept missing values are used, the fact that the data was not collected may have independent information value and should not be ignored. Methods of data transformation and of modeling the missing data are necessary to cope with this issue.

Imaging and complexity of medical data: With advances in medicine and imaging technology the complexity of data has increased fourfold (Ronald and Anand, 1996; Dhar and Tuzhilin, 1993). Increasingly clinical procedures are employing imaging as a tool of choice for diagnosis. There needs to exist efficient mining in databases of images which are much more difficult than mining purely numerical data. As an example cardiac ECG and EEG signals can generate several gigabytes of data daily. Some procedures may produce two dimensional or three dimensional images. They almost frequently accompany other clinical information such as physician's comments or lab analysis. This requires low cost, highly efficient storage devices and new highly efficient tools to analyze the data. It is very difficult for humans to process gigabytes of information in contrast to deciphering images due to the fact that humans can identify, relate and recognize patterns and trends. The information stored become less relevant if it cannot be comprehended easily. Imaging and visualization tools will be indispensable as imaging, X-ray, cat scans, etc., become more prevalent and necessary diagnostic techniques and the challenges to interpret and comprehend them in data mining will grow exponentially.

Ethical, legal and privacy issues: Clinical and medical data is primarily focused on humans and thereby becomes a primary target for abuse and misuse. Effective legal and ethical frameworks are in place to prevent their misuse. Inadvertently the ability to compromise privacy (Han and Chang, 2002) followed by litigation is a major concern among the medical community. Adding to the complexity is the question of data ownership (Ronald and Anand, 1996). Several terabytes of data is collected across the world across heterogeneous databases in varying format without a common standard. The question that has perplexed the medical community is determining the owner of the datathe patient, physicians or insurance companies?

Very recently an article was published by the Boston Globe, 2007 titled "Know your Customer". It talked about lawmakers around the country are taking a hard look at data mining companies that keep detailed records what prescription drugs are prescribed by nearly every doctor in the USA reputed medical data mining whose databases are updated weekly stripped of patients name and sold to drug companies. Recent leaks of personal data at the Health and Human Services (HHS) division of federal government once again brought to the forefront the burning questions of privacy and legal liabilities. On the other hand are ethical issues with stem cell research some in favor of betterment of human kind and others to leave it unmined. Yet, although, medicine is based upon science, there are certain tests that may not be performed and certain conclusions may not be drawn, because medicine is more than a pleasure or convenience rather deals with life and deaf.

Health professionals interpretation: The doctors interpretation of tests conducted, imaging and other clinical data is generally written in free text that is very difficult to comprehend and difficult to standardize which poses a great challenge for miners. Many a time professionals from same field cannot agree upon interpreting a patient's condition but also use different names to describe same disease. Very recently at a medical data mining project at the Duke University which was moving the computer based patient records into a data warehouse the highest error rate with unusable records was use of free text. Furthermore the data entered was free text in place of code from the data dictionary.

RESULTS AND DISCUSSION

Interpretation of mining result set: One of the biggest challenges for data miners mining medical data is interpreting the results from discovery vs noise. Without the help of health professional it becomes difficult to interpret the discovery. Unlike other fields medical data is diverse, complex and to a non healthcare professional hard to interpret. Therefore, it demands a powerful partnership between the domain specialist and the data miner. On the other hand is the difficulty in developing a sufficiently detailed understanding of both medicine and data mining to build a system using current conventions which meet the requirement or simply a facilitate a mapping in the simplest form.

Clinical data: There are several data types such textual documents, semantic metadata and management data. All these three different approaches can be integrated in a single system which build for predication diseases platform.

Structured data: Structured data is data follows a predefined schema and relational database system such as patient medical record number MRN, patient names, identification numbers, dates and diagnosis codes. The advantage of structured data is the existing tools and web frameworks. It is easier to collect and exchange between systems because it is standardized, pre-defined, computer-readable and typically quickly accessible from a database. Structured data can prepopulate fields in electronic templates and be selected from pull-down menus, numeric data, Yes/No and static data.

Semi-structured data: Semi-structured data is the entities belonging to the same class may have different clinical attributes even though they are grouped together. The advantage of semi-structured data is the type of data may be defined for a part of the data and it is also possible that a data instance has more than one type. It is a form of structured clinical data that does not conform to the formal structure of data models associated with relational databases or other forms of data tables but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Semi-structured data such as information include Meta data or schema such as XML or HTML.

Unstructured data: Unstructured medical data is the fact that no identifiable structure within this kind of data is available which that typically requires a human touch to read, capture and interpret properly. The advantage of unstructured data is text documents is full text search that is completely decoupled from the data. It includes machine-written and sometime handwritten information on unstructured paper forms, email messages attachments, and typed transcription. In our proposal the Doctor Notes and Doctor patient documentation are sorting as unstructured medical data.

Clinical Decision Support System (CDSS): In making of clinical decisions, healthcare providers rely heavily on CDSS to analyze data. CDSS is commonly used to support business management; it can also be relied on as an adaptation of decision support system (Burkle *et al*., 2001). In improving final results; physicians, nurses and other medical professionals rely on CDSS to prepare diagnosis and also to review the same diagnosis. With relevant clinical research, medical professions may conduct an Information Retrieval to examine on the patient's medical history. Potential events such as disease symptoms and drug interactions are predicted by such analysis. In determining the best course of care, some physicians choose to rely on their own professional experience avoiding over-consultation on their CDSS (Chen and Liu, 2004).

Clinical support system is faced by some pros and cons in its implementation. Integrating the CDSS to a healthcare organization clinical flow is the biggest challenge that CDSS faces. This is mainly due to the already existing complexity in the clinical workflow. Lack of interoperability in reporting and electronic health record software, brought about by standalone product operations, makes it difficult to incorporate the resulting data, mainly due to sheer number in the clinical research and medical trials being published on ongoing basis. There is a significant strain on application and infrastructure maintenance when it comes to incorporation of large amount of data into the existing system (Kuperman *et al*., 2007).

Mixing of incompatible medications which are adverse can be identified and avoided by both nurses and physicians thanks to increased CDSS popularity (Kuo and Fuh, 2011). Caretakers who also receive prompts from other technology systems can also be overwhelmed by alerts triggered by CDSS. Agency for Healthcare Research and Quality (AHRQ) commissioned a study on the effectiveness of CDSS, this agency found out that inappropriate use of CDSS can be more harmful that when it is not deployed. EHR systems have built a growing number CDSS functions. Before working alongside their EHR systems, providers should plan for and eliminate any overlapping alerts prior to selling out a standalone CDSS. An act that required all healthcare providers to demonstrate the meaningful use of health IT by 2015 or face a reduction in Medicare reimbursements, beginning 2016, saw the use of clinical decision support system increase after its passage.

This act was referred to as Health Information Technology for Economic and Clinical Health (HITECH) (Kawamoto *et al*., 2005). Diagnostic test ordering rule as well as ability to track compliance with this rule was one of the meaningful uses that this acts required providers to implement. Specialty or high-priority conditions also applied with the rule. In case of wrong diagnosis, missed or given the wrong dosage of medication, some providers deploy CDSS to flag patients (Iqbal *et al*., 2011) to serve as a basis for improvement initiatives, population health management report, receives all errors in the problem list. Pearson *et al*. (2009) claimed that in improving and streamlining the quality of healthcare delivery, CDSS have been recognized as promising tools in influencing healthcare provider performance. Decision Support Systems (DSS) brought to birth the CDSS. DSS combined individuals and computers capabilities to improve quality of decisions (Donzelli, 2006). DSS popularity and use in the healthcare domain have been mainly contributed by its functionality and capability. A DSS offering support to physicians, minimizing practice variation and improving patient care is defined as CDSS. Throughout their inception in the medical arena in early 1970s (Pearson *et al*., 2009) observed that CDSS have evolved immensely to support the workflow of clinicians and improved the effectiveness of decision outcomes. CDSS technology still remains the most promising technology due to its ability to:

- Enhance clinical decision-making process of healthcare providers and
- Supporting of evidence-based practices despite the several challenges facing it in use and adoption in the healthcare setting

Kawamoto *et al*. (2005) noted from this regard that CDSS provide clinicians with patient-specific assessments or recommendations to aid in clinical decision making. outlined three examples of CDSS technologies:

- Computerized Physician Order Entry (CPOE) systems; this provide patient-specific recommendations that are part of order entry process
- Outpatient systems: This attaches care reminders to the charts of patients in need of specific preventive care services
- Laboratory alerting systems; this system, pages physicians when critical laboratory values are detected

The O'Kane points out that architecture component of CDSS consist of three main areas:

- Inference/reasoning engine; this combines patient's data with knowledge base data
- User communication/interaction; this consist of simple ways on how data is fed into the system and extracting the results to the user
- Knowledge base; this is made up of rules, guidelines and probabilistic models

Based on the analysis of past works on CDSS, the following can be synthesized:

- The CDSS need to use engine to retrieve clinical data from huge data
- The CDSS need to generate some of analysis to support decisions in the hospital for healthcare
- All the researchers proposed data mining to that aims but data mining need to generate some of knowledge based patterns of clinical which is difficult to find it based on our initial searching for health care
- The information retrieval can play a vital role to retrieve the clinical data need to analysis based on some criteria need to decisions makers

Hence, there is a need to choose information retrieval. As such the next section will explain in detail the process of information retrieval.

Clinical decision making: Quality of decision making by healthcare providers is significantly impacted by CDSS. To aid clinical decision making, CDSS provide clinicians with patient specific assessments or recommendations (Kawamoto *et al*., 2005). Quality decision making should not be taken as an easy endeavor. Neill *et al*. (2005) maintain that due to the complexity that arises in the decision making process a clinical officer should pertain the following:

- Knowledge
- Reliable information input and
- Supportive environment

Where cues are used to assign patients to one number of potential categories which consist of classification task this is defined by Buckingham (2002) as clinical decision making. Use of CDSS as a supportive tool has facilitated achievement of quality complex clinical decisions. Linking intuitive explanations of clinical expertise with empirical data analysis would indeed enhance judgment accuracy. Buckingham (2002) proposed a gelatean model, this model would indeed improve the use of CDSS to support quality decision making in clinical practice. This relationship between clinicians and computers was identified as symbiotic.

In explaining this symbiotic relationship, Buckingham (2002) stated that, computer's side of symbiosis comes with its powers in data storage and analysis while the clinicians are responsible for using their psychological validity. To ensure that all information emanating from the CDSS is interpreted well by the attending clinicians, efforts must be made to ensure that a form of enhanced judgment exists. Combining experimental knowledge with the use of CDSS so that symbiotic relationship can be established is a great way to enhancing clinical judgment.

Clinical practice guideline: Systematically developed systems that help both the medical practoneers and the patient on the appropriate health care for specific clinical circumstance is what (Kotze and Brdaroska, 2004) defines as clinical practice g uidelines. Kotze and Brdaroska (2004) point out that unless clinical practice guidelines are effectively implemented and integrated into the clinical setting they will have little influence upon clinician practice and patient outcome.

Table 2: Summary of clinical decision support system and their problems

Researchers	Data set	Evaluation	Arts	Problem	Deliverable
Kiran Reddy	various sources including	He do not use evaluations	varieties of clinical diagnosis support software systems available in the market	They could not make big impact on healthcare or clinicians	Developed clinical diagnosis software systems
Boris and Milan	Proposed hospital data	Proposed popular evaluations	Manual tasks	The elimination of manual tasks and easier extraction of data directly from electronic records	Electronic system of medical records to generated some tasks
Shree *et al*. (2014)	Questioners	Inter-RAI	Humans processed	Large amount of data cannot be processed by humans in a short time to make diagnosis	Analyse how the data mining can be used in health sector and discusses
Niaksu (2015)	36 AV and 35 LVOT echocardiography images	Similarity measure	CRISP-MED-DM	The application of data mining in healthcare and medicine. When applying data mining in medicine, additional problems such as varied information representation formats, semantic interoperability and patient privacy have to be resolved	Developing the supporting diagnosing models and medical data processing methods

Use of CDSS should be looked as one way of effectively integrating clinical practice guidelines into medical practice. Kwok *et al*. (2009), maintain that use of CDSS has facilitated clinician's adherence to clinical practice guidelines, thereby improving patient outcomes. Incorporation of clinical practice guidelines into computer-based decision support systems has been enabled by:

- Computers ability to store, search and sought large amounts of data rapidly
- Ever expanding knowledge, access and use of computers

By making it easy for programmers to develop rule-based or core-based reasoning in relation to the advices emanating from the CDSS as well as its demand for clinical practice guidelines, CDSS has become an easy incorporative tool. The framework in which encoded programming rules are encoded and used in the development of CDSS is provided by the encoded rules in the clinical practice guidelines. There is strict adherence to asthma clinical guidelines, improved clinical documentation and discharge in management plans for asthma management as noted by Kwok *et al.* (2009) this is all due use of integrated and dynamic Electronic Decision Support System (EDSS).

Previous works on CDSS and health care: Kiran Reddy has studies varieties of clinical diagnosis support software systems available in the market and his findings revealed that the software can impact on the healthcare. Based on that finding he suggests some guidelines to developed clinical diagnosis. Milovic and Milan (2012) used the data mining as classifiers to electronic system of medical records to perform some tasks. Their finding revealed that their system can deliver high performance comparing with manual tasks. Shree *et al*. (2014) analyzed how the data mining can be used in health sector and discusses the importance of resident assessment instrument so that they evaluate their suggestion as questioners and they proof how much data mining can support healthcare. Naiksu (2015) developed the support diagnosing models and medical data processing methods, his methods proposed as image classifier get high performance which leads to support healthcare as CDSS. Table 2 introduce a brief detail of clinical decision support system. Based on the analysis of past works on CDSS and healthcare, the following can be synthesized:

- Most of them focus on how to propose the technologies for support healthcare
- Some of them suggest data mining as classifiers to support health care
- All the researchers proposed data mining to that aims but information retrieval need to generate some of knowledge of clinical which is difficult to find it based on use data mining which are not use the ranking in their function as well as information retrieval method
- The Information Retrieval can play a vital role to retrieve the clinical data need to analysis based on some criteria need to decisions makers

CONCLUSION

This study started by providing an outline of some of the research already done while pointing out limitations where necessary for STM and signal detection. It presented an outline of what is implied by signal detection by regulatory and intelligence agencies with special emphasis on health care industry. More detail on medical device signal detection which forms the case study addressed by the research. Some of the key differences between drugs in the form of chemicals and pharmaceuticals with respect to devices have been highlighted. This should help in understanding and appreciating the need for specific approaches for signal detection in each domain and its sub-domains. Also, having introduced the two key concepts of this thesis, namely signal detection and STM, the next section provides an overview of the research questions addressed and contributions made. In addition, having outlined some of the relevant work in the area of STM and signal detection. The extracting structured data from unstructured data need to apply several techniques and procedure such as NLP. Doctor notes and documents is unstructured data contains information such as proposed diseased, patient risk, treatments, addendum text, etc. Finally, utilize structure data and unstructured data to enhance the proposed predication framework to find the predicate disease and patient risk.

REFERENCES

Adriaans, P. and D. Zantinge, 1996. Data Mining. Addison-Wesley Company, Harlow, England,.

Avram, H.D., 1968. The MARC pilot project: Final report. Library of Congress, Washington, D.C., USA.

Bailey, J., F. Bry, T. Furche and S. Schaffert, 2005. Web and semantic web query languages: A survey. Proceedings of the 1st International Conference on Reasoning Web Schaffert, July 25-29, 2005, Springer, Berlin, Germany, ISBN:978-3-540-27828-3, pp: 35-133.

Belkin, M. and P. Niyogi, 2004. Semi-supervised learning on Riemannian manifolds. Mach. Learn., 56: 209-239.

Berner, E.S., D.E. Detmer and D. Simborg, 2005. Will the wave finally break? A brief view of the adoption of electronic medical records in the United States. J. Am. Med. Inf. Assoc., 12: 3-7.

Berners-Lee, T., J. Hendler and O. Lassila, 2001. The semantic web. Sci. Am., 284: 34-43.

Breuker, J. and R. Hoekstra, 2004. Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law proceedings. LLB Thesis, University of Amsterdam, Amsterdam, Netherlands.

Buckingham, C., 2002. Psychological cue use and implications for a clinical decision support system. Med. Inf. Internet, 27: 237-251.

Burkle, T., E. Ammenwerth, H.U. Prokosch and J. Dudeck, 2001. Evaluation of clinical information systems: What can be evaluated and what cannot?. J. Eval. Clin. Pract., 7: 373-385.

Chen, S.Y. and X. Liu, 2004. The contribution of data mining to information science. J. Inf. Sci., 30: 550-558.

Cleverdon, C., 1970. Evaluation tests of information retrieval systems. J. Doc., 26: 55-67.

Dhar, V. and A. Tuzhilin 1993. Abstract-driven pattern discovery in databases. Trans. Knowl. Data Eng., 5: 926-938.

Donzelli, P., 2006. Decision support system for software project management. IEEE. Software, 23: 67-75.

Frawley, W.J., G.Piatetsky-Shapiro and C.J. Matheus, 1992. Knowledge discovery in databases: An overview. AI Magazine, 13: 57-70.

GHRK., 2015. Developing reliable clinical diagnosis support system. General Hospital Reddy Kiran, Boston, Massachusetts.

Gatterbauer, W., P. Bohunsky, M. Herzog, B. Krupl and B. Pollak, 2007. Towards domain-independent information extraction from web tables. Proceedings of the 16th International Conference on World Wide Web, May 08-12, 2007, ACM, New York, USA., ISBN:978-1-59593-654-7, pp: 71-80.

Gilbert, M. and T. Friedman, 2006. The New Data Integration Frontier: Unifying Structured and Unstructured Data. Gartner Research, Stamford, Connecticut, USA.

Han, J. and K.C. Chang, 2002. Data mining for web intelligence. Comput., 35: 64-70.

Horrocks, I., P.F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof and M. Dean, 2004. SWRL: A semantic web rule language combining OWL and RuleML. W3C Member Submission. http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/.

Hotho, A., R. Jäschke, C. Schmitz and G. Stumme, 2006. Information retrieval in folksonomies: Search and ranking. In: Proceedings of the 3rd International Conference on European Semantic Web, June 11-14, 2006, Springer, Berlin, Germany, pp: 411-426.

Iqbal, A.M., M. Shepherd and S.S.R. Abidi, 2011. An ontology-based electronic medical record for chronic disease management. Proceedings of the 2011 44th Hawaii International Conference on System Sciences, January 4-7, 2011, IEEE, Halifax, Canada, ISBN:978-1-4244-9618-1, pp: 1-10.

Jansen, B.J. and A. Spink, 2006. How are we searching the world wide web? A comparison of nine search engine transaction logs. Inform. Process. Manage., 42: 248-263.

Jaschke, R., A. Hotho, C. Schmitz, B. Ganter and G. Stumme, 2008. Discovering shared conceptualizations in folksonomies. Semant. Sci. Serv. Agents World Wide Web, 6: 38-53.

Kawamoto, K., C.A. Houlihan, E.A. Balas and D.F. Lobach, 2005. Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. BMJ., 330: 765-765.

Kolari, P. and A. Joshi, 2004. Web mining: Research and practice. Comput. Sci. Eng., 6: 49-53.

Kotze, B. and B. Brdaroska, 2004. Clinical decision support systems in psychiatry in the information age. Australas. Psychiatry, 12: 361-364.

Kuo, K.L. and C.S. Fuh, 2011. A rule-based clinical decision model to support interpretation of multiple data in health examinations. J. Med. Syst., 35: 1359-1373.

Kuperman, G.J., A. Bobb, T.H. Payne, A.J. Avery and T.K. Gandhi *et al*., 2007. Medication-related clinical decision support in computerized provider order entry systems: A review. J. Am. Med. Inform. Assoc., 14: 29-40.

Kwok, R., M. Dinh, D. Dinh and M. Chu, 2009. Improving adherence to asthma clinical guidelines and discharge documentation from emergency departments: Implementation of a dynamic and integrated electronic decision support system. Emergency Med. Australas., 21: 31-37.

MedDRA., 2009. Medical dictionary for regulatory activities. MEDDRA, Manchester, UK. http://www.meddramsso.com/index.asp.

Manning, C.D., P. Raghavan and H. Schutze, 2008. An Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK., ISBN-13: 9780521865715, pp: 482.

Marcu, D., 1999. Discourse Trees are Good Indicators of Importance in Text. In: Advances in Automatic Text Summarization, Inderjeet, M. and T.M. Mark (Eds.). MIT Press, Cambridge, Massachusetts, USA., pp: 123-136.

Matheus, C.J., P.K Chan and S.G. Piatetsky, 1993. Systems for knowledge discovery in databases. IEEE. Trans. Knowl. Data Eng., 5: 903-913.

Matsuo, Y. and M. Ishizuka, 2004. Keyword extraction from a single document using word co-occurrence statistical information. Int. J. Artif. Intell. Tools, 13: 157-169.

Maybury, M., 1999. Generating Summaries from Event Data. In: Advances in Automatic Text Summarization, Mani, I. and T.M. Mark (Eds.). MIT Press, Cambridge, Massachusetts, USA., pp: 265-281.

Meza, A.B., M. Nagarajan, L. Ding, A. Sheth and I.B. Arpinar *et al*., 2008. Scalable semantic analytics on social networks for addressing the problem of conflict of interest detection. ACM. Trans. Web (TWEB.), 2: 1-7.

Milovic, B. and M. Milan, 2012. Prediction and decision making in health care using data mining. Intl. J. Public Health Sci., 1: 69-78.

Mobasher, B., 2007. Kunstliche intelligenz. Spec. Issue Web Min., 3: 41-43.

Morris, A.H., G.M. Kasper and D.A. Adams, 1992. The effects and limitations of automated text condensing on reading comprehension performance. Inf. Syst. Res., 3: 17-35.

Neill, O.E.S., N.M. Dluhy and E. Chin, 2005. Modelling novice clinical reasoning for a computerized decision support system. J. Adv. Nurs., 49: 68-77.

Niaksu, O., 2015. Development and application of data mining methods in medical diagnostics and healthcare management. PhD Thesis, Vilnius University, Vilnius, Lithuania.

Pearson, S.A., A. Moxey, J. Robertson, I. Hains and M. Williamson *et al*., 2009. Do computerised clinical decision support systems for prescribing change practice? A systematic review of the literature (1990-2007). BMC. Health Services Res., 9: 154-167.

Pierrakos, D., G. Paliouras, C. Papatheodorou and C.D. Spyropoulos, 2003. Web usage mining as a tool for personalization: A survey. User Model. Adapted Interact., 13: 311-372.

Rath, G.J., A. Resnick and T.R. Savage, 1961. The formation of abstracts by the selection of sentences. Am. Documentation, 12: 139-141.

Rijsbergen, K.V., 1979. Information retrieval. University of Glasgow, Glasgow, Scotland. http://www.dcs.gla.ac.uk/Keith/Chapter.1/Ch.1.html.

Ronald, B.J., T. Khabaza, K. Willi, P.S. Gregory and S. Evangelos, 1997. Mining business databases. Commun. ACM., 39: 28-42.

Ronald, B.J., and T. Anand, 1996. Advances in Knowledge Discovery and Data Mining. In: Process of Knowledge Discovery in Databases, Fayyad, U.M. (Ed.). University of Michigan, Ann Arbor, Michigan, ISBN:9780262560979, pp: 37-57.

SIGIR., 2009. Special interest group on information retrieval. SIGIR, Paris, France. http://www.sigir.org/index.html.

Salton, G., A. Singhal, M. Mitra and C. Buckley, 1997. Automatic text structuring and summarization. Inf. Process. Manage., 33: 193-207.

Sarawagi, S., 2008. Information extraction. Foundations Trends Databases, 1: 261-377.

Shadbolt, N., W. Hall and T. Berners-Lee, 2006. The semantic web revisited. IEEE Intell. Syst., 21: 96-101.

Sheth, A., 2005. Enterprise applications of semantic web: The sweet spot of risk and compliance. Proceedings of the 1st IFIP Working Conference on Industrial Applications of Semantic Web, August 25-27, 2005, Springer, Jyvaskyla, Finland, pp: 47-62.

Sheth, A., C. Ramakrishnan and C. Thomas, 2015. Semantics for the semantic web: The implicit, the formal and the powerful. Intl. J. Semantic Web Inf. Syst. 1: 1-18.

Shree, S.D., A. Kanimozhi and G.N.K.S. Babu, 2014. Application of data mining techniques in health care industry. Int. J. Comput. Sci. Inform. Technol. Res., 2: 243-247.

Sieg, A., B. Mobasher and R. Burke, 2007. Web search personalization with ontological user profiles. Proceedings of the 6th ACM Conference on Information and Knowledge Management, November 06-10, 2007, ACM, New York, USA., ISBN:978-1-59593-803-9, pp: 525-534.

Smyth, P. and R.M. Goodman, 1992. An information theoretic approach to rule induction from databases. IEEE Trans. Knowl. Data Eng., 4: 301-316.

Srivastava, A. and M. Sahami, 2009. Text Mining: Classification, Clustering and Applications. 1st Edn., CRC Press, Boca Raton, Florida, USA., Pages: 279.

Srivastava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD Explorat., 1: 12-23.

Uthurusamy, R., 1996. From Data Mining to Knowledge Discovery: Current Challenges and Future Directions. In: Advances in knowledge Discovery and Data Mining, Fayyad, U.M., S. Padhraic and U. Ramasamy (Eds.). American Association for Artificial Intelligence, Menlo Park, California, USA., ISBN:0-262-56097-6, pp: 561-569.

Zaremba, M., M. Moran and T. Haselwanter, 2006. Applying semantic web services to virtual travel agency case study. Proceedings of the 3rd International Conference om European Semantic Web, June 11-14, 2006, ESWC, Budva, Montenegro, pp: 1-2.