Methods and Trends in Information Retrieval in Big Data Genomic Research

Joseph M. De Guia, Madhavi Deveraj

Abstract— This paper described information retrieval (IR) and the common methods of finding, extracting, and mining information in genomic research through text mining, and natural language processing (NLP). There was a surge of genomic information from the different literature and the production of genome datasets that catapulted the development of several tools for analyzing and presenting new found knowledge in the biomedical and genome research. This paper presented the recent research trends, survey, reviews, experiments, and concepts in information retrieval applied to text, images and object features in big data genomic research. The method used is exploratory survey research in IR uses in genomic research that presents the concepts, methods, evaluation results and next steps described by the key researchers.

Keywords: Information retrieval, text mining, natural language processing, big data, genome, genomic research.

I. INTRODUCTION

The science of searching information that deals with the search of documents containing text and content-based data (meta-data description), databases of texts, and multimedia objects is information retrieval. An example of the IR system are web search engines, online libraries, databases of the electronic journals, books and `other documents. IR system processes the query of users that search for any collections of data retrieved from relevant documents or provide results based on the rank. The IR researchers also explored the issues of how to make data transmission retain its original content and the same way if the data is compressed and reconstructed from the lossless compression. Notable researchers in IR focused on text information [1,2] transformation, analysis, retrieval of information, [3,4] modern information retrieval principles and overview.

The IR as a research area became an important aspect of how people access and find data and information in a certain process that makes it faster and more accurate. The history of IR started from the classic manual search was implemented through the library system using card catalog to locate library collections by taking the card and locating the shelve of the book or material [5]. The introduction of the computer facilitated the information retrieval research group by [1] in 1960. The progression made into a large-scale retrieval systems by the National Institute of Standards and Technology (NIST) in 1992 with the US Department of Defense look at the information retrieval community of interest and provided resources for the Text Retrieval Conference (TREC) to evaluate the state-of-the art text retrieval methods [6]. Recent trends in IR methods in content-based system for objects and multimedia features were proposed to improve the algorithms and tasks involved in the retrieval of images in large-scale image databases and datasets described in the papers [7,8,9,10].

The web became the dominant source of information for most of the users and researchers looking for raw, simple, and complex information resources. The growth of data that are too complex to be processed and its conceptual characteristics such as volume, variety, velocity, and veracity in many systems uses advanced tools to reveal meaningful information and insights. Regardless of any data and information that is largely made of text information. IR can be leveraged to access the information at a speed, with accurate and actionable results. It was estimated by IDC and Statista in 2025 there will 163 zettabytes volume of data [11,12] or generated knowledge can be essential uses for big data. The IR and in combination with other technologies and methods, it became the source of reference and tool of choice for most of the users and researchers. The advanced technology of IR became the part of the scientific discovery process where the knowledge in the vast collections of literature and data repository became available to any researchers.

In biomedical and genomic research, the human genome is a sequence of the human genes and codified as described by the Human Genome Project [8]. The diversity of the human codes, genome science and technologies became possible through systematic genome-wide searches and genome sequencing technologies [9]. This technology analyzed many types of diseases such as cancer and other dreadful diseases. The International Cancer Genome Consortium analyzed more than 25,000 cancer genomes as of 2013 [10]. There was a rapid expansion of the cancer genome data sets also accelerated the genetic analytical tools for genome association studies and analysis through microarray. The result of these analyses was maintained through different online repositories and reported in scientific and research journals such as the PubMed of the National Center for Biotechnology Information (NCBI) and other life sciences, bioinformatics, and genome science journals [10]. MEDLINE is the journal citation database that has 25 million references. PubMed has over 28 million citations of biological articles and ever increasing every year. While PubMed Central is the full-text journal articles has over 3 million articles from PubMed [11,12]. The IR

Revised Manuscript Received on July 18, 2019.

Joseph M. De Guia, School of Information Technology, Mapua University, Muralla St., Intramuros, Manila, Philippines. (E-mail: jmdeguia@mapua.edu.ph)

Madhavi Deveraj, School of Information Technology, Mapua University, Muralla St., Intramuros, Manila, Philippines. (E-mail: mdevaraj@mapua.edu.ph)

Retrieval Number: I11090789S219/19©BEIESP
DOI : 10.35940/ijitee.I1109.0789S219

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

systems in biomedical and genome research are based on the process and forms where the main goal is to find the content.

This paper surveyed researches of IR models and its applications relevant to biomedical and genome research. The literature in IR and related processes and tasks in biomedical and genome studies were analyzed using Cited Reference Explorer (CRE). The information is based on the growth of research journals or articles related to IR published in PubMed/Medline was used as the source using the relevant terms available in the title, keywords, and abstract. The selected papers were published on 2000-2017. The paper also presented the introduction to IR theory, models, concepts, applications and summary of the papers published in this knowledge area. In this paper, we presented the different research survey, reviews, concepts, and tools in IR applied to biomedical and genomic research in section 2. The survey presented the evaluation and next steps described by the key research in section 3. In section 4, an evaluation summary of the IR systems, tools, and tasks are presented. Finally, in the conclusion described advances and current state-of the-art contribution made by the researchers and the next steps to look at the future of IR technologies.

II. RELATED WORKS AND DISCUSSION

2.1 The Current State of Information Retrieval in Biomedical and Genomic Research

The growth of the published journal articles and related information in PubMed/MEDLINE is exponential. In this current state with the increasing number of publications that is too large for researcher to read and to catch up it is difficult to determine what is current and the state-of-the-art results in a certain discipline or domain. It is essential for researcher to exploit automated text search in the literature to discover the interesting point of retrieving and extracting what has been published. In Fig. 1 shows the number of published articles and journals and with the rate of increasing publication in Information Retrieval applied to biomedical and genomic research. Published journals and articles from 2000 to 2017 are shown. Using the Cited Reference Explorer (CRE) it shows the cited references of the papers considered on this study against the publication year (2010-2017) as referred as well in PubMed. It shows that the peak in the given period was 2016 where it has more than 12,000 papers have interest in IR systems in biomedical and genomic research.

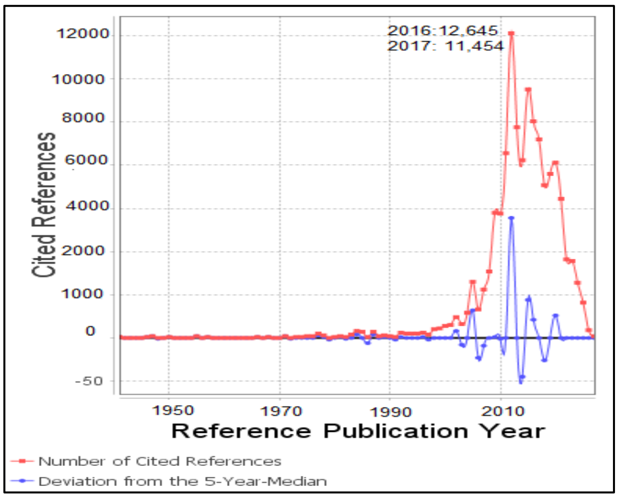

Figure 1. The number of published papers and the rate publication in IR domain from 2000-2017 indexed in PubMed.

2.2 Overview of Information Retrieval in Biomedical and Genome Research

Information Retrieval (IR) in biomedical and genomic research deals with the text retrieval from many of the published literature in PubMed, PMC, MEDLINE, and other repositories. However, there are new categories of content that has significance in all aspects of research. The IR systems now can search and retrieve images, video, gene and protein sequences, chemical structures, etc. The useful way to classify the information in terms of its structure and function of content. The content can be divided into bibliographic, full-text, annotated, and aggregated formats. The bibliographic content also known as literature references such as MEDLINE which contain references in biomedical articles, editorials, and letters to the editors. There are a number of fields such as title, abstract, indexing terms, and other identifying attributes. Other resources similar to MEDLINE are EMBASE (European MEDLINE) for non-English journals. Similarly, web catalogs include HealthFinder, HON Select, Translating Research into Practice (TRIP), and Open Directory, all related to health and clinical information. Another content category is full-text which has the full text content and supplementary data such as results in tables, figures, images, raw data that can have hyperlinks pointing to other information in the web. The available web resources, such as Ovid and MDConsults, Stat-ref, and online Mendelian Inheritance in Man (OMIM) which list the information about genomic cases of disease. Similarly, online repositories and websites of cancer and related diseases such as the National Cancer Institute (NCI) and Centers for Disease Control (CDC) that aggregates related information about the disease. In addition, there are known health websites for consumers and body of knowledge from the American Health Information Management Association (AHIMA). Third category is annotated content that are stored in the database. The database content can also be divided into: images, genomes, citations, EBM, etc. Some examples of these databases include Visible Human, Lieberman eRadiology, WebPath, Pathology Educational Instructional Resource (PEIR), DermIS, VisualDX [13]. The databases in genomics are also available in Nucleic Acid Research (NAR) catalogs such as Molecular Biology Database Collection. The citation database is also available in the Science Citation Index (SCI) or Web of Science, Scopus, Google scholar, and CiteSeer. Other evidence-based medicine (EBM) are Cochrane, Clinical evidence, UptoDate, infoPOEMS, and ACP smart medicine, etc. [14]. Finally, the aggregated content category deals with all types of contents. The MedlinePlus contains topics about health, medicines, dictionaries, directories, etc., linked to PubMed. Another resource is the MerckMedicus developed by a pharmaceutical company and publisher for physicians in the US. Refer to Table 1 for the available IR system search engine.

Table 1. IR system available search engine

IR system	Corpus	Type	Methods	Source
XploreMed	MEDLINE	Research	Explores bibliographic MEDLINE searches	http://www.bork.emblheidelberg.de/xplormed
PubFocus	PubMED	Research	Ranking journal citation	www.pubfocus.com
BioMed Search	PubMED, MEDLINE	Commercial	Retrieval with clustering	www.biomedsearch.com
Textpresso	C. Elegans papers	Research	C. elegans literature information retrieval and extraction tool	www.textpresso.org
MedMiner	PubMED, GeneCards	Research	Extraction of sentences relevant to genes	https://discover.nci.nih.gov/index.jsp

2.2.1 Indexing Methods and Applications

The basic concept in the classic information retrieval is the representation of the keywords or text in the document such as the index term. The other key terms are words or terms in a specific document. Another representation is the concept indexing that identifies key terms and phrases that maps to a vocabulary. The term index and vocabulary makeup the other document indexing such as term-document and term-frequency. The term-document records the occurrence of the terms appear in the document. While the term-frequency is based on the observation as a term-weight where how many times each distinct term occurs in the document. The full text logical view of the document presents challenges to poor search experience from full set to a set of index terms. These representative keywords or terms such as stop words, stemming, accents, spacing, and noun groups to form the controlled vocabulary. However, in the domain specific vocabulary or thesaurus of biomedical knowledge of the Unified Medicine Language System (UMLS) [15] the IR models can also pose a challenge. The semantic meaning and categories of the index terms is more extensive in biomedical domain specific area. Other issues are practices in the biomedical knowledge that represents different granularity of the terms when compounded matching a term in the thesaurus with multiple meanings, using a set of acronyms, and the negated concept of term from a medical narrative. Another approach of the indexing tasks is applied to controlled terminologies. The terminologies can be applied into three types: hierarchical, synonym, and related terms. The MeSH terminology uses canonical headings and terms that contain the three terminology types. The MeSH terminologies can also link to UMLS and its sources, Metathesaurus, semantic network, and specialist lexicon [15]. The rules for the canonical form for the MeSH headings is used for this concept. Each of the headings has the concept unique identifiers. These identifiers are linked to the terms, strings, and the atomic instance of the string in the vocabulary. In addition, the synonyms or relations of each concepts, terms, and string can also have no-relation to the word index.

In database search using keywords or terms as part of the query the issue is that if the users will find the result of the search relevant and non-relevant. The relevance ranking is a very useful method to provide the best match and ranked highest in the search result. Relevance ranking use document vector approach. This approach was being used by PubMed and MEDLINE search engine for key biomedical and genomic topics [14,15]. Similarity measures in document vector was also derived on the common terms accessed from several databases. Similarly, clustering of the document search which are similar in the database use document-vector approach. The IR integration in traditional database management systems or RDBMS is available. The degree of integration in the SQL data manipulation allowing Boolean and relevance ranking can be specified in all indexed term or fields in the database. Implementation of specialized content in genome in several databases on the web such as NCBI contain mostly of textual content, hypertext mark-up language (HTML) pages and hyperlinks. Online databases use search engine that index web pages and facilitate full-text search as well as approximate searching. The manual indexing is another challenge for bibliographic, full-text, and annotated content resources. The index can have two or more headings and subheadings of document and text attributes. Similar in a book index, web index can be a form of a catalogs and aggregations. The manual indexing in web can be deployed using the Dublin Core Metadata Initiative (DCMI) [16] another method is using Open Directory Project [17]. The DCMI was used as the standard indexing in web as approved by the International Organization for Standardization (ISO). In the web metadata information, the standard catalog used is resource description framework (RDF). In the standard web interchanging metadata RDF is in the form of Extensible Markup Language (XML). In the automatic indexing implemented in the IR systems, is a combination of human and word indexing. In this method, the term weighting based on the product of inverse document frequency (IDF) and term frequency (TF). This method was used in the TREC tasks which probabilistic theory in expression. Another method is using the semantic-equivalence of words in documents or the latent semantic index (LSI) using the singular-value decomposition (SVD). Language modeling was also introduced and used in another TREC tasks by [18] and improved by [19] for smoothing processes. Similarly, another improvement with the topic mapping to UMLS Metathesaurus was introduced in the TREC Genomic Track. The link method in commercial application used by Google search engine is the PageRank algorithm. This algorithm put weights on the number of other pages that has a link of the cited pages. Some illustrative examples of these implementations available to researchers are online databases such as Gene Expression Omnibus (GEO) Database, GEO Datasets, GEO Profiles and other useful related databases in the NCBI portal [11]. Refer to Table 2 for the IR system knowledge-based resources.

Table 2. IR system available knowledge-based resources

IR system	Domain	Contents	Characteristics	Source
UMLS	Biomedical content	Million terms	Uses upper ontology with SNOMED, MeSH, GO	www.nlm.nih.gov/research/umls
GO	Genetic terms	Thousand terms	GO terms mapped with MeSH	www.geneontology.org
MeSH	Biomedical terms	Thousand terms	Used for automatic indexing and manual indexing of MEDLINE	www.nlm.nih.gov/mesh

2.2.2. Retrieval Methods and Applications

In the retrieval methods, there are two approaches being deployed into IR systems such as exact-match and partial-match search. The exact-match search provides the user to all the possible documents that exactly matches the criteria of the search using Boolean search method. The exact-match retrieval uses the primary and secondary content for the search terms and attributes in combination with the Boolean operators, AND, OR, NOT. The partial-match method is simply using simple terms or few terms for search and retrieval tasks. This method uses the natural language query or the vector-space model that rank the closeness of the query in the documents relevant to the user.

IR systems in this area, as deployed by NLM is PubMed system using automated tools (PubMed Tools) such as MeSH, citation matcher, topic-specific queries, BioSample, Assembly, Gene, etc. The PubMed system allow basic search and provides content that are close to the query of the user using the lexical-stat retrieval method. A specialized query can be performed using advanced search and search builder. Similar to Google PageRank, PubMed has the ability to rank and sort search results. This process also utilizes the relevant term weighting method of TF and IDF and the temporal attribute of the document. When the document of interest has been identified it is presented in text format HTML and PDF for reading and printing. Other tools available are the links to articles and citations, notifications, and responses to the author through auto-email. In Table 3 this illustrates the IR system and available tools suing supervised learning from tagged corpora, statistical rules, and NLP tasks.

Table 3. IR system available tools and resources

IR system	Entities	Type	Method	Source
AbGene	genes, proteins	Research	statistically extracted rules	ftp.ncbi.nlm.nih.gov/pub/tanabe
MedBLast	proteins	Research	Document retrieval	http://medblast.sibsnet.org
AliasServer	proteins	Research	Proteins alias handler	http://cbi.labri.fr/outils/alias/index.php
PubCrawler	text, terms	Research	Text and document alerting and updating	http://pubcrawler.gen.tcd.ie/

2.3 Information Retrieval and Text Mining Applications

The discovery of interesting knowledge from a dataset is data mining process. The search and patterns of interesting knowledge from the data sets is applied to any data mining programs. Data mining apply different algorithmic methods or machine learning to uncover the hidden patterns of interesting knowledge. These algorithms include common machine learning techniques such as supervised and unsupervised learning and statistical pattern. Data mining explore and derive information from different data and its insights. In IR, text mining involves the identification of a large and structured set of texts in any file system, databases, the web on any content repositories for analysis. The application of machine learning techniques to develop a model and structure of the information of text of interest as described in [20]. In addition, natural language processing (NLP) is an essential task working in combination with text-mining. In the genomic studies, a number of text mining applications were discussed [21]. These applications automate the extraction of information for proteins, genes, functional relationships to domain research through published articles, journals or text documents. These text-mining applications improve access to information and knowledge for biomedical and gene research domains.

The Metathesaurus of UMLS is a MetaMap program [15] an application that maps text searched from biomedical literature. The paper described MetaMap of UMLS Metathesaurus mapping the biomedical text using symbols, NLP, and computational linguistics approaches. In the paper of [17] developed an online lexicon of annotated biomedical abbreviations in Medstract of MEDLINE. The application of text mining and NLP in the algorithm measured the performance of precision 80% and recall 83% to all MEDLINE database high scoring abbreviations in the Medstract corpus. Another annotation for chemical named entity recognition (NER) using NLP for chemistry papers corpus was described in the paper [22]. Using LingPipe [23] for inter annotator techniques for chemicals to recognize unknown words using tokenization ChemTok [24] was tested and measured with precision 67% and recall 63% [23]. GATE is software architecture for language engineering [25] a framework used in developing applications with language resources, processing resources, visual resources using text mining and NLP techniques. The architecture framework Unstructured Information Management Architecture (UIMA) [26] supports the NLP processing that can facilitate application layers through service-oriented deployments. A gene-mention finding module LingPipe used in biomedical text mining. The paper of [27] evaluated the performance of LingPipe of n-gram language model and text classification tagging extraction system with 99% recall in Hidden Marcov Model (HMM). The Biological Text Knowledge Services (BioTeKS) [28] is a UIMA system for analyzing biomedical text in MEDLINE content and repositories. This system used text mining and NLP annotators used for biomedical entity extraction and its relations into ontologies.

A system for supporting textual annotation layers for bioscience NLP processing using layered query language (LQL) in syntactic and ontology hierarchy integrated in SQL was developed [29]. A search approach for biomedical lexical ontologies identifying abbreviation and its definitions was developed by [29,30]. The algorithm was evaluated using 1000 MEDLINE abstracts against the Medstract corpus with 96% precision and 82% recall rates. A Biomedical Named Entity Recognizer (ABNER) a text mining tool using NLP for molecular biology [31]. The algorithm used for sequence labeling and segmenting of data is conditional random field (CRF) in the Machine Learning for Language Toolkit (MALLET) [32] implementation. The evaluation of ABNER system two tagging modules indicated 69% precision and recall of 72% in the NLPBA corpus [33]. While the BioCreative [34] corpus has 75% precision and a recall of 66% [27]. The results are competitive against published result of same evaluation and can be potential application to some specific corpus

. Table 4. IR system using text-mining tools and resources

IR system	**Entities**	**Type**	**Method**	**Source**
ABNER	Proteins, DNA,RNA, cells	Research	Supervised learning	cs.wisc.edu/~bsettles/abner
Lingpipe	Genes, proteins and others	Commercial	IE tool based on supervised learning	alias-i.com/lingpipe/web/download.html
ChemTok	Genes, proteins and others	Research	rule based tokenizer term classification for chemical NER	sfnet_oscar3-chem/downloads/chemtok/1.0.1/chemtok-1.0.1.tar.gz/

2.4. Information Retrieval in Microarray Gene Analysis and Applications

The advances in the large datasets of genes and its repositories available make use of genomic researches to link the genes of interest with the relevant literature. This process is great challenge for researcher to make sense of the gene and literature combined for the individual and common genes of the whole dataset. Some of the applications for large gene collections such as the microarray and its processing are presented.

Text-mining methods were also employed for genes in microarray to automatically link the information of gene to articles in databases. The application GEISHA [35] do the mining of terms related with the genes and provide a statistical analysis. Another application is the MedMiner [36, 37] a tool used for interpreting microarray data. In addition, Entrez [38] is a visualization tool to map the network of gene information. The microGENIE [39] uses semi-automatic query processing for the large genes database in PubMed. A tool in PubMed is the Gene Expression Omnibus (GEO) [40] a public repository of gene data which supports the submission, storage, and retrieval of gene expressions and genomic hybridization experiments. Another database in the NCBI is the Reference Sequence or RefSeq [41] that holds the collection of sequences of genomic data, transcripts, and proteins. Galaxy [42] another database that has the integration of genomic sequences, alignments, and functional annotations. Finally, the Bioconductor [43] is an open source software and development tool for statistical and visualization in R programming. In this software is the BioMart package or module that is designed for large scale databases such as the gene information and gene products.

The functional analysis of key biological processes and functions methods [44] were used in gene information retrieval. To improve the process of information retrieval, Integrative Functional Analysis (IFA) was added to use minimal parameter setting for analyzing the gene expression.

In mining the maximal differential co-expression biclusters, finds the reasonable patterns in a microarray gene expression for normal and cancer cells [45]. The pair of samples of microarray gene expression which produces the differential co-expression genes will create a relational graph of differential weighted undirected pattern. This pattern is to explore the pruning technique for the maximal bicluster. The DECluster can identify the differential co-expression biclusters with significant biological clusters. The performance of the algorithm was determined based on running and computation time.

The identification of the peripheral blood in the gene for the CTC in breast cancer was the interest of this study by [46]. The approach used in monitoring the CTC in breast cancer and its indicative genes were limited for isolation and detection. The investigation of breast cancer and periphery blood microarray datasets were used to identify the periphery blood that will be compared based on the two-stage procedure made from normal and cancer tissue samples. The results of the investigation were evaluated using statistical and biological comparisons (empirical Bayes) where it classifies the signatures of the normal and cancer samples.

In querying gene expression records where volume of gene expression data is available slow response time was the result of the query and database operations. In [47] developed a data model for the high dimensional microarray data to optimize the database performance and scalability using Oracle. Using a key value storage HBase this schema resulted to three times decrease in retrieval time of the high dimensional data and optimizing the performance of the data store for gene expressions.

Discovering the gene-disease relationship is a challenge for bioinformatics. The approach of [48] using GeneWizard tools were explored to realize the gene-disease relationship. These tools were explored to generate the hypothesis from microarray data and its relationship with the published literatures. In Table 5 it lists the summary of IR systems available for microarray, gene analysis applications.

Table 5. IR system in microarray gene analysis applications

IR system	Entities	Type	Method	Source
GEISHA	Genes	Research	Text-mining tool to assist microarray analysis	http://www.pdg.cnb.uam.es/blaschke/cgi-bin/geisha
MedMiner	Genes	Research	Extraction of sentences relevant to genes	http://discover.nci.nih.gov/textmining/main.jsp
PubGene	Genes	Research	Text-mining tool for microarrays	http://www.pubgene.org/
MicroGenie	Genes	Research	Text-mining for microarrays	http://www.cs.vu.nl/microgenie
GEO	Genes	Research	Platform, sample, series for gene expressions and hybridization data	http://www.cs.vu.nl/microgenie

2.5. Information Retrieval uses in NLP Applications

Natural Language Processing (NLP) is a processing of human language by a computer. In the application to biologist, biomedical, and genomic research growth is essential to perform this systematic survey of the large-scale information in research journals, databases of genes and proteins. NLP application is a challenge for most of the researchers and developers that make sense of text and documents analysis contained on these resources in biomedical and genomic domain. Some of the challenges are in the area of morphology, syntax, and disambiguation for the entity or terms for genes, proteins, chemicals, reactions, networks, ontology, etc. Some of the applications of NLP in this domain includes the automation of searching published literature, gene annotation and extraction, database curation and population.

Mining biomedical and genomic related literature and data for gene-disease relationship use text mining and NLP approaches to discover the knowledge in biomedical and gene research. A survey of NLP techniques in bioinformatics [49] provided the different tools and algorithm techniques to solve specific problems in bioinformatics. The prediction of protein-to-protein interaction, gene-disease relationship can utilize text mining and NLP. The survey explored the text mining algorithms and noted that the recent text and NLP techniques for bioinformatics should be verified. Utilizing NLP techniques in semantic annotations and ontologies in searching biomedical researches were explored by [50]. Information extraction techniques using text and NLP can result to the discovery of targeted biomedical resources and enhances the accuracy of service discovery. The framework developed [51] was able to answer clinical questions with the tools available utilizing NLP techniques with ontologies. The evaluation of the framework from the experiments performed in the web-based tool European Union Adverse Drug Reaction (EUADR) [52], having precision (40%) and recall (73%) respectively; using the NLP framework precision (100%) and recall (14%); Free text precision (25%) and recall (59%). In the second test, NLP framework has 100% precision and 17% recall.

A survey of NLP in a system architecture framework was described by [51]. The NLP techniques for biomedical research were explored and compared. The general architecture incorporated the two main components; background knowledge and framework (method, tools, system). The general architecture is conceptually and was explored in the UMLS using rule-based approaches NLP. In biomedical research, application and in practice, rule-based NLP was easy and effective and can be adopted.

In the molecular image domain described and evaluated BioMedLEE using NLP, automated information extraction. The lexicon of BioMedLee was extended including organisms and cell lines and was evaluated based on the following categories: process, probes, phenotype, gene-gene product, imaging instrument, organism, and cell lines. The performance was evaluated using precision and recall of HT-29 cells were 70% and 74.5%. In the BioMedLEE was adopted and developed a system using the NLP engine for clinical information extraction. The genotype-phenotype was extracted in the BioMedLEE and the evaluation showed a precision of 64% and 77% recall against the experts in reporting the phenotype information.

The use of NLP in genomic research, information retrieval, extraction and semantics are fundamental in the biomedical literature and genome data databases where NLP techniques were applied. The challenges of discovering the associations and relationship of biological text and sequences of data from different biological domain and gene expression data explored ways to integrated bio-NLP from the traditional NLP techniques. This integration of algorithms and toolkits available for knowledge discovery in different genome was explored.

2.6. Information Retrieval and Datamining Applications

There has been a rapid wealth of information and advances in data mining to discover the unknown and knowledge in the area of biomedical and genomic research. The data lakes of genes and proteins as well the network of its activity needs to be managed and new method of discovering knowledge is essential for data mining experts. The integration of the databases and data analysis processes has been formalized using a variety of tasks through ontologies, annotations, and the improved workflows for classification, regression, and clustering methods. The surveyed papers are focused on the cancer genome and the data mining tasks deployed on how data management, analysis, and discovery were made. The integration of these processes as part of the IR and data mining tasks is essential for the identification and classification of cancer genome for the intelligent data analysis and knowledge discovery systems.

The paper of [56] reviewed the different dataset available in The Cancer Genome Atlas (TCGA) and tools in analysis. This system was also developed to analyze the dataset available in TCGA. There are many online portals available to describe the analysis with visualization of the cancer datasets retrieved from TCGA. Some of the illustrative web tools in analyzing the TCGA datasets presented were Bioportal, GDAC, canEvolve, PROgeneV2, UCSC Cancer browser, UZH Cancer browser. The comparison was presented by [58] and based on the visualization, differential gene analysis, and interpretation according to the gene selected. Mining the cancer genome [59] presented the cross-comparison of gene data with the tumor suppressor in lymphoma sample (EPHA7) and the screening process. The paper also described the importance components of cancer gene discovery through the cancer genome analysis and unbiased genetic screening. The cross referring will infer the functional filter for the gene data and its relevance to the disease.

COSMIC is a public portal that curate information about the somatic mutations in cancer [58]. The portal having a database of genome-wide somatic mutations of cancer samples with codes and annotations. The portal provides a browser and the visualization of the cancer sample analyzed.

In the gene function discovery was described and how this can be resolved by genome-wide high-throughput (HTP) techniques. The genome-wide experiments were made based on the predicted genes in a genome sample. The review of the databases that integrate HTP data for gene-protein-linked – omics; single or multi-organisms – omics. The grand challenge for the gene and function discovery was still at large and the integration of the datasets was also questioned without the standardization.

In the ngs.plot developed visualized the patterns of DNA and sequence of data using the NGS technology in epigenomes. The tool develop was described to fill the gap between data and information to protein regulators functional elements and the phenotype outputs.

2.7. Information Retrieval in Content-based systems

Information retrieval in content-based systems learning the textual relevance of documents to many of the images and other media formats. The learning models were deployed in many text processing and multimedia applications that is doing text and as well as computer vision and image processing that uses vector representation of text, image, and speech formats.

In optimizing the image retrieval system evaluation is top precision whereas the similarity function method was nonexistent and therefore can fill this space. In their proposal, a similarity learning model was used in the experiment for the image database. The maximum top precision similarity (MTPS) learning method has precision higher than 0.18 against state-of-the-art algorithms.

Image retrieval used visual saliency model and their proposed multi-feature fusion: cognitive load and cognitive level complexity classification in combination with the group sparse logistic regression model. This process improved the accuracy of image annotation and the extracted feature to describe the image more effective. Finally, combining these as process method enhance the overall performance of the image retrieval system. The experiment performed confirms the effectiveness of the model being proposed where the image database precision and recall registered an increased performance in the retrieval system compared with other systems.

III. INFORMATION RETRIEVAL SYSTEM AND USER EVALUATIONS & RESULTS

It is essential to determine if the IR system or tasks achieves the objectives, performance, and accuracy of the results. A methodology that focused on research repeatability, comparability, and viability of the results are also part of the evaluation. There are two methodologies to system evaluation used in the KDD challenge cup and TREC. In KDD involves the development of reliable valid measurement and metrics with the test collection. While in TREC an ongoing large-scale evaluation use and generate reusable test collections with the exemplar of ideas, exchange and adoption of best practices, and the platform of technology transfer. Another method of evaluation focused on the users and the information needs of the researchers or the user-centered evaluations.

The most widely used evaluation measurement for most of the systems are recall and precision. In IR systems, recall is the number or fraction of all relevant documents retrieved from the database or document repository. While precision is the number or fraction of all retrieved documents that are relevant. In relation to ranking the mean average precision (MAP) which is the average of every relevant document retrieved. The best-known challenge evaluation is exemplified in TREC. The IR community and experts focused on improving the effectiveness of IR systems. There are well known IR system evaluations from TREC series that includes: General IR tasks Genomic track [67] improvement of the MEDLINE, Medical Records Track, etc. Come of these evaluations are focused on tasks in classifying documents, tagging genes and proteins, recognizing entities and relations, retrieving relevant medical images, etc. Most of the IR systems has the precision rates of 70-90% while the recall is 70%. In each of the TREC series, it was noted on its objectives and domains evaluated for every task it has varying rates of precision and recall. In the biomedical and genomic domain, there is 13% drop in the test accuracy in the evaluation of tagged training corpus with training. In other tasks that makes tagging difficult, term characteristics and occurrence with annotators has 75%-90% agreement with genes and proteins.

Retrieval Number: I11090789S219/19©BEIESP
DOI : 10.35940/ijitee.I1109.0789S219

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

IV. CONCLUSION

This paper explored the application of IR in searching, text processing, data mining, and application of NLP techniques in biomedical and genomic research. The surveyed papers presented the models, processes, evaluation, and state-of-the-art outputs. Despite these successes in IR systems there are still open challenges from the vast ocean of data that increase every day. The growing information needs for searching, storing, processing, and transmitting in the available technology and infrastructure platform still involve the better use and improvement of IR systems and its users. There are signs of progress and some papers described as trends of IR applications in the following tasks in text mining, summarization, data mining, coreference resolution and normalization, question answering, image analysis and learning, and systems evaluation.

The genome data, research journals, and applications built on the key processes of bioinformatics increase its presence and use by life science researchers and in the biomedical studies and genomic research. The vast amount of data and information available in different databases available for access, retrieval, analysis, processing, and transforming into new knowledge brought about the plethora applications developed. The mature text mining, natural language, pattern recognition, and machine learning approaches facilitate the discovery of new genome, protein, sequences, diseases related to aid in the prognosis, diagnosis and treatment. The advancement of IR systems and its technologies was evident as presented in this paper producing hundreds of possibilities to integrate, architect, and develop a unified system and can be standardized to help the researcher in finding the answers to each questions, hypothesis, and proof to improve the knowledge in the molecular level, and well-being of human and life progress.

This survey paper is only highlighting the very available journals retrieved from PubMed and other sources there are more beyond the IR topics and its related systems domain. It is recommended to add more knowledge in this survey paper with the recent work to be presented in the future. It is a very active and well researched domain with a community of experts and professionals that will continue to contribute towards progressive future of IR and its related field.

V. DECLARATIONS

Ethics approval and consent to participate
Not applicable
Consent for publication
Not applicable
Availability of data and material
Not applicable
Competing interests
Not applicable
Funding
This research is supported by Mapua University DRIVE institutional research initiative.

VI. AUTHORS' CONTRIBUTIONS

The content of the research was formulated and investigated by the first author (De Guia) and the second author (Devaraj) provided the research advice, review and manuscript improvement feedback.

VII. AUTHORS' INFORMATION

Joseph M. De Guia is a PhD student from Mapua University, Manila Philippines. He is currently a faculty member of the School of Information Technology of same university. Joseph finished his master’s degree in Information Technology (MSIT) at Carnegie Mellon University and Computer Science (MSCS) at Mapua University. His research interests are data mining, machine learning, health and bioinformatics, information security, big data analytics, IT infrastructure and digital innovation. His research papers in digital health and enterprise architecture has been presented in international and local conferences.

Dr. Madhavi Devaraj received doctoral degree in computer science from Dr. A.P.J Abdul Kalam Technical University, Lucknow, India. She has also completed master’s in computer applications and MPhil in Computer Science from Madurai Kamaraj University, Madurai, India. Currently, she is distinguished professor in computer science department at Mapua University, Manila, Philippines. She has been assistant professor in computer science department at Invertis University, India and Babu Banarasi Das University, India, previously. Her research interests include Text analytics, Scientometric Analysis, Opinion Mining, Sentiment Analysis, Information Extraction, Neural Networks, Artificial Intelligence, Machine Learning and Big Data Analysis.

VIII. ACKNOWLEDGEMENTS

This work is supported by Mapua University through the Directed Research for Innovation and Value Enhancement (DRIVE) institutional research program.

IX. REFERENCES

1. Salton G. (1989). Automatic Text Processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley; Reading, MA.
2. Van Rijsbergen CJ. (1979). Information Retrieval. Butterworths; London, UK.
3. Baeza-Yates R., Ribeiro-Neto B. (1999). Modern Information Retrieval. Addison-Wesley Longman; Harlow, UK.
4. Witten IH., Moffat, A., Bell, TC. (1999). Managing Gigabytes. Morgan Kaufman; San Francisco, CA.
5. Sanderson, M., Croft, WB. (2012). The History of Information Retrieval Research. Proceedings of the IEEE, vol. 100, no. Special Centennial Issue, pp. 1444-1451. http://doi.org/10.1109/JPROC.2012.2189916
6. NIST. US Commerce Department. Text Retrieval Conference (TREC). Accessed on June 15, 2018 from https://trec.nist.gov/
7. Liang, R.Z., Shi, L., Wang, H., Meng, J., Wang, J.J.Y., Sun, Q. and Gu, Y. (2016). Optimizing top precision

performance measure of content-based image retrieval by learning similarity function. In Pattern Recognition (ICPR) 23rd International Conference, pp. 2954-2958.

8. Wang, H., Li, Z., Li, Y., Gupta, B.B. and Choi, C. (2018). Visual saliency guided complex image retrieval. Pattern Recognition Letters.
9. Wang, J., Wang, H., Zhou, Y. and McDonald, N. (2015), October. Multiple kernel multivariate performance learning using cutting plane algorithm. In Systems, man, and cybernetics (SMC), 2015 IEEE international conference, pp. 1870-1875.
10. [Wang, H. and Wang, J. (2014). An effective image representation method using kernel classification. In 2014 IEEE 26th international conference on tools with artificial intelligence (ICTAI), pp. 853-858.
11. Reinsel, D., Gantz, J., Rydning, J. (2017). Data Age 2025: The Evolution of Data to Life-Critical Don’t Focus on Big Data; Focus on the Data that’s Big. IDC White Paper by Seagate. Accessed on Sep 2018 from https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf
12. Statista, The Statistics Portal (2018). Volume of data/information created worldwide from 2005 to 2025 (in zetabytes) Accessed on Sep. 2018 from https://www.statista.com/statistics/871513/worldwide-data-created/
13. National Human Genome Research Institute (2016). An Overview of the Human Genome Project. Accessed on June 15, 2018 from https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/
14. Stewart, B., Wild, C. (2014). World Cancer Report. International Agency for Research on Cancer. WHO Press. Retrieved on June 15, 2018 from http://publications.iarc.fr/Non-Series-Publications/World-Cancer-Reports/World-Cancer-Report-2014
15. Omics International (n.d.). Cancer Genomics. Journal of Clinical and Medical Genomics. Accessed on June 15, 2018 from https://www.omicsonline.org/scholarly/cancer-genomics-journals-articles-ppts-list.php