International Journal of Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-3)
Research Article | **March 2015**

An Overview of Text Mining Techniques and Methodologies Used in Bioinformatics Domain

R. Savitha*, Dr. R. Porkodi
Department of Computer Science, Bharathiar University,
Coimbatore, Tamil Nadu, India

Abstract—

The field of biological research is growing rapidly. The sheer volume in biomedical literature makes it difficult to access and locate the needed information. Support for publications in biomedical area is very challenging. Surplus amount of unstructured data in textual format is available online on websites such as PubMed. Searching a reliable data in a particular domain is difficult and there are no substantial methods to retrieve and analyse very large amount of datasets in an effective manner. The goal of biomedical text mining is to allow researchers to identify needed information more efficiently from sources such as Medline abstracts, abbreviations from scientific texts, etc. They use various manipulating methods and techniques to achieve the required results. This paper reviews various techniques proposed by various researchers such as new techniques, algorithms, tools and methodologies like MyMed, DiscoTEX, dictionaries, etc. These techniques are assessed on the extracted information and knowledge in an efficient manner to find out the relationships among the information. These techniques are employed to lessen the burden of information overload by applying it to the vast data source).

Keywords— Text mining, Text mining Tools, Bioinformatics, Medline abstracts.

I. INTRODUCTION

Text mining is a process of discovering knowledge and information that is difficult to retrieve by analysing unstructured texts from large sets of databases. Information extracting is the main objective of text mining. Development of a text mining system includes phases of design, construction, integration and evaluation. The work includes gathering information, developing corpora, annotating them with information and assessing parts of them to specify the use of language, collecting NLP components that analyse the language usages and then integrating the NLP components and required resources into a system. Evaluation is done by evaluating the system against the corpora during the evaluation phase.

Text mining is the application of techniques from machine learning, natural language processing (NLP), information extraction and statistical/mathematical approaches to automated extraction of useful knowledge from text [9]. Text mining of biomedical literature has been applied successfully to various biological problems. Many studies have been focusing on protein-protein and gene-protein interactions [14].

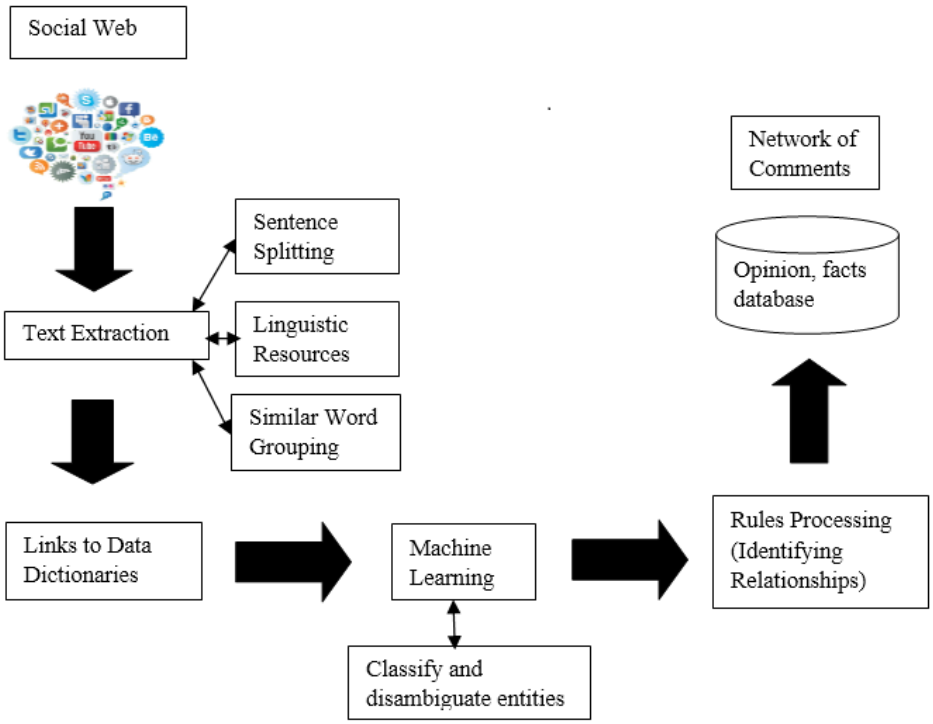

Fig. 1 Text mining process

II. BIOINFORMATICS IN TEXT MINING

Biomedical text mining (also known as BioNLP) refers to text mining applied to texts and literature of the biomedical and molecular biology domain. It is a rather recent research field on the edge of natural language processing, bioinformatics, medical informatics and computational linguistics. There is an increasing interest in text mining and information extraction strategies applied to the biomedical and molecular biology literature due to the increasing number of electronically available publications stored in databases such as PubMed.

The main developments in this area have been related to the identification of biological entities (named entity recognition), such as protein and gene names as well as chemical compounds and drugs in free text [1], the association of gene clusters obtained by microarray experiments with the biological context provided by the corresponding literature, automatic extraction of protein interactions and associations of proteins to functional concepts (e.g. gene ontology terms). Even the extraction of kinetic parameters from text or the subcellular location of proteins has been addressed by information extraction and text mining technology. Information extraction and text mining methods have been explored to extract information related to biological processes and diseases [2].

Biomedical Informatics domain represents the common knowledge between biomedical domain and bioinformatics domain. Each subclass of this entity inherits characteristics from its domain and properties related with the biomedical informatics domain [15].

This paper gives an overview of text mining tools in bioinformatics domain. Text mining processes includes the following steps,

- Information Retrieval: There are also advanced knowledge information retrieval systems besides conventional information retrieval systems, which integrate data from different resources into a single context to enhance our understanding of complex biomedical systems.
- Named Entity Recognition and Relation Extraction: In the extraction of knowledge [24], named entity recognition is considered the most important step, which identifies specific terms, such as gene, protein, disease, and drug. Several computing technologies have been employed for biomedical term identification. Dictionary-based approaches, rule-based approaches and machine learning approaches are the three major categories of current biomedical named entity recognition technique.
- Knowledge Discovery: Knowledge including facts, information, descriptions refers to the theoretical or practical understanding of a domain or a subject [26]. Knowledge discovery is the creation of knowledge from large volumes of structured or unstructured data. The knowledge obtained may become additional data that can be used for further usage and discovery. Knowledge discovery is a very important part of data mining. Text mining, also referred as text data mining, is a branch of data mining that particularly deals with text.
- Hypothesis Generation: Scientific hypothesis is like a scientific imagination which is based on existing evidence and knowledge. Hypothesis generation is to get unproved inference with clues hidden in the text while knowledge discovery means extracting novel knowledge. The biomedical literature is a treasure trove of potential information for making biomedical inferences and generating new hypotheses. Hypothesis generation is an important task in text mining, which is very helpful for biomedical researchers who want to infer unknown biomedical facts that can be used to guide the design of experiments or explain existing experimental results. This task is gradually receiving much more attention from researchers.
- Gene/Protein name and Synonym Dictionary Creation: Gene and protein names are mentioned as gene symbols, protein names, synonyms, gene names, and typographical variants. Some of them are ambiguous. To reduce the synonyms and phases representing the same gene, normalization with dictionary based approach is necessary. Precision comes down when false positives are larger in number, which is caused by short names. String matching approach and terminological resources is being integrated into the dictionary based methods to locate gene that is mentioned. This approach uses a list of biological entity names and identifies their occurrence in text by using various substring matching techniques. There are various methods to construct a dictionary, it can be manually constructed, automatically generated, also manually constructed first and then can be semi-automatically extended. Spelling variation is another problem of dictionary creation. A word with a slight variation in spelling can be considered as two different terms when exact matching technique is used. Approximate string matching method helps in sorting out this problem when there is a surface level similarity.
- Keywords Extraction from Biomedical Literature: The end result in almost all the researches are available in text form, as in MEDLINE abstracts, comment fields of relevant reports, as in GenBank feature table annotations. Such information is useful for analysis, classification of proteins, extraction of protein-protein interactions, new functional relationships, maintaining information of material and methods. However, information in text form cannot be used in computerized systems. IE extracts information from MEDLINE abstracts and other sources and converts it to computer readable form.
- MEDLINE Abstracts Keywords Extraction. Text mining is an alternative to manual information extraction. Off late, most reports on text mining of scientific literature have used the MEDLINE repository for tasks such as extracting protein interactions and for benefiting researchers. The text processing units from which MEDLINE extraction is carried out are full abstracts, component sentences, or phrases. By using PUBMED interface, the user can submit a query to the database consisting of the AND of two biochemical terms, and abstracts in MEDLINE containing both terms are returned.

- Gene-Name Normalization: The recognition and normalization of gene mentions in biomedical literature are crucial steps in biomedical text mining. Gene name recognition, entity mapping, disambiguation and filtering are the major components.
- Gene Normalization: Gene normalization is assigning a unique identifier from a database to the gene mentions. Using these identifier, information can be gathered from external databases such as interactions, pathways, sequences and protein structures. Normalizing gene mentions in articles is difficult because of the ambiguity of the high gene mentions in biomedical publications. In biomedical scientific articles, taxonomical entities are used besides concrete species mentions as references to different group of organisms. Species taxonomies are hierarchical systems (trees) of living creatures and therefore provide a classification of species. Some papers also investigate the added value of the utilization of taxonomic entity mentions in the inter-species gene normalization task.
- Tokenization: Tokenizing is an operation that splits the document into a sequence of tokens. Tokenization mode is defined by splitting points. Split points are chosen differently depending on the mode. The incoming document will be split into tokens on each of these characters. The common words that are hurdles for text mining such as prepositions, articles, and pronouns are considered stopwords. Every text document has words which are unimportant for text mining application. It reduces the text data and improves the system performance.
- Filtering: Filtering helps to design data sources and mining. Filters can be created to use a part of the data for training and testing different models. Filter is used by length, content, English, dictionary. Tokens are filtered by length. Min chars is the minimal number of characters that a token must contain to be considered. Max chars is the maximal number of characters that a token must contain to be considered.

MEDLINE, the principal online bibliographic citation database of NLM's is used internationally to provide access to the world's biomedical journal literature. The decision to index a journal for this service is done by the Director of the National Library of Medicine based on scientific policy and scientific quality. The Literature Selection Technical Review Committee (LSTRC) has been established to review journal titles and their quality.

The current MEDLINE database includes 2,208,948 abstracts as of 2014. There are 2,617,986 records that are computer readable in MEDLINE database as shown in Fig 2. Several approaches have been proposed on ontology productivity. This paper incorporates several information retrieval methodologies and experimental results.

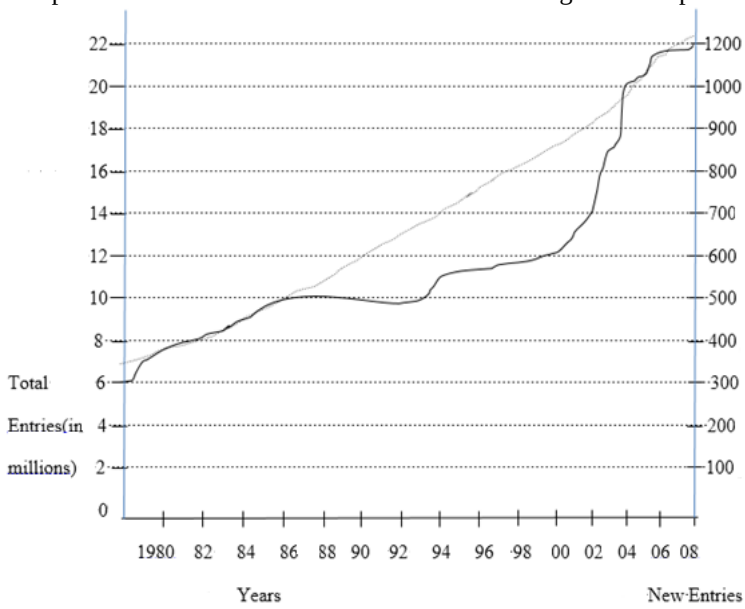

Fig. 2 MEDLINE entries (in thousands)

PubMed [18] is one of the best biomedical databases which contain more than 20 million citations on biomedical articles from MEDLINE and life science journals [19]. This is a web-based search portal for users and also serves as an application program interface for developers. Textpresso [19, 20] uses ontology and returns concepts like cell, or phenotype, classes of relations of objects like association, regulation, and related descriptions. GoPubMed [21, 22] classifies literature abstracts according to gene ontology and shows the ontology terms that are related to the query words. Users can explore PubMed search results with an ontology viewer.

This paper is organized as follows, section I describes introduction, section II explains the role of text mining in the field of bioinformatics and text mining processes. Section III reviews the related literatures. Section IV describes the biomedical tools for text mining and section V describes the conclusion of this paper.

III. REVIEW OF LITERATURE

Khaled Khelif, Rose Dieng-Kuntz, and Pascal Barbry [31] propose a model for generation of ontology-based semantic annotations which is called MeatAnnot and MeatSearch to draw advanced inferences on these annotations.

They say that this helps biologist who mainly work on DNA microarray experiments. This is a system to support people working on DNA microarray experiments for validation of their experiments and interpretation. They also state that CORESE query language that they have used in MEAT is similar to SPARQL

Hong Yu,a, Vasileios Hatzivassiloglou, Andrey Rzhetsky, and W. John Wilbur [28] have developed a software system called "GPmarkup" that identifies gene/protein terms in MEDLINE abstracts automatically, along with that, a knowledge source of paired gene/protein symbols and full names is also generated. They say that many of the pairs in their knowledge source do not appear in current GenBank database, so their methods may also be used for automatic lexicon generation.

Jeffrey T. Chang, HinrichSchutze, Ph.D. and Russ B. Altman, M.D., Ph.D. [32] have developed an algorithm to match abbreviations in text with their expansions. Their algorithm uses a statistical learning algorithm and logistic regression to score abbreviation expansions based on their resemblance to a training set of human-annotated abbreviations. They have applied it to Medstract, a corpus of MEDLINE abstracts in which abbreviations and its expansions are manually annotated. They have then run the algorithm on all abstracts in MEDLINE creating a dictionary. They have tested the coverage of the database with an independently created list and measured the recall and precision as 83% and 80% respectively when searched against the China Medical Tribune dictionary. It should be noted that they have not linked their dictionary with any external dictionaries. The algorithm is also run against Medstract Gold Standard.

Un Yong Nahm and Raymond J. Mooney [14] propose that text mining and Information Extraction (IE) are both topics of recent interest. Text mining is knowledge discovery from databases (KDD) techniques to unstructured text. This paper describes a system called DiscoTEX, which combines IE and KDD methods to perform the task of text mining, discovering prediction rules from natural language corpora. Initially, DiscoTEX is developed by integrating an IE module based on Rapier and a rule-learning module, Ripper. On applying these techniques to a corpus of computer job postings from an Internet newsgroup, good results were obtained. Due to a growing interest in the text mining field, few working systems and detailed experimental evaluations has been going on. By using existing IE and KDD technology, text mining systems can be developed rapidly and can be evaluated on existing IE corpora. In this paper, they presented an approach to using an automatically learned IE system to extract structured databases from a text corpus, and then mining this database with traditional KDD tools. Their preliminary experimental results demonstrate that the knowledge discovered from such an automatically extracted database is close in accuracy to the knowledge discovered from a manually constructed database.

R. Porkodi and Dr. B. L. Shivakumar [33] say that text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. The research aspect in text mining includes information extraction, information search and retrieval, text document categorization, text document clustering and text summarization. This research work focuses on information extraction and mining of association rules in which the information extraction is used to identify and extract textual information of entities and their relationships. The mining of association rules is extraction of semantic knowledge among entities that are extracted in Information Extraction phase. Their research is designing information extraction approach to extract entities from Medline abstracts and to mine association rules among them. The work uses Bioinformatics databases, Gene Ontology, MeSH ontology and Medline abstracts related to Alzheimer's disease.

Rania A. AbulSeouda and Mai S. Mabrouk [34] state that hepatocellular carcinoma (HCC) is the third leading cause of cancer-related mortality all over the world. Chromosomal copy number alterations can lead to activation of oncogenes and inactivation of tumor suppressors in human cancers. Identifying cancer-specific CNAs will helping in understanding the molecular basis of tumorgenesis and facilitate the identification of HCC biomarkers using CNA. They propose TMT-HCC system, a tool for text mining the biomedical literature for hepatocellular carcinoma, (HCC) biomarkers identification. TMT-HCC provides a way to identify molecular biomarkers of HCC to inform diagnosis and treatment driver genes with causal roles in carcinogenesis to detect regions under frequent alterations in genomes in case of cancers (CNAs). TMT-HCC also extracts protein–protein interactions from the scientific papers. The results prove that the integration of genomic and transcriptional data offers potential for identifying novel cancer genes in HCC pathogenesis.

Jeffrey T. Chang, HinrichSchutze, Ph.D. and Russ B. Altman, M.D., Ph.D. [35] have developed an algorithm to match abbreviations in text with their expansions. Their algorithm uses six statistical learning algorithm and logistic regression to score abbreviation expansions based on their resemblance to a training set of human-annotated abbreviations. They have applied it to Medstract, a corpus of MEDLINE abstracts in which abbreviations and its expansions are manually annotated. They have then run the algorithm on all abstracts in MEDLINE creating a dictionary. They have tested the coverage of the database with an independently created list and measured the recall and precision as 83% and 80% respectively when searched against the China Medical Tribune dictionary. It should be noted that they have not linked their dictionary with any external dictionaries. The algorithm is also run against Medstract Gold Standard.

Some of the biomedical tools used for text mining that is surveyed in this paper are ALICE, GRNS, DiscoTEX, TMT-HCC, GPmarkup, and KID algorithm.

IV. BIOMEDICAL TOOLS FOR TEXT MINING

A. ALICE

This is an algorithm, which is developed by Hiroko Ao and Toshihisa Takagi, Ph.D. [29]. They have constructed this support system called ALICE (Abbreviation LIfter using Corpus-based Extraction). This algorithm extracts

abbreviations and its expansions from any literature using heuristic pattern matching rules. This system is said to have three phases and identifies 320 abbreviation-expansion patterns. The subtly compiled heuristics enables it to extract abbreviations without reducing precision. It is also said that ALICE retrieval is more accurate from a literature, also recognizes unidentified abbreviation in a paper by constructing an abbreviation dictionary.

1) Procedure: This consists of three phases which are inner search, outer extraction, and validity judgment [29]. In the inner search phase, the trigger for ALICE to go into this phase is pair of parenthesis. The algorithm checks whether the inner is a candidate abbreviation or expansion. The acceptance condition is constructed based on the presence of a space before left parenthesis and characters of the inner word. The discard conditions are to see whether the inner front word, first inner word is included in the corresponding stop word lists. Also, checks word categories.

This consists of three phases which are inner search, outer extraction, and validity judgment [29]. In the inner search phase, the trigger for ALICE to go into this phase is pair of parenthesis. The algorithm checks whether the inner is a candidate abbreviation or expansion. The acceptance condition is constructed based on the presence of a space before left parenthesis and characters of the inner word. The discard conditions are to see whether the inner front word, first inner word is included in the corresponding stop word lists. Also, checks word categories.

In the outer extraction phase, about 16 templates are used to extract an outer from its left chunk. They are constructed based on how abbreviations are formed from their expansions. The discard conditions were constructed to see whether the outer and first outer word is included in their corresponding stop word lists. Also, to check the length of the string in the outer.

In the validity judgment phase, about 5 acceptance conditions and 14 discard conditions are used to judge the validity of a set of an inner and its outer through 11 steps. They have also been constructing a system known as PETER and ALICE is a part of it. PETER is developed to select relevant papers from PubMed search.

2) Result: A Corpus of 1000 abstracts with titles randomly selected from the MEDLINE database. The recall is 95% and precision is 97% [29].

B. GRNS

The framework called Gene Regulatory Networks System is developed by Yong-Ling SONG, and Su-Shing CHEN [30]. The GRNS automatically mines biomedical literature and construct gene regulatory networks based on extracted information and some existing domain-specific knowledge. GRNS also detects informative sentences and save them and they are sorted by a heuristic sentence ranking score.

1) Procedure: The first step is information extraction [30]. Then comes, strain table analysis. The strain table is analysed and genotype and strain number is extracted from the table. The third step is gene, protein, and discriminating words recognition. After strain table analysis tokenization, sentences splitter and Part Of Speech tagging is done. The gene and protein names discrimination and term recognition is performed. The next step is relation and phenotype identification. Cascaded finite state automata is used to recognize gene regulatory relation and phenotype information. Then, unrecognized sentences detection and ranking. If a sentence includes gene/protein names and fails to match the existing pattern, they assign that sentence to a template candidate. Sentence ranking schema is also used. The rankings are, document relevance score-Sd, location document relevance score-S1, sentence ranking score.

Automatic construction and visualization of regulatory networks: The first step is information extraction [30]. Then comes, strain table analysis. The strain table is analysed and genotype and strain number is extracted from the table. The third step is gene, protein, and discriminating words recognition. After strain table analysis tokenization, sentences splitter and Part Of Speech tagging is done. The gene and protein names discrimination and term recognition is performed. The next step is relation and phenotype identification. Cascaded finite state automata is used to recognize gene regulatory relation and phenotype information. Then, unrecognized sentences detection and ranking. If a sentence includes gene/protein names and fails to match the existing pattern, they assign that sentence to a template candidate. Sentence ranking schema is also used. The rankings are, document relevance score-Sd, location document relevance score-S1, sentence ranking score.

2) Result: For 200 randomly selected papers about Pseudomonas aeruginosa Genome [30] and filter out to display the data on the type III secretion subsystem the result is, strain number gave 93% precision and 92% recall. Genotype gave 90% precision and 89% recall. Gene regulatory gave 91% precision and 79% recall. Phenotype gave 87% precision and 74% recall.

C. DISCOTEX

DiscoTEX is a system proposed by Un Yong Nahm and Raymond J. Mooney [14] that combines IE and KDD methods to perform a text mining task, discovering prediction rules from natural-language corpora. An initial version of DiscoTEX is constructed by integrating an IE module based on Rapier and a rule-learning module, Ripper. They present encouraging results on applying these techniques to a corpus of computer job postings from an Internet newsgroup. There is a growing interest in the general topic of text mining; however, there are few working systems or detailed experimental evaluations. By utilizing existing IE and KDD technology, text-mining systems can be developed relatively rapidly and evaluated on existing IE corpora. In this paper, they presented an approach to using an automatically learned IE system to extract a structured database from a text corpus, and then mining this database with traditional KDD tools. Their preliminary experimental results demonstrate that the knowledge discovered from such an automatically extracted database is close in accuracy to the knowledge discovered from a manually constructed database.

1) Procedure: Information extraction, they have extracted a database from postings to the USENET newsgroup, austin.jobs [14]. They have collected 1000 postings and classified them as relevant and irrelevant by human expertise. Naïve-Bayes text categorizer is used to identify relevant documents. The categorizer gave 99% accuracy. RAPIER, a machine learning system for information extraction from natural language texts is used to construct IE module for DiscoTEX.

2) Result: About 600 computer science job postings to the newsgroup austin.jobs were collected. DiscoTEX gave classification accuracy of 92.7% and all-absent strategy had an accuracy of 92.5% [14].

D. TMT-HCC

Rania A. Abul Seouda and Mai S. Mabrouk [34] have proposed TMT-HCC system, a tool for text mining the biomedical literature for hepatocellular carcinoma, (HCC) biomarkers identification. TMT-HCC provides a way to identify molecular biomarkers of HCC to inform diagnosis and treatment driver genes with causal roles in carcinogenesis to detect regions under frequent alterations in genomes in case of cancers (CNAs). TMT-HCC also extracts protein–protein interactions from the scientific papers. The results prove that the integration of genomic and transcriptional data offers potential for identifying novel cancer genes in HCC pathogenesis.

1) Procedure: The TMT-HCC system starts full text retrieval from databases, which has NCBI and PubMed published papers [34]. Then it tags the full texts and stores in a local database. The tagged papers are pre-processed to replace synctactic constructs (tokenizer). Then parsing is done using Stanford parser. The system then identifies bio-interaction terms.

TMT-HCC is implemented with C#. GUI is developed using Microsoft visual studio. Dictionaries are implemented using SQL server 2005.

2) Result: TMT-HCC system search database for protein and gene names. Protein name is highlighted in blue, DNA is highlighted in green, and a gene name is highlighted in brown. In case of no entity existed, the system will link to the UniProtKD for searching [34].

3) Evaluation: To evaluate the TMT-HCC, it is compared to the existing systems such as ABNER by entering 10 sentences to the parser, six of them are right and four are wrong input.

E. GPMARKUP

Hong Yu,a, Vasileios Hatzivassiloglou, Andrey Rzhetsky, and W. John Wilbur [28] have developed a software system called "GPmarkup" that identifies gene/protein terms in MEDLINE abstracts automatically, along with that, a knowledge source of paired gene/protein symbols and full names is also generated. They say that many of the pairs in their knowledge source do not appear in current GenBank database, so their methods may also be used for automatic lexicon generation.

Their method consists of 6 subsections like mapping gene/protein symbols to full names from MEDLINE abstracts, filtering out abbreviation-full name pairs, marking up gene/protein terms, evaluating GPmarkup, and measuring out the percentage of GP symbols in MEDLINE abstracts.

1) Procedure: Their method section consists of six sub-sections [28]. The first subsection is mapping gene/protein symbols to full names as well as abbreviations to full names. Second is generating a knowledge source of paired abbreviations and full names from MEDLINE abstracts. The third is filtering out other abbreviation-full name pairs to produce a knowledge source of paired gene/protein symbols and full names. Fourth subsection is marking up gene/protein terms in MEDLINE abstracts, then evaluating GPmarkup. The final subsection is measuring the percentage of defined gene/protein symbols in MEDLINE abstracts.

2) Result: About 100 abstracts from 782,560 MEDLINE abstracts were extracted, which has "protein" keyword. GPmarkup has shown 73% recall and 93% precision in marking up gene/protein terms in MEDLINE abstracts [28].

F. KID ALGORITHM

Stephanie Heinen, Bernhard Thielen and Dietmar Schomburg [36] present a rule and dictionary based text mining algorithm for the extraction of kinetic data. "KID the Kinetic Database" is the database which has the extracted information about the kinetic enzyme from 17 million Pubmed abstracts.

They present this text mining algorithm for the extraction of kinetic information such as KM, Ki, kcat etc. as well as associated information such as enzyme names, EC numbers, ligands, organisms, localizations, pH and temperatures. Using this rule- and dictionary-based approach, it was possible to extract 514,394 kinetic parameters of 13 categories from about 17 million PubMed abstracts and combine them with other data in the abstract.

The database may be of valuable help in the field of chemical and biological kinetics. It is completely based upon text mining and therefore complements manually curated databases.

1) Method: The algorithm is divided into two parts, identification of entities in the text and a rule-based linkage of these units [36].

2) Result: A manual verification of approximately 1,000 randomly chosen results yielded a recall between 51% and 84% and a precision ranging from 55% to 96%, depending of the category searched. The results were stored in a database and are available as "KID the Kinetic Database" via the internet [36].

TABLE I SURVEY OF THE TEXT MINING TOOLS

Tools	Description	Datasets Used	Accuracy

ALICE [29]	Extracts abbreviations and its expansions from any literature. [29]	1000 abstracts with titles selected randomly from MEDLINE database. [29]	High[29]
GRNS [30]	Mines biomedical literature and constructs gene regulatory networks. [30]	200 selected papers about Pseudomonas aeruginosa genome. [30]	High[30]
DiscoTEX [14]	A system that combines IE and KDD methods to perform a text mining task. [14]	600 scientific job postings to the newsgroup austin.jobs. [14]	Moderate[14]
TMT-HCC [34]	A tool for text mining the biomedical literature for hepatocellular carcinoma (HCC) biomarkers identification. [34]	NCBI and PubMed published papers. [34]	Moderate[34]
GPmarkup [28]	A system that identifies gene/protein terms in MEDLINE abstracts. [28]	100 abstracts from 782,560 MEDLINE abstracts, which has "protein" keyword. [28]	Moderate[28]
KID Algorithm [36]	A text mining algorithm for the extraction of kinetic data. [36]	17 million PubMed abstracts. [36]	Moderate[36]

TABLE II COMPARISON OF THE TEXT MINING TOOLS

Text Mining Tools	**Merits**	**Demerits**
ALICE [29]	Better than the existing systems. [29]	Cannot make clear distinction between synonyms and expansions. [29]
GRNS [30]	Ranking the template candidates gives good result in evaluating sentences. [30]	There is no structured strain table data directly. [30]
DiscoTEX [14]	Accuracy is close to the accuracy of manually constructed database. [14]	Similar slot fillers are collapsed manually in the extracted data. [14]
TMT-HCC [34]	Works with multithreads that enables multitask on same time. [34]	System is less scalable when new data is added to the system. [34]
GPmarkup [28]	All of the combinations are used for substitution. [28]	Might miss some gene/protein symbols, full names, and abbreviations. [28]
KID Algorithm [36]	Fast calculation time and a high precision compared to the other algorithms. [36]	Quality of the identification is limited by amount and quality of the entries. [36]

The above mentioned six biomedical text mining tools have both advantages and limitations with them. The advantage of using the ALICE is that it is far more efficient than the existing systems. Several limitations that the existing systems have, have been overcome by the support system, ALICE. This also has the disadvantage of having difficulty in making clear distinction between synonyms and expansions. It is also impossible to retrieve expansions divided by enumeration. The GRNS framework ranks the template candidates using sentence ranking schema and gives good result in evaluating the sentences. The limitation of GRNS is that there is no structured data directly and the extracted information does not give information such as start and end of tables, columns, and rows. The knowledge discovered from automatically extracted database by DiscoTEX system is close in accuracy to the knowledge discovered from a manually constructed database. Procedure of selecting slot used in rule mining is done manually. Similar slot fillers are collapsed manually in the extracted data, this is the main disadvantage of this system. The advantage of TMT-HCC is that it works with multithreads that enable users to perform multitask at the same time. The disadvantage of TMT-HCC is that it is less scalable as every year new data will be added to the system. This is a challenge to accommodate the increase in data without affecting the performance of the system significantly. GPmarkup includes all of the combinations for substitution when a full name has more than one word which has many abbreviations to it. It can also miss some gene/protein symbols, full names, and abbreviations when authors do not follow the guidelines for naming genes and proteins. This may also miss abbreviations and full names which are introduced through syntactic patterns. The advantage of using KID algorithm is fast calculation time and a high precision of the received information when compared to the other algorithms. Since they are dictionary based, the quality of the identification is limited by amount and quality of the entries. In contrast, removing false positives from the dictionary will negatively impact on recall.

V. CONCLUSIONS

There is seemingly growing interest in the field of text mining. By integrating the proposed methods, techniques and algorithms, discovering knowledge from vast text, corpora becomes easier. The extraction of required information and knowledge is made easier by building ontology, tools, and dictionaries. On reviewing some of the papers related to advanced text mining and data mining techniques and tools such as BONSAI, DiscoTEX, MyMED, MARLIN, GRNS,

this paper have found that they have shown good performances through their precision and recall values. Performance of some of the text mining tools such as ALICE, GRNS, DiscoTEX, TMT-HCC, GPmarkup, KID algorithm have been surveyed. The performance of ALICE and GRNS have been high when compared to the performance of DiscoTEX, TMT-HCC, GPmarkup and KID algorithm. They have shown relatively a moderate performance in the process of text mining. This paper have studied and analysed the techniques used, outcomes, challenges, shortfalls of the algorithms and methods used in text mining and how they can be made effective. XML provides many features than plain text making it necessary to explore further. The major challenge in the field of text mining for the next decade is development of text mining tools to benefit the researchers, which should have greater access to full text collections that uses it. As such, increase in precision or recall by itself cannot be equated to user's success in searching task. More work in the area of creating metrics and methods is required to design potentially efficient systems.

REFERENCES

[1] M. Krallinger & A. Valencia, *Genome Biology*, 6: 224 (2005) [PMID: 15998455].
[2] 1. Aronson AR, Bodenreider O, Demner-Fushman D, et al. *From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches*. In: Biological,Translational, and Clinical Language Processing, Prague, Czech Republic: *Association for Computational Linguistics*, 2007;105–12.
[3] Jeyakumar Natarajan1 and Jawahar Ganapathy. Functional gene clustering via gene annotation sentences, MeSH and GO keywords from biomedical literature. © 2007 Biomedical Informatics Publishing Group. ISSN 0973-2063. Bioinformation 2(5): 185-193 (2007).
[4] http://www.ncbi.nih.gov/entrez/
[5] http://bioinformatics.weizmann.ac.il/cards/
[6] http://ca.expasy.org/sprot/
[7] http://www.cse.ucsc.edu/centers/cbe/Genome/
[8] http://www.gene.ucl.ac.uk/hugo/
[9] Saurav Sahay, Baoli Li, Ernest V. Garcia , Eugene Agichtein and Ashwin Ram. *Domain Ontology Construction from Biomedical Text*. 2007 International Conference on Artificial Intelligence (ICAI'07), Las Vegas, Nevada, USA, CSREA Press.
[10] L. Smith, T. Rindflesch, and W. J. Wilbur. *Medpost: a part of speech tagger for biomedical text*. Bioinformatics, 20(14), 2004.
[11] Juana Maria Ruiz-Martinez, Rafael Valencia-Garcia, Jesualdo Tomas Fernandez-Breis, Francisco Garcia-Sanchez, Rodrigo Martinez-Bejar. *Ontology learning from biomedical natural language documents using UMLS*. J.M. Ruiz-Martínez et al. / Expert Systems with Applications 38 (2011) 12365–12378.
[12] Studer, R., Benjamins, V. R., & Fensel, D. (1998). *Knowledge engineering: Principles and methods. Data Knowledge and Engineering*, 25(1–2), 161–197.
[13] Gomez-Perez, A., Moreno, A., Pazos, J., & Sierra-Alonso, A. (2000). Knowledge maps: *An essential technique for conceptualisation. Data and Knowledge Engineering*, 33(2), 169–190.
[14] Robert Stevens, Carole A. Goble, and Sean Bechhofer. *Ontology-based Knowledge Representation for Bioinformatics*. _robert.stevens carole seanb_@cs.man.ac.uk. September 26, 2000.
[15] Paulo Gottgtroy, Prof. Nik Kasabov, Stephen MacDonell.. *An ontology driven approach for knowledge discovery in Biomedicine.* Paulo.gottgtroy, nkasabov, stephen.macdonell,}@aut.ac.nz. http://www.kedri.info.
[16] Sabin Corneliu Buraga, Liliana Cojocaru, Ovidiu Cătălin Nichifor. *Survey on Web Ontology Editing Tools*. PERIODICA POLITECHNICA, Transactions on AUTOMATIC CONTROL and COMPUTER SCIENCE Vol.NN (ZZ), 2006, ISSN 1224-600X.
[17] Fei Zhu a,b, Preecha Patumcharoenpol c,d, Cheng Zhang a, Yang Yang a,b, Jonathan Chan c, Asawin Meechai e, Wanwipa Vongsangnak a, Bairong Shen. *Biomedical text mining and its applications in cancer research*. 1532-0464/$ - see front matter _ 2012 Elsevier Inc. All rights reserved.
[18] Pubmed. <http://www.ncbi.nlm.nih.gov/pubmed/>
[19] Muller HM, Kenny EE, Sternberg PW. *Textpresso: an ontology-based information retrieval and extraction system for biological literature*. PLoS Biol 2004;2:e309.
[20] Textpresso. http://www.textpresso.org/
[21] Doms A, Schroeder M. Go. *PubMed: exploring PubMed with the Gene Ontology*. Nucleic Acids Res 2005;33:W783–6.
[22] GoPubMed. http://www.gopubmed.org/
[23] Leser U, Hakenberg J. *What makes a gene name? Named entity recognition in the biomedical literature*. Brief Bioinform 2005;6:357–69.
[24] Cohen AM, Hersh WR. *A survey of current work in biomedical text mining*. Brief Bioinform 2005;6:57–71.
[25] Frawley William J, Piatetsky-Shapiro G, Matheus Christopher J. *Knowledge discovery in databases: an overview*. AI Mag 1992;13:57–70.
[26] Nikolai Daraselia, Sergei Egorov, Andrey Yazhuk, Svetlana Novichkova, Anton Yuryev and Ilya Mazo. *Extracting Protein Function Information from MEDLINE Using a Full-Sentence Parser*. Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics.
[27] Mikio Yoshida, Ken-Ichiro Fukuda and Toshihisa Takagi. PNAD-CSS: *A workbench for constructing a protein name abbreviation dictionary*. Bioinformatics Ontology. Vol. 16 no. 2 2000. Pages 169-175.

[28] Hong Yu,a, Vasileios Hatzivassiloglou, Andrey Rzhetsky, and W. John Wilbur. *Automatically identifying gene/protein terms in MEDLINE abstracts*. H. Yu et al. Journal of Biomedical Informatics 35 (2002) 322–330.

[29] Hiroko Ao, Msc, Toshihisa Takagi, Ph.D: ALICE: An Algorithm to Extract Abbreviations from MEDLINE. Journal of the American Medical Informatics Association Volume 12 Number 5 Sep/Oct 2005.

[30] Yong-Ling SONG1, Su-Shing CHEN. *Text Mining Biomedical Literature for Constructing Gene Regulatory Networks*. Interdiscip Sci Comput Life Sci (2009) 1: 179–186.

[31] Khaled Khelif, Rose Dieng-Kuntz, Pascal Barbry. *An Ontology Based Approach to Support Text Mining and Information Retrieval in the Biological Domain*. Journal of Universal computer Science, 12(2007), 1881-1907.

[32] Jeffrey t. Chang, Hinrich Schütze, Ph.D., Russ B. Altman, M.D., Ph.D. *Creating an Online Dictionary of Abbreviations from MEDLINE. Journal of the American Medical Informatics Association*, Volume 9 Number 6 Nov/Dec 2002.

[33] R. Porkodi, Dr. B. L. Shivakumar. *Associations among entities related to Alzheimer Disease using Ontology based approach*. IOSR Journal of Engineering (IOSRJEN) ISSN: 2250-3021 Volume 2, Issue 8 (August 2012), PP 21-32 www.iosrjen.org.

[34] Rania A. Abul Seouda, Mai S. Mabrouk. *TMT-HCC: A tool for text mining the biomedicalliterature for hepatocellular carcinoma (HCC) biomarkers identification*. Computer methods and programs in biomedicine 112 (2013) 640–648. www.intl.elsevierhealth.com/journals/cmpb.

[35] Jeffrey t. Chang, Hinrich Schütze, Ph.D., Russ B. Altman, M.D., Ph.D. *Creating an Online Dictionary of Abbreviations from MEDLINE*. Journal of the American Medical Informatics Association Volume 9 Number 6 Nov/Dec 2002.

[36] Stephanie Heinen, Bernhard Thielen and Dietmar Schomburg. *KID-an algorithm for fast and efficient text mining used to automatically generate a database containing kinetic information of enzymes*. Heinen et al. BMC Bioinformatics 2010, 11:375. http://www.biomedcentral.com/1471-2105/11/375.

[37] Shilpa Dang, PeerzadaHamid Ahmad. *Text Mining: Techniques and its Application*. IJETI International Journal of Engineering & Technology Innovations, Vol. 1 Issue 4, November 2014 ISSN (Online): 2348-0866 www.IJETI.com.

