A Review of Towered Big-Data Service Model for Biomedical Text-Mining Databases

Alshreef Abed
Department of Computer Science and Technology, Wuhan University of Technology Wuhan, China

Jingling Yuan
Department of Computer Science and Technology, Wuhan University of Technology Wuhan, China

Lin Li
Department of Computer Science and Technology, Wuhan University of Technology Wuhan, China

***Abstract*—The rapid growth of biomedical informatics has drawn increasing popularity and attention. The reason behind this are the advances in genomic, new molecular, biomedical approaches and various applications like protein identification, patient medical records, genome sequencing, medical imaging and a huge set of biomedical research data are being generated day to day. The increase of biomedical data consists of both structured and unstructured data. Subsequently, in a traditional database system (structured data), managing and extracting useful information from unstructured-biomedical data is a tedious job. Hence, mechanisms, tools, processes, and methods are necessary to apply on unstructured biomedical data (text) to get the useful business data. The fast development of these accumulations makes it progressively troublesome for people to get to the required information in an advantageous and viable way. Text mining can help us mine information and knowledge from a mountain of text, and is now widely applied in biomedical research. Text mining is not a new technology, but it has recently received spotlight attention due to the emergence of Big Data. The applications of text mining are diverse and span to multiple disciplines, ranging from biomedicine to legal, business intelligence and security. In this survey paper, the researcher identifies and discusses biomedical data (text) mining issues, and recommends a possible technique to cope with possible future growth.**

Keywords*—Big data; biomedical data; text mining; information retrieval; feature extraction

I. INTRODUCTION

Currently, the field of biomedical research is booming, a lot of biomedical knowledge is in unstructured form in the form of text file, and now the field has witnessed exponential trend increase; there is a need to solve the contradictions between massive growth of information and knowledge of text slowly and in a credible manner to identify useful patterns in the text which is still a challenge. In recent years, biomedical text mining technology which is one branch of an efficient automatic access to new exploration-related knowledge has witnessed significant progress [1].

Biomedical information is increasing rapidly in size, and helpful outcomes come into sight daily in research publications. However, automatically taking out useful information from such a stupendous quantity of documents is a difficult task because these documents are unstructured and are revealed in natural language. To enable data mining and knowledge discovery techniques, documents should be in the structured format [2].

The problem faced by the biological researchers is on how to effectively find out the useful and needed documents in such an information-overload environment. Traditional manual retrieval method is impractical. Furthermore, online biological information exists in a combination of different forms, including structured, semi-structured and unstructured forms [3]. It is impossible to keep abreast of all developments. Computational methodologies increasingly become important for research [4]. Text mining techniques which involve the process of information retrieval, information extraction and data mining provide a means of solving this by Ananiadou et al. [5].

The volume of published knowledge in the biomedical region is produced at an unprecedented pace. Biomedical researchers need to explore the big amount of scientific publications to examine findings related to certain biomedical entities such as proteins, diseases, etc. In the biomedical domain, simple keyword based matching may not be adequate because biomedical entities have synonyms and ambivalent names. Biomedical text mining relates to automatically identifying biomedical entities from a given text and to associate them to their correlating entries in knowledge bases. Biomedical text mining enables researchers to recognize useful information more efficiently. Two elementary functions of information extraction are Named entity recognition and Relation extraction. Named entity recognition deals with detecting the name of entities. Relation extraction refers to uncovering the semantic relations between entities [2].

The number of articles that are added to the literature databases is growing at a fast pace [6]. Retrieval of relevant information from literature databases and combining this information with experimental output is time-consuming and requires careful selection of keywords and drafting of queries. This is often a biased and time-consuming process, resulting in incomplete search results, preventing the realization of the full potential that these databases can offer [7]. Automated processing and analysis of text (referred to as Text Mining (TM)) can assist researchers in evaluating scientific literature. Nowadays, TM is applied to answer many different research questions, ranging from the discovery of drug targets and biomarkers from high-throughput experiments [8]–[13] to drug repositioning, the creation of a state-of-the-art overview of a certain disease or therapeutic area and for the creation of

domain-specific databases [14], [15]. Due to the heterogeneous nature of written resources, the automated extraction of relevant biological knowledge is not trivial. As a consequence, TM has evolved into a sophisticated and specialized field in the biomedical sciences where text processing and machine learning techniques are combined with mining of biological pathways and gene expression databases.

The rest of the paper is organized as follows: Section II has discussed the purpose and overview of text mining, the significance of biomedical text mining, task, models and methods used. It presents the definitions of the concepts explored in this study. Section III discussed the previous study which is related to text mining, biomedical text mining, biomedical data with feature extraction approach and biomedical data mapping technique. It critically evaluates methodologies that were available at the time of this research. Section IV discussed research methods used by the reviewed articles. Analysis and discussion are covered in Section V wherein the contextual settings of the reviewed articles are examined. The study findings, conclusion, and recommendation for further research are discussed in Section VI.

II. Purpose of the Study

Due to the rapid growth of data and text, information extraction is a difficult task, especially in biomedical databases [16]. Additionally, the diversity, complexity and volume of the information that need to be mined present challenges in the biomedical domain impacts the biomedical discovery process, stifling researchers working towards novel hypotheses to address critical questions [17]. Subsequently, such information extraction depends on the flexible formulation and common methods for heterogeneous data integration and indefinable discovery of knowledge sources that highly depend on a particular scientific question. It truly influences the effective techniques of storage, extraction and permitting sympathetic of the molecular substructures of biological processes. For this purpose, this paper briefly overviews the major challenges in these areas and discusses the recommendations and implications of this research.

A. Overview of Text Mining

Text mining or text analytics is an umbrella term describing a range of techniques that seek to extract useful information from document collections through the identification and exploration of interesting patterns in the unstructured textual data of various types of documents - such as books, web pages, emails, reports or product descriptions. A more formal definition restricts text mining to the creation of new, nonobvious information (such as patterns, trends or relationships) from a collection of textual documents. Typical text mining tasks include activities of search engines, such as assigning texts to one or more categories (text categorization), grouping similar texts together (text clustering), finding the subject of discussions (concept/entity extraction), finding the tone of a text (sentiment analysis), summarizing documents, and learning relations between entities described in a text (entity relation modeling) [18].

The utilization of the web has expanded the obtainability and access to publications that are the foremost in various cases on data over-burdening [19]. Specifically, biomedical data sets have increased rapidly in large computerized stores [20]. Therefore, searching and organizing these data is always considered as time-consuming and cost ineffective. For example, in the digital library, the MEDLINE is a fast-growing biomedical database, and the information within this data set is stored in text form. Recently, it has stored more than 18 million indexed articles. So the usability and obtainability of this data have become precarious to the researchers and students who are working in the biomedical area [21]. The quick advancement of these accumulations makes it progressively troublesome for analysts to get the required information in a helpful and proficient way.

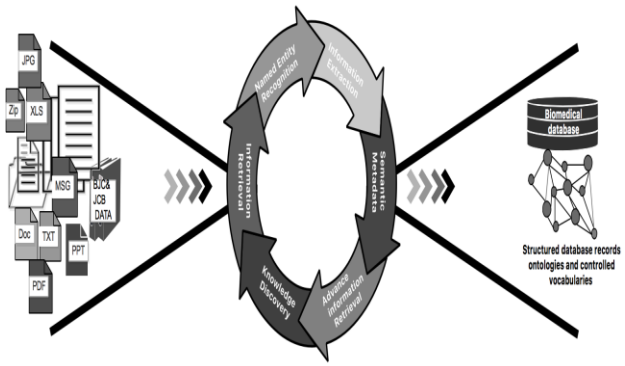

Fig. 1. Text mining eco-system for biomedical data.

Subsequently, the relationship amongst various medical conceptions from medical collected works is a foremost issue for many biological researchers. But, the data gathering level confines the incorporation into choice data frameworks for two reasons. Firstly, it requires more time from therapeutic specialists to create and maintain the learning base. Secondly, sharing and reusing the approved information base is troublesome due to the absence of clearness [22]. The goal is to obtain consistent data, and its extraction is one of the essential objectives of biomedical text mining groups [23]. The term text mining is used when exploring the objects stored in an unstructured data set and offers the capability towards managing and analyzing [24] the large sets of data in an effective manner [25]. While also realizing the significant relationships or correlations amongst variables in the huge dataset [26]. The smart data retrieval system is essential in operating non-standardized entries in order to access the data [27]. Subsequently, there is a robust need to create strategies for programmed extraction of pertinent data from the collected works, which is composed in natural language [28]. Therefore, in this study, the text mining method is towards discovering additional useful information in a more effective way. “Fig. 1” shows the overview of text mining process from the biomedical database.

B. Text Mining

Text mining refers to the automated extraction of knowledge and information from the text by revealing relationships and patterns that are present, but not obvious, in a document collection. Subsequently, it uses a wide range of

utilities including information extraction, text clustering, sentiment analysis, text categorization, document summarization, named entity recognition and question answering and the seven interdisciplinary fields based on computational linguistics: artificial intelligence, data mining, natural language processing and information retrieval [29].

The goal of text mining is to derive implicit knowledge that hides in unstructured text and present in an explicit form. This generally has four phases: information retrieval, information extraction, knowledge discovery, and hypothesis generation. Information retrieval systems aim to get desired text on a certain topic; information extraction systems are used to extract predefined types of information such as relation extraction; knowledge discovery systems help us to extract novel knowledge from the text; hypothesis generation systems infer unknown biomedical facts based on text, as shown in "Fig. 2". Thus, the general tasks of biomedical text mining include information retrieval, named entity recognition and relation extraction, knowledge discovery and hypothesis generation [30].

The text mining-associated text document and database models [31] are identified as:

- Information recovered from web archives with a population of data set patterns.
- Disclosure of data presented in the text as well as the capacity for XML or social groups.
- Incorporation and questioning of content information after it has been stored in databases.
- Deduplication of a data set through utilizing standard information mining strategies like clustering.

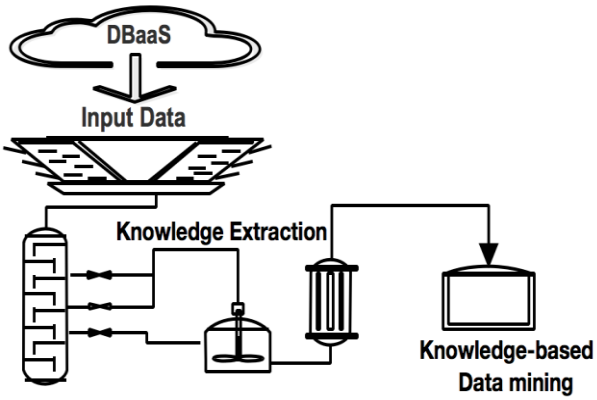

Fig. 2. BDaaS Utilization Model for knowledge extraction.

C. Models and Methods Used in Text Mining

To solve text mining issues, previously many researchers have suggested new methods for relevant information retrieval according to a user's requirement [32]. Based on the information retrieval process, there are four methods: term, phrase, pattern taxonomy and the concept-based method.

D. Biomedical Literature Mining

The era of applying text mining approaches to biology and biomedical fields came into existence in 1999. It was first applied to the biomedical domain for gene expression profiling [33], as well as the extraction and visualization of protein-protein interaction [34]. It emerged as a hybrid discipline from the edges of three major fields, namely, bioinformatics, information science, and computational linguistics. Biomedical literature mining is concerned with the identification and extraction of biomedical concepts (e.g., genes, proteins, DNA/RNA, cells, and cell types) and their functional relationships [35]. The major tasks include 1) document retrieval and prioritization (gathering and prioritizing the relevant documents); 2) information extraction (extracting information of interest from the retrieved document); 3) knowledge discovery (discovering new biological event or relationship among the biomedical concepts); and 4) knowledge summarization (summarizing the knowledge available across the documents). A brief description of the biomedical literature mining tasks is listed as follows.

E. Biomedical Text Mining Tasks

Document Retrieval: The process of extracting relevant documents from a large collection is called document retrieval or information retrieval [36]. The two basic strategies applied are query-based and document-based retrieval. In query-based retrieval, documents matching with the user specified query are retrieved. In document-based retrieval, a ranked list of documents similar to a document of interest is retrieved.

Document Prioritization: The retrieved documents are usually prioritized to get the most relevant document. Many biomedical document retrieval systems achieve prioritization based on certain parameters including journal-related metrics (e.g., impact factor, citation count) [37] and MeSH index [38], [39] for biomedical articles. The similarity between the documents is estimated with various similarity measurements (e.g., Jaccard similarity, cosine similarity) [40].

Information Extraction: This task aims to extract and present the information in a structured format. Concept extraction and relation/event extraction are the two major components of information extraction [41], [42]. While concept extraction automatically identifies the biomedical concepts present in the articles, relation/event extraction is used to predict the relationship or biological event (e.g., phosphorylation) between the concepts [43], [44].

Knowledge Discovery: It is a nontrivial process to discover novel and potentially useful biological information from the structured text obtained from information extraction. Knowledge discovery uses techniques from a wide range of disciplines such as artificial intelligence, machine learning, pattern recognition, data mining, and statistics [45]. Both information extraction and knowledge discovery find their application in database curation [46], [47] and pathway construction [48], [49].

Knowledge Summarization: The purpose of knowledge summarization is to generate information for a given topic from one or multiple documents. The approach aims to reduce the source text to express the most important key points through content reduction selection and/or generalization [50]. Although knowledge summarization helps to manage the

information overload, state of the art is still open to research to develop more sophisticated approaches that increase the likelihood of identifying the information.

Hypothesis Generation: An important task of text mining is hypothesis generation to predict unknown biomedical facts from biomedical articles. These hypotheses are useful in designing experiments or explaining existing experimental results [51].

Text mining for biomedical literature often involves two major steps. a. First, it must identify biomedical entities and concepts of interests from free text using natural language processing techniques. Many text mining algorithms have been applied to this problem. For example, some morphological clues to recognize the heartache like obesity, blood pressure. b. And then, the converted information is extracted from the text or unstructured documents into the standardized data set, and data mining is applied to the data source. "Fig. 3", shows the typical text Mining Process.

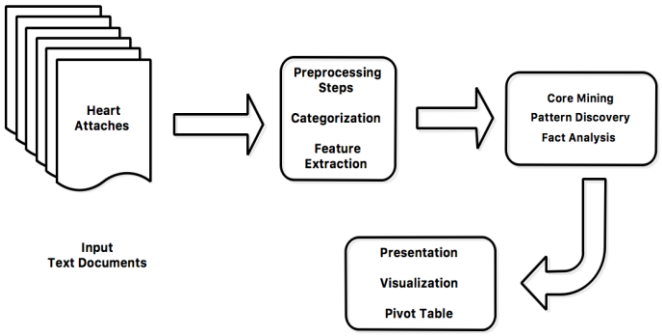

Fig. 3. Typical text mining process.

III. Related Works

This section provides a brief summary of text mining followed by most recent studies that have been conducted with regard to text mining in the field of biomedical research texts.

A. Biomedical Text Mining Review

Information extraction or (IE) covers the recognition of biomedical identities in biomedicine for extracting information pertaining to a disease, its treatment and its proteins and extracting the association (s) between these identities. The association between two different entities is extracted through different methods. Previous studies related to the extraction of useful information from databases are discussed as follows.

Tan and Lambri [52] suggested a framework for the purpose of selecting a suitable ontology for a specific application for biomedical text mining. Subsequently, an experiment was put forth for biomedical ontology in the context of a gene normalization system by utilizing the framework. Inside the references of the framework, the results of the assessment directed us to a comparatively firm option of ontology for our module. Furthermore, the researchers have planned to evaluate this framework with more applications and ontologies.

Qi et al. [53] conducted a survey about text mining in the realm of bioinformatics with a focus on the application of text mining. During the course of this study, the primary research focus of text mining in bioinformatics was supported through exhaustive examples. This study, in particular, matched the requirement for a state-of-the-art area of text mining in bioinformatics, primarily due to the swift advancement in both the fields of text mining and bioinformatics. The full ability of this area has remained underutilized.

A framework of a probabilistic combination nature for the purpose of precisely linking citation information with the content-based information retrieval weighting model is suggested by Yin et al. [54]. Through a case study, they were able to observe the model of linking information that was available in the citation graph. Extensive parameter tuning can possibly be done away with through this framework. However, this basically tested the suggested combination framework in the context of a biomedical literature corpus; they researchers of the opinion that the basic premise of their paper could be absorbed for literature retrieval in other areas.

Also, Tari et al. [55] explained the Gene Properties Mining Portal as one that permits retrieving gene-centric data from literature through text mining. This portal acts as a node for scientists to discern vital relationships in an effective and efficient manner from literature. But, the precision of the relations that were extracted were influenced by many issues, for instance, by limiting the methods of extraction in addition to the quality of the sources.

Bchir and Ben Abdessalem Karaa [56] proposed a method for the purpose of extracting relations between disease and drug. To begin with, they deployed Natural Language Processing methods for prepossessing abstracts. Later, features were extracted in the form of a set of preprocessed abstracts. To conclude, a disease drug association was extracted through the utilization of a disease-drug Association through a machine learning classifier. But, they ended up extracting associations among drugs and diseases, with a need to additionally extract other relations among other concepts.

Mala and Lobiyal [57] relied on ontologies for extracting concepts and offered an algorithm to locate and identify concept-based clusters. They then went on to label semantic weightage for all terms for every document. They resorted to using a tagging mechanism commonly known as POS (Part of Speech Tagging) to locate nouns in addition to utilizing Rapid Miner for text mining method such as text processing. The use of medical ontologies can also enhance the outcome of this method.

Roth et al. [58] had an objective to extract from biomedical literature information that was supportive of Protein-Protein Interactions (PPIs) that were of a predictable nature. The demonstrated results of the relation extraction show that an f-score of 0.88 was witnessed on the HPRD50 corpus, and the similarities in semantics that were calculated with an angular distance were also proved to be statistically considerable.

Jimeno Yepes and Berlanga [59] suggested an innovative technique to create word-concept probabilities from knowledge bases (KBs), which could then act as a foundation for numerous text mining jobs. The findings indicated that this

technique secured enhanced accuracy when compared with other state-of-the-art methods, particularly in the context of the MSH WSD data set. However, the present refinement implementation does not attempt to recognize or locate new synonyms for prevailing concepts; rather, it only attempts to tag the data by quantifying the frequency of usage within a specific concept. It does not attempt to locate or unearth new concepts that are not found in the knowledge base. It would be worthwhile to evaluate information extraction methods in order to locate and recognize new synonyms [60] for concepts that are both prevailing and new.

Meaney et al. [61] debated the changes and patterns in the use of techniques in statistics and epidemiology found in medical literature from the last 20 years. Furthermore, the research proposed a method to improve the text-mining approach and incorporated advanced retrieval techniques to gauge the ratio of articles. This method referred to a specific technique that was statistical or epidemiological : this is where further study needs to be undertaken by the team [62], [63]. A statistical machine translation approach [64] and a Bayesian information extraction network for the Medline abstract [65] are used in the proposed text mining system to deal with this problem.

B. Text Mining Methods

Berardi et al. suggested a framework that assists biologists in automatically extracting information from machine-readable documents or texts. These extraction models were later used on unobserved texts in automatic mode. They reported an application that was a real world dataset compiled by publications, which were in turn chosen to aid biologists in annotating an HmtDB database.

An extension of the Okapi retrieval system that was effectual for mining biomedical text has been suggested by [66]. This lead to two advantages in the system when compared with other models. First, this method is uncomplicated to implement and is not tagged to any domain. Secondly, it has proven its competence and effectiveness in TREC Genomics experiments. Despite the fact that the suggested extension is effective in discerning the subtle variations in the verbiage of a biological entity, it does not offer any comprehensive solution to encompass all variations in that lexicon. But, this algorithm cannot serve to be its identifying factor. Henceforth, such variations would be discerned through a query expansion algorithm.

A text summarization algorithm that used scientific literature in biomedicine which discerns the focal topic of biomarker cancer discoveries and all information in the literature that is deemed vital was suggested by Islam et al. [67]. The purpose of this study, however, needs to be directed towards extracting more specialized information on protein structure and image data mining. Also, the system needs to be optimized to handle large loads with quick response and must support multiple databases.

Liu et al. [68] introduced a study regarding names in the Bio Thesaurus, which was, in turn, collated from multiple databases present in a free-text by utilizing a data set that was automatically created from cross-referencing in the UniProtKB. The findings proved that using different resources to put together synonyms for biological identities can result in optimized coverage for nomenclature present in the text while utilizing matching that is able to be adjusted. But, flexible matching creates more ambiguous situations for English words that are common. This results in the need to narrow down the confusion between common English vocabulary and biological identity nomenclature through corpus-based word sense disambiguation.

Leroy et al. [69] created text mining tools that indicate co-occurrence relations among concepts. Engaging subsets of relations are mined through statistical measures. In addition, the researchers proved the manner in which these relations were directed had an effect on the amount of interest. To summarise, the numerous relations and their assistants were quantified. The differences in direction had a remarkable effect on the number of relations, and it also included the firm support of different types of graphs. The consequences of directionality on bigger graphs were not considered, however.

Salahuddin and Rahman [70] attempted to analyze and collate biomedical data from hypertext documents by utilizing text mining methods with the assistance of biomedical ontology. The matching and layout of the biomedical entity from the Metathesaurus were performed through a query on a keyword. However, this study focused on data in documents alone. Documents contain both textual information and visual imagery, and hence, there is a need to take into consideration the relevance of images in medical documents and attempt to give ranking to the documents based on the combined textual and image content.

Ronquillo et al. [23] proposed a program for automatic categorization of biomedical text. The results achieved pertaining to performance and execution timing are more positive when compared to the results obtained earlier and used in Weka, and what is known as the baseline system. This system has certain limitations, however, especially when it needs to show the difference between texts regarding hearing loss classified as syndromic and nonsyndromic. For the purpose of improving categorization, this method will be used to locate and indicate symptoms and genes that are related to both types of hearing loss.

Hou et al. [71] proposed two options to help in directing the relation between genes and diseases (a) utilizing proximity relationship among genes and diseases, and (b) using GO terms that are prevalent among genes and diseases for the purpose of comparing similarity. Experiments demonstrate that relations using GO terms function better than utilizing word proximity. This proves that GO terms serve as a better option for good gene-disease association. But, this only concentrated on the aspect of the relationship. Additionally, there is a need to focus on applying prediction of gene-disease relationships apart from the OMIM database.

A text mining technique that extracts numerous entities from biomedical text had proposed by Javed and Afzal [72] where candidate terms are discerned through the application of an algorithm known as the C-Value. These candidate terms and prevalent terms used in Seed/Ontology are labeled in the corpus. By resorting to the assessment of profiles that were

lexical and contextual in the comparison between candidate terms and the prevailing Seed/Ontological Terms, it was possible for them to discover novel ideas and assess them. This study required an enhancement to the categorization of included measures that resembled each other, such as Word Net to discern the link between two terms.

The summary of text mining methods is presented in Table 1 where each method is briefly identified and then analyzed. Other methods which include knowledge extraction and data mapping techniques have been classified in the next sections along with they're summary in Table 2. The evaluation includes some major limitations in each method which need to be recovered for potential researches and experiments. The summary of previous studies related to biomedical data mapping techniques are discussed in Table 3. Finally, the recommendation and implication of this research are discussed in Table 4.

TABLE I. SUMMARY OF PREVIOUS STUDIES RELATED TO TEXT MINING APPROACH

Ref.	Method	Results	Advantage	Limitation
Berardi et al. [73]	Text extraction rule	Automatic information extraction	Fast and simple	Extract abbreviations and acronyms.
Ming Zhong and Xiangji [66]	Okapi retrieval approach	An effective TREC Genomics experiments	Simple implementation	Cannot serve to be identifying factors
Islam et al. [67]	text summarization algorithm	Discerns the focal topic of biomarker cancer discovery	Simple implementation	Does not support multiple and public databases
Liu et al. [68]	Text classification approach	Nomenclature text optimization	Flexible in text matching	High confusion between common English and biological vocabulary
Leroy et al. [69]	Text mining tool	Multiple text mapping	Secure and support different types of graphs	Does not support bigger graphs
Salahuddin and Rahman [70]	Ontology-based text mining	Documents identification	Effectiveness for fewer parameters.	document data only
Ronquillo et al. [23]	Text classification approach	Small data sets classification	High Performance and execution time	Does not identify some symptoms and genes
Hou et al. [71]	Text mining approach	Utilizing word proximity using GO terms function	Very secure	Does not support multiple and public databases
Javed and Afzal [72]	Text mining methodologies	Automatic biomedical text extraction	High efficiency.	Does not enhance similarity measures

C. Knowledge Extraction Methods

Jahiruddin et al. [74] introduced an innovative Biomedical Knowledge Extraction and Visualization framework (BioKEV) which is used to discern and isolate vital information components from biomedical text documents. The method of information extraction was based on NLP or Natural Language Processing methods and analysis that were also based on semantics. Additionally, it was suggested that a ranking system for documents needed to be in place to refer to retrieved documents in the same relevant order as queried by the user. Furthermore, they improved the format of the query processing module to render it compatible with a high degree of efficiency when searching biomedical queries of a complex nature.

Sharma et al. [75] concentrated on discovering the task and extracting relations that were witnessed between certain bioentities, like green tea and cancer of the breast. Additionally, a verb-centric algorithm was suggested to be put in place. This system locates and extracts the primary verb(s) observed in a sentence; hence, there is no requirement for a separate set of rules or patterns. The algorithm was assessed in numerous datasets and observed an average of F as 0.905, which is considerably more than what had been previously achieved.

However, a framework called Feature Coupling Generalization (FCG) for the purpose of developing novel features from untagged data has been suggested by Li et al. [76]. This framework chooses Example-Distinguishing Features (EDFs) and Class-Distinguishing Features (CDFs) to recognize the gene entity name (NER), extract the protein-protein interaction (PPIE) and classify the gene ontology (GO). Additionally, the performance of baselines that are under supervision was improved by 7.8 %, 5.0 %, and 5.8 %, respectively, in all three tasks. But this study does not justify the reason for the workings of FCG and the reasons that determine EDFs' and CDFs' qualities.

Holzinger et al. [77] proposed a Sequence Memorizer Based Model (SMBM) that had its roots in what was known as the generative model to oversee its functioning. This method resorted to the utilization of the generative strategy in order to avoid the option of selecting work that was time-consuming. While ensuring the advantages of models that were generative in nature, the functionality of this technique can be compared to that of the Maxent model.

Holzinger et al. [77] offered a way to assess knowledge discernment of disease-disease relationships for rheumatic diseases. Also, they resorted to utilizing a Point wise Mutual Information (PMI) calculation to identify a relationship's strength. The output indicates concealed knowledge in articles

pertaining to rheumatic diseases that were indexed by MEDLINE, and which could be used by medical experts and researchers for the purpose of making medical decisions. This study also needs to concentrate on collecting the names of diseases, nomenclature/codes of diagnosis and treatments to observe the extent to which identification of diseases in the searched content can be improved through screening for diagnosis and treatment of such diseases.

Pereira et al. [78] developed an integrated approach for the reconstruction of Transcriptional Regulatory Networks (TRNs), which retrieve the relevant data from important biological databases and insert the result into a unique repository named KREN. Further, they integrated this into the Note software system, which allows some methods from the Biomedical Text Mining field, including algorithms for Named Entity Recognition (NER), extraction relationships between biological entities and extraction of all relevant terms from publication abstracts. Finally, this tool was extended to allow the reconstruction of TRN using scientific literature.

Landge and Rajeswari [79] conducted an overview of the comparative analysis of numerous techniques employed in determining the relation between chemical entities, and also reviewed the comparative analysis of numerous text mining methods. Further, they suggested to using a parallel approach to text mining towards minimizing the time needed by their method. Conventional algorithms can be parallelized and applied to mine and extract information and knowledge from a large data set.

TABLE II. SUMMARY OF PREVIOUS STUDIES RELATED TO KNOWLEDGE EXTRACTION METHOD

Ref.	Method	Results	Advantage	Limitation
Tangtulyangkul et al.[80]	Keyword mining scheme	External-source based knowledge accumulator	Reduce Information overload	clinical records only
Sharma et al. [75]	Verb-centric algorithm	Biomedical entities identification	Handled complex sentence	private data sets only
Liu et al. [68]	FCG framework	Supervised data learning utility	High supervised baselines performance	Does not confirm FCG, EDFs and CDFs qualities
Holzinger et al. [77]	Sequence memorizer Based Model	Natural language recognition	High performance.	Does not integrate sequence memorized into machine learning model.
Holzinger et al. [77]	Decision-making approach	Medical decision-making processes	High efficiency	Disease names only
Pereira et al. [78]	Transcriptional Regulatory Networks	Gene scientific corpora	Robust extraction	Does not validate the regulatory model.
Landge and Rajeswari [79]	Reviewed text mining approach towards chemical entities	Chemical data machine learning algorithms	High efficient for small samples	Does not use parallel approach

D. Biomedical's Data Mapping Techniques

Cano et al. [81] suggested an approach that was hybrid in nature for the purpose of mining or unearthing the vast knowledge that was accumulated in the scientific literature. This method has its foundations on the utilization of effectively mining text through tools that work in tandem with precise and collaborative human duration. To demonstrate the effectiveness of this method, this study requires quantification of the time that is reduced in performing tasks. This leads to an observable upgrade in the state of information regarding the remaining portion of the knowledge content and ensures that active learning techniques are put into use for assigning priority to the annotation process.

Yang and Dong [82] proposed a mapping-based approach by first mapping bio-entities to terms in an established ontology Medical Subject Headings (MeSH). Specifically, they present two approaches to mapping biomedical entities identified using the Unified Medical Language System Met thesaurus to MeSH terms. The first approach utilizes a special feature of the MetaMap algorithm, and the second employs an approximate phrase-based match to map entities directly to MeSH terms. These two approaches deliver comparable results with an accuracy of 72% and 75%, respectively, based on two evaluation datasets.

Mohammed and Nazeer [83] suggested an enhanced system of text mining that was focused on the method of matching patterns and heuristics that reduced space and increased the recall and accuracy. The system recalls, f-factor and precision were assessed through three metrics. The output of the experiments resulted in a recall of 98.68% and precision of 98.68%.The system has a drawback, though, in that it placed restrictions on the format of candidate acronym-definition pairs, which means that they needed to appear as either an acronym.

Ji et al. [84] created a Map Reduce algorithm to calculate the strength of association among two biomedical terms

witnessed in biomedical documents. Additionally, they evaluated if the algorithm was scalable by utilizing 3,610 documents retrieved from biomedical journals. Further, they demonstrated that this algorithm was linearly scalable when measured in the context of the number of nodes in a cluster. This method was only tested on a limited number of clusters with a reduced dataset, therefore leaving an additional need to assess the scalability of the algorithm in the context of the dataset size. Moreover, the algorithm needed to be enhanced in efficiency and accuracy.

TABLE III. SUMMARY OF PREVIOUS STUDIES RELATED TO BIOMEDICAL DATA MAPPING TECHNIQUES

Ref.	Method	Results	Advantage	Limitation
Cano et al. [81]	Hybrid mining approach	Scientific literature knowledge mining	High efficient.	Does not quantify the time reduction
Yang and Dong [82]	Mapping-based approach	Biomedical entities mapping and identification	High efficient for dimensional data	Low accuracy.
Mohammed and Nazeer [83]	Pattern matching method	Space reduction heuristics	High accuracy.	Low acronym definition
Ji et al. [84]	Map Reduce	Interestingness calculation	Linear scalability.	Tested on a small number clusters with low- scale datasets.

IV. MATERIALS AND METHODS

A comprehensive literature search of mining for text information in a medical database was conducted using a database such as Google search, Elsevier, IEEE, and Springer digital library and other literature sources. The searches were restricted to the years 2005 to 2016. The retrieved articles had text summarizations, like clinical, biomedical and medical summarization. This kind of search approach was applied to the web supplement. Additionally, searching through the collected database investigated the references of the included articles with an uncommon spotlight on past pertinent surveys. As seen in this review, the search retrieved a total of 112 potentially suitable articles to fulfill the inclusion criteria required for this review. Here, this study included unique examinations concentrated on the created and assessed text summarization techniques in the therapeutic areas, together with a summarization of electronic health records and biomedical collected works. “Fig. 4” shows the study flow diagram based on the PRISMA guidelines for reporting systematic reviews. However, the studies that met any of the following norms were excluded: images and multimedia summarization without a text summarization component, summarization of substance outside the biomedical area and non-English may have missed frameworks that compress content in different languages.

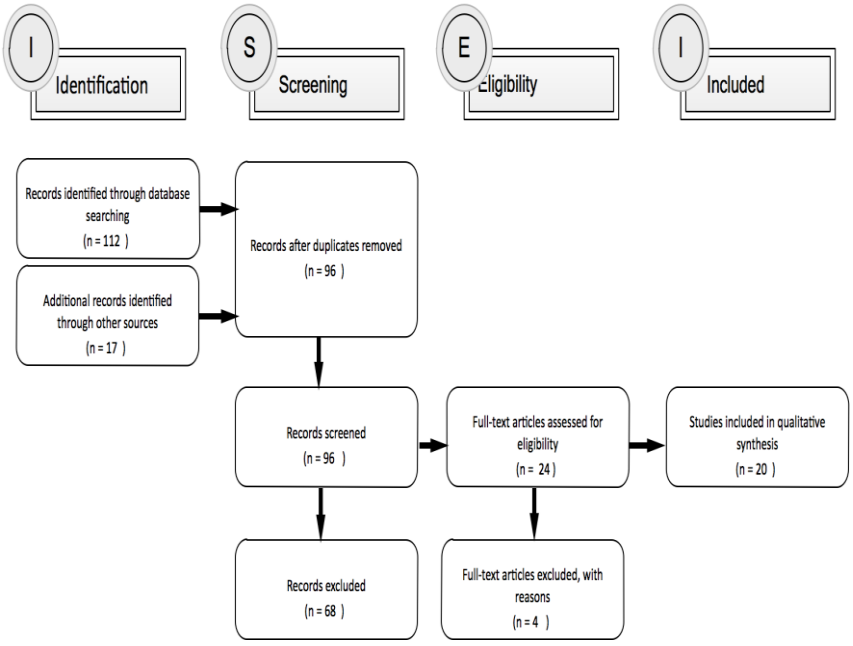

Fig. 4. Analysis of referred article.

V. RESULTS AND DISCUSSION

From the above review, the studies of Tan and Lambri, Salahuddin & Rahman, Yang & Dong, [52], [70], [82] have suggested a structure for choosing a suitable ontology for a specific biomedical text mining application. However, this study needs to focus more on handling the complex biomedical words. Additionally, this study only concentrated on document data. The archive is improved with both printed data and pictures. Therefore, this study needs to consider the significance of pictures in medicinal reports and attempt to rank archives both on the premise of printed data and picture data [67], [70]. Also, a focus on gaining higher mapping accuracy should be included.

Some studies focused on event extraction applications in biomedical text, such as [85], [86]. On the other hand, some concentration was held on security-based event extraction applications, such a sin [87]. Hogenboom et al. [88] reviewed event extraction methods from the text for decision support systems. They extracted the biologist's data automatically from the text, though some researchers had proposed [68], [72], [73] a mining based framework. However, this study needs to focus on reducing redundancy in data as well as improving the classification by adding similarity measures in order to extract the biomedical term proposed [71], [84] association rule mining approach. However, this method was tested on a fewer number of the clusters with low-scale datasets. also, there is a need for further refinement of this algorithm to improve the overall efficiency and accuracy. Landge and Rajeswari [79] reviewed the comparative analysis of various text mining methods to find an association amongst various chemical entities. They also discussed that text mining algorithms take a large amount of time [81] for the huge data sets. For this purpose, they suggested using the parallel approach of text mining towards minimizing the time over huge datasets.

Few of the previous studies have proposed a framework for identifying key information components from biomedical text documents, such as [55], [82] and [74], [84]. But, the precision of the extricated relations was influenced by various issues, for example, the impediment of the extraction designs and the nature of the sources. Liu et al. (2007) Aimed to study the Bio Thesaurus. Nonetheless, the examination of sets with names was neglected, which demonstrates that there are a few equivalent words in the content that were neglected to be caught in the Bio Thesaurus [59], [60]. [54], [56], [77], [80] all extracted keywords from biomedical records. However, this study only focused on the biomedical literature corpus and could be adapted to literature retrieval in other domains [89], [90]. Hou et al. Sharma et al. [71], [75] focused on mining associations amongst bioentities, like breast cancer and green tea. However, this study only concentrated on a specific data set. Furthermore, this would take a shot at the undertakings of categorization and relationship integration [23], [72], [76] which proposed an algorithm for categorizing biomedical text in an automatic manner. However, this system needs to improve the classification to achieve a higher performance [58]. A deep validation process in order to compare this method with the existing regulatory model is still necessary. Meaney et al. [61] recommended, enhancing the text mining technique towards a retrieval approach or highly sophisticated preprocessing [35],which could be utilized to evaluate the extent of articles referring to a given epidemiological or statistical technique [62], [63].

It is hence clear that biomedical text mining has great potential. However, that potential is yet unrealized. In the following years, text mining should be able to evaluation validate the results of analytical expression methods in identifying significant groupings of data [91]. Text mining researcher should co-operate with biology researchers in this interdisciplinary area. The following are some of the potential "New Frontiers" in biomedical text mining: Question-answering, Summarization, Mining data from full text (including figures and tables), User-driven systems, Evaluation [92] Now, this is an exciting time in biomedical text mining, full of promise.

VI. CONCLUSION

In this research, the researcher discussed and analyzed text mining techniques for biomedical data retrieving from the pool of documents on the web. From the literature, the biomedical record recovery strategy demonstrates about ideal results. In any case, the significance of a web report significantly relies on upon client's need that implies how much applicable the web record is as indicated by the client question. More effective text processing approach will provide an ideal result for the retrieval of the document from the web. Proficiency in processing mainly depends on time, but the calculation of time for ranking is a critical issue in implementation. As the web contains a large number of reports, offline estimation approach is not estimated effectively by any of existing approaches. Due to the complexity of Natural language processing, there is a broad examination in this field. So in future, it is necessary to concentrate more towards an effective method for capturing the meaning as well as relationships of words present in the document.

Based on the above review, future studies need to focus on:

- Cognitive aspects of text summarization which include visualization techniques, and evaluations of the impact of text summarization systems in work settings.
- Need to enable summarization corpora and reference standards to support the development of summarization tools in various applications.
- The increasing interest of users in efficiently retrieving and extracting relevant information, the need to keep up with new discoveries described in the literature or in biological databases, and the demands posed by the analysis of high-throughput experiments, are the underlying forces motivating the development of text-mining applications in molecular biology. Those technologies should provide the foundation for future knowledge-discovery tools able to identify previously undiscovered associations, something that will assist in the formulation of models of biological systems.
- Need to enable publicly available summarization corpora and reference standards to support the development of summarization tools.
- Need to improve the classification and mine the data towards getting higher performance

TABLE IV. RECOMMENDATIONS AND IMPLICATIONS OF THIS RESEARCH

Recommendation	Definition
Text Summarization	Further research is required in the subjective parts of text summarization, together with visualization method and the assessments towards the effect of text summarization systems in work settings.
Summarization Tool	Need to permit the reference standards and summarization corpora towards supporting the advancement of summarization tools in different applications.
Databases	The expanding enthusiasm of clients in productively recovering and separating important data, the need to stay aware of new disclosures depicted in the collected work or inorganic databases. Also, the requests postured by the investigation of high-throughput investigates, investigation are the basic powers spurring the improvement of text mining applications in sub-atomic science. Those innovations ought to provide an establishment of future information disclosure devices ready to distinguish already unfamiliar affiliations that will help with planning models of organic frameworks.
Higher Performance	Need to concentrate towards increasing the classification and mining the data for the attainment of higher performance.

ACKNOWLEDGMENT

This research project is supported by the ministry of higher education in Saudi Arabia and the National Natural Science Foundation of China (Grant No: 61303029,61602353), National Social Science Foundation of China (Grant No: 15BGL048), 863 Program (2015AA015403), Hubei Province Science and Technology Support Project(2015BAA072).

REFERENCE

[1] K. Meijing, Z. Le, and W. Jun, "Review of Research on Biological literature text mining," worldcomp, 2017. [Online]. Available: http://worldcomp-proceedings.com/proc/p2014/BIC3034.pdf. [Accessed: 21-Jun-2017].

[2] R. P. Saste and S. S. Patil, "Extraction of incremental information using query evaluator," in 2014 First International Conference on Networks & Soft Computing (ICNSC2014), 2014, pp. 324–328.

[3] M. Ghanem, A. Chortaras, Yike Guo, A. Rowe, and J. Ratcliffe, "A grid infrastructure for mixed bioinformatics data and text mining," in The 3rd ACS/IEEE International Conference onComputer Systems and Applications, 2005., 2005, pp. 185–188.

[4] G. C. Black and P. E. Stephan, "Bioinformatics: Recent Trends in Programs, Placements and Job Opportunities," Biotech, 2004. [Online]. Available: http://biotech35.tripod.com/private/ReportBioinfSloan_June04.pdf. [Accessed: 21-Jun-2017].

[5] S. Ananiadou, D. B. Kell, and J. Tsujii, "Text mining and its potential applications in systems biology.," Trends Biotechnol., vol. 24, no. 12, pp. 571–9, Dec. 2006.

[6] W. W. M. Fleuren and W. Alkema, "Application of text mining in the biomedical domain," Methods, vol. 74, pp. 97–106, Mar. 2015.

[7] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," Nat. Rev. Genet., vol. 7, no. 2, pp. 119–129, Feb. 2006.

[8] C. Plake, L. Royer, R. Winnenburg, J. Hakenberg, and M. Schroeder, "GoGene: gene annotation in the fast lane," Nucleic Acids Res., vol. 37, no. Web Server, pp. W300–W304, Jul. 2009.

[9] Z.-X. Huang, H.-Y. Tian, Z.-F. Hu, Y.-B. Zhou, J. Zhao, and K.-T. Yao, "GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords," BMC Bioinformatics, vol. 9, no. 1, p. 308, 2008.

[10] A. Kentsis, F. Monigatti, K. Dorff, F. Campagne, R. Bachur, and H. Steen, "Urine proteomics for profiling of human disease using high accuracy mass spectrometry," PROTEOMICS - Clin. Appl., vol. 3, no. 9, pp. 1052–1061, Sep. 2009.

[11] F. Al-Shahrour et al., "FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments," Nucleic Acids Res., vol. 35, no. suppl_2, pp. W91–W96, Jul. 2007.

[12] A. S. Haqqani, J. Kelly, E. Baumann, R. F. Haseloff, I. E. Blasig, and D. B. Stanimirovic, "Protein Markers of Ischemic Insult in Brain Endothelial Cells Identified Using 2D Gel Electrophoresis and ICAT-Based Quantitative Proteomics," J. Proteome Res., vol. 6, no. 1, pp. 226–239, Jan. 2007.

[13] W. W. M. Fleuren et al., "CoPub update: CoPub 5.0 a text mining system to answer biological questions," Nucleic Acids Res., vol. 39, no. suppl, pp. W450–W454, Jul. 2011.

[14] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki, "Integrated Text Mining and Chemoinformatics Analysis Associates Diet to Health Benefit at Molecular Level," PLoS Comput. Biol., vol. 10, no. 1, p. e1003432, Jan. 2014.

[15] D. G. Jamieson, M. Gerner, F. Sarafraz, G. Nenadic, and D. L. Robertson, "Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database," Database, vol. 2012, p. bas023-bas023, Apr. 2012.

[16] W. Hersh, "Evaluation of biomedical text-mining systems: Lessons learned from information retrieval," Brief. Bioinform., vol. 6, no. 4, pp. 344–356, 2005.

[17] G. Gonzalez et al., "Text and Data Mining For Biomedical Discovery," 2014.

[18] M. Truyens and P. Van Eecke, "Legal aspects of text mining," Comput. Law Secur. Rev., vol. 30, no. 2, pp. 153–170, Apr. 2014.

[19] G. Leroy et al., "Genescene: Biomedical Text And Data Mining," in Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL'03), 2003.

[20] S. Bleik, M. Song, A. Smalter, J. Huan, and G. Lushington, "CGM: A biomedical text categorization approach using concept graph mining," in 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop, 2009, pp. 38–43.

[21] PubMed, "http://www.ncbi.nlm.nih.gov/pubmed," 2016. .

[22] S. L. Achour, M. Dojat, C. Rieux, P. Bierling, and E. Lepage, "A UMLS-based Knowledge Acquisition Tool for Rule-based Clinical Decision Support System Development," J. Am. Med. Assoc., vol. 8, no. 4, pp. 351–360, 2001.

[23] F.-I. Ronquillo, C. P. de Celis, G. Sierra, I. da Cunha, and J.-M. Torres-Moreno, "Automatic classication of biomedical texts: experiments with a hearing loss corpus," in 4th International Conference on Biomedical Engineering and Informatics (BMEI), 2011, pp. 1674–1679.

[24] A. Browne, B. D. Hudson, D. C. Whitley, M. G. Ford, and P. Picton, "Biological data mining with neural networks: implementation and application of a flexible decision tree extraction algorithm to genomic problem domains," Neurocomputing, vol. 57, pp. 275–293, Mar. 2004.

[25] D. Luo et al., "Searching association rules of traditional Chinese medicine on Ligusticum wallichii by text mining," in 2013 IEEE International Conference on Bioinformatics and Biomedicine, 2013, pp. 162–167.

[26] Y. He et al., "Using association rules mining to explore pattern of Chinese medicinal formulae (prescription) in treating and preventing breast cancer recurrence and metastasis," J. Transl. Med., vol. 10, no. Suppl 1, p. S12, 2012.

[27] A. Li, Q. Zang, D. Sun, and M. Wang, "A text feature-based approach for literature mining of lncRNA?protein interactions," Neurocomputing, vol. 206, pp. 73–80, Sep. 2016.

[28] Y. Liu, M. Brandon, S. Navathe, R. Dingledine, and B. J. Ciliax, "Text mining functional keywords associated with genes," Stud. Health Technol. Inform., vol. 107, pp. 292–296, 2004.

[29] T. Magerman, B. Van Looy, B. Baesens, and K. Debackere, "Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents," katholieke Universiteit Leuven, 2011.

[30] F. Zhu et al., "Biomedical text mining and its applications in cancer research," J. Biomed. Inform., vol. 46, no. 2, pp. 200–211, Apr. 2013.

[31] A. Stavrianou, P. Andritsos, and N. Nicoloyannis, "Overview and Semantic Issues of Text Mining," SIGMOD Rec., vol. 36, no. 3, pp. 23–34, 2007.

[32] S. V. Gaikwad, A. Chaugule, and P. Patil, "Text Mining Methods and Techniques," Int. J. Comput. Appl., vol. 85, no. 17, pp. 42–45, 2014.

[33] L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter, and J. N. Weinstein, "MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling.," Biotechniques, vol. 27, no. 6, pp. 1210–4, 1216–7, Dec. 1999.

[34] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia, "Automatic extraction of biological information from scientific text: protein-protein interactions.," Proceedings. Int. Conf. Intell. Syst. Mol. Biol., pp. 60–7, 1999.

[35] M. Krallinger and A. Valencia, "Text-mining and information-retrieval services for molecular biology.," Genome Biol., vol. 6, no. 7, p. 224, 2005.

[36] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology Behind Search, 2nd ed. Boston: Addison Wesley, 2011.

[37] Y. Lin, W. Li, K. Chen, and Y. Liu, "A Document Clustering and Ranking System for Exploring MEDLINE Citations," J. Am. Med. Informatics Assoc., vol. 14, no. 5, pp. 651–661, Sep. 2007.

[38] S. J. Darmoni et al., "Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases," J. Med. Libr. Assoc., vol. 100, no. 3, pp. 176–183, Jul. 2012.

[39] M. Petrova, P. Sutcliffe, K. W. M. B. Fulford, and J. Dale, "Search terms and a validated brief search filter to retrieve publications on health-related values in Medline: a word frequency analysis study.," J. Am. Med. Inform. Assoc., vol. 19, no. 3, pp. 479–88, 2012.

[40] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval International Student Edition. Chennai: Cambridge University Press, 2008.

[41] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition.," Pac. Symp. Biocomput., pp. 652–63, 2008.

[42] K. Raja, S. Subramani, and J. Natarajan, "A hybrid named entity tagger for tagging human proteins/genes.," Int. J. Data Min. Bioinform., vol. 10, no. 3, pp. 315–28, 2014.

[43] M. Torii, C. N. Arighi, G. Li, Q. Wang, C. H. Wu, and K. Vijay-Shanker, "RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information.," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 12, no. 1, pp. 17–29.

[44] K. Raja, S. Subramani, and J. Natarajan, "PPInterFinder--a mining tool for extracting causal relations on human proteins from literature," Database, vol. 2013, p. bas052-bas052, Jan. 2013.

[45] J. Natarajan, D. Berrar, C. J. Hack, and W. Dubitzky, "Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications.," Crit. Rev. Biotechnol., vol. 25, no. 1–2, pp. 31–52.

[46] K. E. Ravikumar, K. B. Wagholikar, D. Li, J.-P. Kocher, and H. Liu, "Text mining facilitates database curation - extraction of mutation-disease associations from Bio-medical literature.," BMC Bioinformatics, vol. 16, p. 185, Jun. 2015.

[47] S. Matos et al., "Mining clinical attributes of genomic variants through assisted literature curation in Egas," Database, vol. 2016, p. baw096, Jun. 2016.

[48] S. Subramani, R. Kalpana, P. M. Monickaraj, and J. Natarajan, "HPIminer: A text mining system for building and visualizing human protein interaction networks and pathways.," J. Biomed. Inform., vol. 54, pp. 121–31, Apr. 2015.

[49] J. Czarnecki, I. Nobeli, A. M. Smith, and A. J. Shepherd, "A text-mining system for extracting metabolic reactions from full-text articles.," BMC Bioinformatics, vol. 13, p. 172, Jul. 2012.

[50] R. Mishra et al., "Text summarization in the biomedical domain: A systematic review of recent research," J. Biomed. Inform., vol. 52, pp. 457–467, Dec. 2014.

[51] F. Zhu et al., "Biomedical text mining and its applications in cancer research.," J. Biomed. Inform., vol. 46, no. 2, pp. 200–11, Apr. 2013.

[52] H. Tan and P. Lambri, "Selecting an ontology for biomedical text mining," in Proceeding BioNLP '09 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, 2009, pp. 55–62.

[53] Y. Qi, Y. Zhang, and Min Song, "Text Mining for Bioinformatics: State of the Art Review," in 2009 2nd IEEE International Conference on Computer Science and Information Technology, 2009, pp. 398–401.

[54] X. Yin, J. X. Huang, and Z. Li, "Mining and modeling linkage information from citation context for improving biomedical literature retrieval," Inf. Process. Manag., vol. 47, no. 1, pp. 53–67, Jan. 2011.

[55] L. Tari et al., "Mining Gene-centric Relationships from Literature to Support Drug Discovery," in 2011 IEEE International Conference on Bioinformatics and Biomedicine, 2011, pp. 639–644.

[56] A. Bchir and W. Ben Abdessalem Karaa, "Extraction of drug-disease relations from MEDLINE abstracts," in 2013 World Congress on Computer and Information Technology (WCCIT), 2013, pp. 1–3.

[57] V. Mala and D. K. Lobiyal, "Concepts extraction for medical documents using ontology," in 2015 International Conference on Advances in Computer Engineering and Applications, 2015, pp. 773–777.

[58] A. Roth, S. Subramanian, and M. K. Ganapathiraju, "Towards extracting supporting information about predicted protein-protein interactions," IEEE/ACM Trans. Comput. Biol. Bioinforma., pp. 1–1, Dec. 2015.

[59] A. Jimeno Yepes and R. Berlanga, "Knowledge based word-concept model estimation and refinement for biomedical text mining," J. Biomed. Inform., vol. 53, pp. 300–307, Feb. 2015.

[60] D. R. Blair, K. Wang, S. Nestorov, J. A. Evans, and A. Rzhetsky, "Quantifying the Impact and Extent of Undocumented Biomedical Synonymy," PLoS Comput. Biol., vol. 10, no. 9, p. e1003799, Sep. 2014.

[61] C. Meaney, R. Moineddin, T. Voruganti, M. A. O'Brien, P. Krueger, and F. Sullivan, "Text mining describes the use of statistical and epidemiological methods in published medical research," J. Clin. Epidemiol., vol. 74, pp. 124–132, Jun. 2016.

[62] D. Jurafsky and J. Martin, Natural Language Processing. Englewood Cliffs, NJ: Pearson, 2008.

[63] C. Manning, Foundations of Statistical Natural Language Processing. Cambridge, MA: The MIT Press, 1999.

[64] A. Bodile and M. Kshirsagar, "Text mining in radiology reports by statistical machine translation approach," in 2015 Global Conference on Communication Technologies (GCCT), 2015, pp. 238–241.

[65] M. Mannai and W. Ben Abdessalem Karaa, "Bayesian information extraction network for Medline abstract," in 2013 World Congress on Computer and Information Technology (WCCIT), 2013, pp. 1–3.

[66] Ming Zhong and Xiangji Huang, "An effective extension to okapi for biomedical text mining," in 2006 IEEE International Conference on Granular Computing, 2006, pp. 615–618.

[67] M. T. Islam, D. Bollina, A. Nayak, and S. Ranganathan, "Intelligent Agent System for Bio-medical Literature Mining," in 2007 International Conference on Information and Communication Technology, 2007, pp. 57–63.

[68] H. Liu, M. Torii, Z. Hu, and C. Wu, "Mapping Gene/Protein Names in Free Text to Biomedical Databases," in Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), 2007, pp. 101–106.

[69] G. Leroy, M. Fiszman, and T. C. Rindflesch, "The Impact of Directionality in Predications on Text Mining," in Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), 2008, pp. 228–228.

[70] S. Salahuddin and R. M. Rahman, "Mining biomedical data from hypertext documents," in 14th International Conference on Computer and Information Technology (ICCIT 2011), 2011, pp. 417–422.

[71] W.-J. Hou, L.-C. Chen, and C.-S. Lu, "Identifying gene-disease associations using word proximity and similarity of Gene Ontology

terms," in 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI), 2011, pp. 1748–1752.

[72] Z. Javed and H. Afzal, "Biomedical text mining for concept identification from traditional medicine literature," in International Conference on Open Source Systems and Technologies, 2014, pp. 206–211.

[73] M. Berardi, D. Malerba, and M. Attimonelli, "Mining Information Extraction Models for HmtDB annotation," in Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), 2006, pp. 207–212.

[74] Jahiruddin, M. Abulaish, and L. Dey, "A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora," J. Biomed. Inform., vol. 43, no. 6, pp. 1020–1035, Dec. 2010.

[75] A. Sharma, R. Swaminathan, and H. Yang, "A Verb-Centric Approach for Relationship Extraction in Biomedical Text," in 2010 IEEE Fourth International Conference on Semantic Computing, 2010, pp. 377–385.

[76] Y. Li, X. Hu, H. Lin, and Z. Yang, "A Framework for Semisupervised Feature Generation and Its Applications in Biomedical Literature Mining," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 8, no. 2, pp. 294–307, Mar. 2011.

[77] A. Holzinger, K.-M. Simonic, and P. Yildirim, "Disease-Disease Relationships for Rheumatic Diseases: Web-Based Biomedical Textmining an Knowledge Discovery to Assist Medical Decision Making," in 2012 IEEE 36th Annual Computer Software and Applications Conference, 2012, pp. 573–580.

[78] R. T. Pereira, H. Costa, S. Carneiro, M. Rocha, and R. Mendes, "Reconstructing transcriptional Regulatory Networks using data integration and Text Mining," in 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015, pp. 1552–1558.

[79] M. A. Landge and K. Rajeswari, "A Survey on Chemical Text Mining Techniques for Identifying Relationship Network between Drug Disease Genes and Molecules," Int. J. Comput. Appl., vol. 146, no. 1, pp. 5–9, 2016.

[80] P. Tangtulyangkul, T. S. Hocking, and C. C. Fung, "Intelligent information mining from veterinary clinical records and open source repository," in TENCON 2009 - 2009 IEEE Region 10 Conference, 2009, pp. 1–6.

[81] C. Cano, A. Labarga, A. Blanco, and L. Peshkin, "Collaborative semi-automatic annotation of the biomedical literature," in 2011 11th International Conference on Intelligent Systems Design and Applications, 2011, pp. 1213–1217.

[82] H. Yang and Y. Dong, "Recognizing hierarchically related biomedical entities using MeSH-based mapping," Tsinghua Sci. Technol., vol. 17, no. 6, pp. 609–618, Dec. 2012.

[83] S. Mohammed and A. Nazeer, "An improved method for extracting acronym-definition pairs from biomedical Literature," in 2013 International Conference on Control Communication and Computing (ICCC), 2013, pp. 194–197.

[84] Y. Ji, Y. Tian, F. Shen, and J. Tran, "High-Performance Biomedical Association Mining with MapReduce," in 2015 12th International Conference on Information Technology - New Generations, 2015, pp. 465–470.

[85] J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski, "Complex event extraction at PubMed scale.," Bioinformatics, vol. 26, no. 12, pp. i382-390, Jun. 2010.

[86] M. Miwa, R. Saetre, J.-D. Kim, and J. Tsujii, "Event extraction with complex event classification using rich features.," J. Bioinform. Comput. Biol., vol. 8, no. 1, pp. 131–46, Feb. 2010.

[87] M. Naughton, N. Kushmerick, and J. Carthy, "Event Extraction from Heterogeneous News Sources," 2006.

[88] F. Hogenboom, F. Frasincar, U. Kaymak, F. de Jong, and E. Caron, "A Survey of event extraction methods from text for decision support systems," Decis. Support Syst., vol. 85, pp. 12–22, May 2016.

[89] C. Jonquet et al., "NCBO Resource Index: Ontology-based search and mining of biomedical resources," Web Semant. Sci. Serv. Agents World Wide Web, vol. 9, no. 3, pp. 316–324, Sep. 2011.

[90] A. B. Can and N. Baykal, "MedicoPort: a medical search engine for all.," Comput. Methods Programs Biomed., vol. 86, no. 1, pp. 73–86, Apr. 2007.

[91] M. Ghanem, Y. Guo, and A. S. Rowe, "Integrated Data Mining and Text Mining In Support of Bioinformatics," in Poster at Proceedings of the UK e-Science All Hands Meeting, 2004, pp. 1–3.

[92] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "New Frontiers in Biomedical Text Mining," Pacific Symp. Biocomput., vol. 12, pp. 205–208, 2007.