A Document Ranking Approach based on Weighted-Gene/Protein in Large Biomedical Documents using MapReduce Framework

[1] K.S.S. Joseph Sastry, [2] Venkata Daya Sagar Ketaraju
[1] *Dept of Computer Science and Engineering*, [2] *Departement of Electronics and Computer Science Engineering*
Koneru Lakshmaiah Education Foundation, Guntur, A.P, India.

***Abstract* - As the size of biomedical documents increases, automatic gene or protein based document indexing and ranking models becomes increasingly important on large biomedical databases. Traditional biomedical single entity based document indexing and ranking models restricts search space on high dimensional feature space. However, traditional biomedical document ranking models do not find and extract the relevant documents using the genes or proteins. Also, traditional ranking models are not efficient to rank the biomedical documents using weighted genes or weighted proteins on large biomedical repositories. We focus on the problem of indexing, extracting and ranking biomedical document sets using gene or protein entities on large databases. In the proposed model, a novel MapReduce based natural language processing framework is designed and implemented on large biomedical databases using weighted gene or protein measures and document ranking score. Experimental results show that the proposed model has high contextual ranking accuracy, less search space and time consumption compared to the traditional biomedical document ranking models.**

Keywords – Document Ranking, Weighted Gene Protein, Large Biomedical Documents, MapReduce Framework

I. INTRODUCTION

Information Extraction (IE) in the biomedical domain is the process of extracting associations between biological entities in document sets. Biomedical text mining is the process of extracting useful and essential knowledge from the biomedical databases for decision making. Most interesting patterns that are extracted from biomedical repositories are Protein-Protein Interactions (PPIs), gene-protein, gene-disease and functional protein annotations. A large number of standard biomedical repositories are used for text mining to improve the ranking efficiency. Therefore, an efficient distributed ranking method is required to enhance the true positive accuracy and ranking performance on large biomedical repositories. Biomedical repositories such as PubMed, Medline are the most popular databases in the field of life sciences and biomedicine. The articles of this database are growing exponentially year by year. Most of these data sets are also interlinked to DBpedia. Still, there is no proper integration of unified medical language system to the gene or protein based document ranking models in clinical text processing.

Natural Language Processing systems are used to identify, extract and encode information of the historical biomedical data for knowledge analysis. Most of the biomedical tools such as MPLUS, MEDSYNDICATE, MetaMap, SemRep, MedLEE, BioMedLEE are used to find and extract the essential entities in the biomedical repositories. BioMedLEE is specific type of MedLEE that usually emphasizes on the process of extracting and structuring of biomedical entities. Additionally, relations in biomedical literature involving phenotype and genotype information are the most important part. MedLEE applies the unified medical language system codes to the basic theories. These biomedical document extraction systems allocate codes from both the unified medical language system and from different resources involving gene or protein entities or ontologies. Again, it involves the process of clinical text analysis and extraction. Apart from these, there are some other important systems like MetaMap Lite, medical language extraction and encoding system. All of these systems have responsibility to extract different kinds of clinical information.

The purpose of natural language processing in biomedical rank search is to identify documents relevant to the user's query. The ranking models exploit the document statistics such as document term frequency, inverse document term frequency, inter-document term frequency, intra-document term frequency, etc, for efficient query processing. Traditional document extraction models are based on MeSH keywords, causing the keyword matching problem in feature extraction process since medical documents are organized at the domain knowledge and concepts. Also, most of the current text extraction models are designed and implemented to extract bi-gram word terms, such as "Cancer disease". But they fail to discover longer keywords such as "Colon based cancer disease". Let us consider an example of Drosophila. Numbers of different genes are termed after some phenotype of mutants and they are known as "white-w", "shaggy-s" and "mind the gap-mtg". Apart from this, it is very complicated to discriminate gene names from protein names. Another issue is the word limitation for the articles. An example can be mentioned here that is, "HZF-7" vs. "HZF-7 protein".All of the name entity recognition of biomedical document extraction process is categorized into two types, they are:-

1. Dictionary-based NER:- A dictionary usually consists of large numbers of words which represent

examples for a particular entity class. It can be constructed with the help of databases. Pattern matching algorithms are implemented to perform matching against names in the key phrases or MeSH terms.

2. Rule-based NER:- Rule based techniques usually dependent on the constraints on the entity name of the biomedical genes or protein structures. These rules or constraints are used to construct rules in order to distinguish different entity classes in biomedical document extraction process.

As the size of the data increases, document search space and feature space increases in sequential ranking models. Therefore, it is essential to implement an efficient and scalable document ranking and feature extraction approach for highly distributed biomedical repositories using big data framework. The 'unstructured' and 'uncertainty' problems are seen in many domain fields such as biomedical repositories, biomedical databases, web mining, health care system, education and technology-intensive companies. Information extraction from biomedical repositories and analyzing this information with an experimental study are time-consuming and require an efficient feature selection and ranking models. Nowadays, text mining is used to answer many different research queries, ranging from the biomarkers, gene discovery, gene-disease prediction and drug discovery from biomedical repositories. Due to the heterogeneous nature of data format, the automatic ranking and feature extraction are not trivial. As a result, text mining has evolved in the field of biomedical systems where text mining techniques and machine learning models are integrated using high computational resources. A large number of document preprocessing techniques have been implemented in the literature on data repositories, which are responsible for transforming the raw information into a specific structured format.

Document pattern mining automatically detects the similar documents using statistical measures on term frequencies, phrase frequencies and sentence frequencies. The majority of the document pattern mining techniques are centred on the feature vector spaces, which are broadly used to train document model for text pattern mining. The similarity between sentences/documents is examined using one of document similarity measures that are based on such a feature vector or word frequencies, for instance, Jaccard measure and the cosine measure. Document ranking is currently a major challenging issue, particularly in areas such as recommender systems, information retrieval, user navigation interests and personalization. Traditionally, text ranking techniques represent documents either in term frequency or document frequency format. Hence, the selection and ranking of features must be carefully done. A large number of ranking models have been proposed in the literature using the relevant terms with limited document features.

II. LITERATURE REVIEW OF RELATED WORKS

X. Chen, A. et.al, developed an open source framework known as DataMed to extract the entities in biomedical datasets [1]. The major objective of this research work is to detect relevant datasets for the process of data reuse. As the size of the biomedical document sets increases, this model require new optimal document ranking model to find and extract useful document from the biomedical databases. DataMed tool is used to index, rank and search entities in different kinds of biomedical datasets. This system contains two important components, those are:-

1. Data ingestion pipeline:- It is responsible to gather and transform actual metadata related information for a model construction. This predefined metadata model is known as DatA Tag Suite (DATS).

2. An advanced search engine:- This search engine is responsible to detect relevant datasets according to the user-specified queries.

S. Sohn, et.al, proposed a model to find disease related patterns on clinical document variation and NLP system portability [2]. The main objective of this research work is to assess clinical disease variations with the help of various electronic medical record systems. Document variations corresponding to asthma among two cohorts are considered to rank the biomedical document sets.

J. Cuzzola, et.al, implemented a model on biomedical databases using UMLS and DBPedia to discovery the association between the document sets [3]. The main objective of this model is to map the unified medical language system entities to DBPedia resources. A common ontology relations from the Simple Knowledge Organization System and Resource Description Framework Schema are used to map and rank the UMLS and DBPedia entities.

E. S. Chen, et.al, proposed an automated disease acquisition of disease-drug knowledge from biomedical documents [4]. The main objective of this research work is to explore and automate the knowledge acquisition in case of biomedical and clinical documents. An integration of text mining and statistical approaches are implemented in order to detect the disease-drug relationships. Two different natural language processing systems are implemented here, those are:- BioMedLEE and MedLEE. These two natural language processing systems have the responsibility to detect all disease and drug entities. The outcome of this method contain ranked lists of disease-drug pairs.

E. Soysal, et.al, implemented a toolkit to find and extract the biomedical documents using the natural language processing pipelines [5]. Traditional clinical NLP tools such as MetaMap and BioNer are used to find key entities from clinical databases. These toolkits are not efficient on large biomedical databases.

K. Raja, proposed a hybrid named entity tagger for tagging human proteins/genes [6]. It is very much important to extract gene or protein names in case of biomedical text.

A large numbers of different taggers are present for the named entity recognition process, but none of the existing models is good for tagging human genes or proteins. Therefore, a hybrid tagger is necessary in order to tag human genes or proteins using UMLs ontologies.

Information about gene, protein and its pathways is analyzed using the MedScan toolkit. The MedScan is a three-tier knowledge extraction system based on a biomedical document parsing model. In the first tier, preprocessor module aimed to tag various biomedical MeSH terms using domain specific concepts. Preprocessor module reads the biomedical XML format of a MEDLINE abstract and parses into individual terms or sentences. In this module, a protein name dictionary is used as a training dataset to filter the protein names and to select the terms or sentences containing at least one gene-protein name. In the second tier, natural language processor performs a set of semantic relationships between terms or sentence structures. It is based on context-free grammar and a lexicon parse tree for MEDLINE protein extraction. In the final tier, knowledge extraction engine acting as a domain knowledge filter for extraction key MeSH-based document information in the form of conceptual graph format [7].

Support vector machine optimization aims at constructing a separate function for domain knowledge extraction. Each document is classified using the hyperplanes and feature vector space. The kernel function used in NLProt is given as Equation 1:

$$<W,X>+c=0 \quad (1)$$

Where W, X are the hyper-plane parameters corresponding to the decision-making function f(x):

$$f(x) =sgn(<W,X>+c) \quad (2)$$

The polynomial kernel function used in the NLPlot is $K(y_i, y_j) = (w_i .(y_i . y_j) + c)^d$

To describe such problem, conditional random fields define the probability conditional distribution p(Y/X) as Equation 3:

$$\Pr ob(Y / X = x) = \frac{1}{Z(X = x)} e^{\sum_{c \in Cls} \sum_{k} W_k f_k (Y_c, X, c)} \quad (3)$$

where c is the set of document categories, Y is the set of conditional field vertices in graph c, f is the function to represent the feature extraction and can be defined as

$$f(Y_c, X = x) = 1; \text{if } X \text{ and } Y \text{ are same gene entities}$$

$$f(Y_c, X = x) = 0; \text{otherwsie}$$

Z(X=x) is the normalization factor over document terms and gene set.

A new document ranking model has been proposed using a weighted comparative model. In this model, ranking is performed using document concepts and key phrases. This approach assumes only the relative similarity between the document keyphrase ranks, but it ignores the relationship between the phrases. Cosine similarity, Euclidean distance, Jaccard coefficient, averaged Kulback-Leibler (KL) Divergence and Pearson correlation coefficient are used in traditional document ranking measures. A hierarchical based document ranking model has been proposed using local document patterns [8] . The traditional pattern recognition model based on hierarchical approach may result in noisy patterns or inconsistent patterns. So, to overcome this issue a novel model known as Instance Driven Hierarchical Ranking Approach is implemented to form a rank hierarchy without extracting the global patterns. This model first discovers the locally significant patterns by using each instance to find its nearest representative to ensure an effective balance between pattern significance and local pattern frequency.

Graph-based Document Ranking Models

A document based graphical ranking method has been proposed, which considers the document sets in different forms. A query sensitive ranking measure for graph-based feature extraction has been implemented. The traditional graph-based model considers the term frequency and gene entities to discover the neighbourhood relationships. This system considers the sentence to sentence edge weight prediction and query sensitive measure in the graph model. Another model has been presented which uses TextRank with some differences and uses the shortest path method to find the nearest feature sets to the TextRank. In the initial phase, the graph model has been built for representing the document and interconnected phrase entities are there in the graph model with meaningful relationships. A weighted graph method has been proposed using the novel approach which includes ranking both phrases and sentence ranking for document feature extraction.

The major steps involved in this methodology are:

- Combines both sentence and phrase ranking methods for similarity computation.
- A phrase or sentence rank is generated based on singular matrix factorization.
- The weighted graph model is implemented to find the sentence relationship in the documents.

This method has been presented in three phases. In the initial step, document structure is represented to every document in the document set; the structure can be represented as an undirected graph. Phrases in the document play a significant role in the sentence formation in the graph model. In the second step, each phrase ranking measure in the document is computed using the ranking technique. Finally, the maximal marginal relevance technique is used to generate the relevant summary [9][10].

III. PROPOSED DOCUMENT RANKING NEW MODEL

Figure 1 presents the overall NLP framework of the proposed model on large biomedical document sets using the MapReduce framework. Initially, biomedical documents are extracted using the web services for document pre-preprocessing. In the mapper phase, document pre-processing operations are performed on each biomedical document using tokenization, POS tagger and stop word removal. In this phase, gene or protein tagger is used to find and extract the genes and proteins in each document for document ranking process.

These gene and protein entities are used to rank the biomedical document using the computed weighted gene score, weighted protein score and document ranking measures. In this process, multiple genes or proteins are used in parallel to minimize the time and search space for document ranking. Finally, all the mappers are combined in the reducer phase to sort the gene or protein based documents. Algorithm 1 describes the biomedical document indexing and extraction process for biomedical repositories such as PubMed and Medline. Lines 1-7 describe the web service connection to biomedical data repository and documents indexing and extraction using URL and document ID. Lines 8-11 describe the gene or protein identification and extraction from the biomedical document using ABNER tagger.

```
Algorithm 1: Biomedical Data Extraction
Input : Biomedical document ID List :BList, URL
Output : Biomedical Document List : BDOC[]
1: Connect to Biomedical data repository.
2: For each biomedical id in Biomedical document BList
3: do
4.          if( id!=NULL)
5.              Document List BDOC[]=getDoc(URL,id);
6.          end if
7. done
8.  for each document d[i] in BDOC
9.   do
10.     Find gene GF[] and protein  PF[] features in each document d[i] using ABNER tagger.
11.done
```

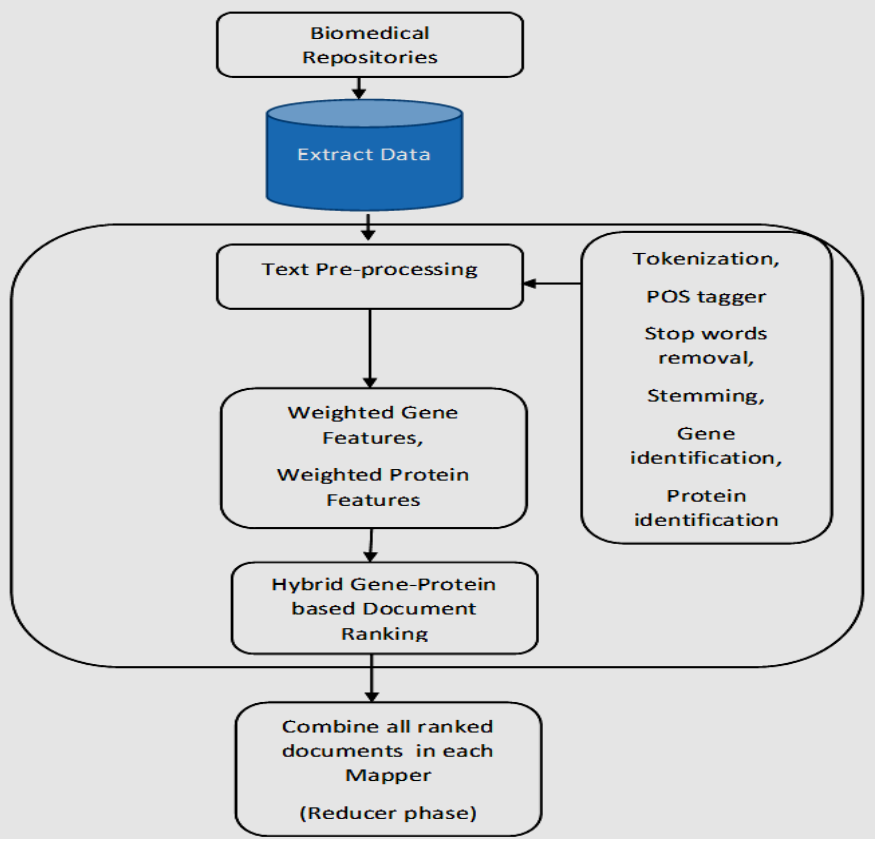

Figure 1: Proposed Map Reduce Framework of Dimensionality reduction approach for Biomedical Databases

First, all documents are tokenized into tokens for POS tagging (Stanford POS tagger was used). Part of speech tags are required by WordNet to identify synonymous words as it assumes that synonyms should have the same POS tag. After tagging the contents of documents, other pre-processing steps including tokenization, stop word removal and stemming are applied. Pre-processing steps cannot be applied prior to POS tagging which requires the original text. The next step is to search the documents with the help of WordNet for terms that are synonymous. Synonyms of a particular concept are all replaced with a representing term in the documents' bag of words. Here, multiple Hadoop mapper interfaces are assigned to each distributed data source for domain knowledge extraction. At first, Named Entity Recognition (NER) method helps in recognizing elements during the document indexing process. Numerous lexical analysis strategies have been floated to shape up a few strategies by which any entity or element can be recognized and perceived. Some of these methodologies are implemented in the territory of molecular biology to identify the names of various genes. From an NLP point of view, much uncertainty originates from different semantic phenomena, for instance, broad lexical varieties (i.e., one name has a few spelling varieties, all speaking the same idea), synonymy (i.e., one concept is represented by several names) and homonymy or polysemy (i.e., one name has many implications and stands for many notions). The basic steps in the document pre-processing are described as below:

- Tokenization: Tokenization is the initial phase of document pre-processing. Here, all the words acceptable by pattern matching algorithms are retrieved from various documents. These groups of words are acceptable to pattern mining algorithms. Several common words which do not affect the pattern extraction process are identified and discarded. Stop-word is defined as the most frequently used words which do not influence the pattern mining process. Some examples of stop-words are- pronouns, prepositions, conjunctions and so on. If the numbers of stop-words are decreased, the pattern mining process is enhanced to a great extent.
- Stemming: Stemming involves all the activities starting from suffix elimination to the production of word stems. Such words are considered to be equal; thus, different stemming approaches have been implemented to convert every word to its root. In other words, this approach is used to combine different words having similar conceptual meaning. For example, words like play, player, played and playing all have similar conceptual meaning. Hence, all these types of words are considered as a single word and the overall size of the dictionary is also reduced which requires less storage capacity and less processing time.
- Pruning: The pruning process eliminates words which are either very rarely used or most frequently used. Words that have very poor or very high frequency create several issues in pattern mining.

Algorithm 2 describes the biomedical document ranking model using the gene or protein weights. Here MapReduce framework is used to optimize the search time and document ranking. In the MapReduce framework, P mappers and reducers are used to pre-process and rank the biomedical documents. Initially, k number of biomedical documents are given to each mapper for pre-processing and ranking computation as shown in lines 1-13. Lines 7-14 represent the document tokenization, stemming, stop word removal and non-special characters removal process. Lines 16-21 represents the gene weight, protein weight and document rank computation. Lines 22-26 represents the threshold verification for gene or protein based documents ranking.

Algorithm 2. MapReduce based Biomedical Document Ranking using Weighted Gene and Protein features
Input: Biomedical documents BDOC[], Gene features GF, Protein Features PF, Number of mappers P and Reducers Q.
Output: Biomedical gene-protein based document ranking.
Procedure:

```
1.      Partitioning the Biomedical document sets BDOC[] into P mappers.
2.      For each mapper in P
3.      Do
4.          For each mapper document md[i]
5.          Do
6.                  Apply Tokenization to md[i] as BDT[].
7.                  For each token t in BDT[]
8.                  Do
9.                      If(t!==NULL)
10.                     Then
11.                         Apply Stopword removal.
12.                         Apply Stemming.
13.                         Remove non-special characters {!#$%}.
14.                     End if
15.                 Done
16.                 Compute weighted gene document frequency to each document as
```

Continued on next page

17. $$WGF(i,j) = \frac{|GF_j| * Prob(GF[i]/md[j])}{Max\{GF_j\}}$$
18. Compute weighted protein document frequency to each document as
19. $$WPF(i,j) = \frac{|PF_j| * Prob(PF[i]/md[j])}{Max\{PF_j\}}$$
20. Document Ranking score(DRS) to each term is computed as
21. $$DRS(i,j) = \frac{\{WGF(i,j) + WPF(i,j)\}}{Max\{|GF_j|,|PF_j|\}} * Max\{tf(j)\}$$
22. If(DRS(i,j)>=Threshold)
23. Then
24. Display all documents related to GF_i,PF_i from the biomedical repositories.
25. Else
26. Delete document from list.
27. Done
28. Done
29. Combine all the biomedical ranked documents in the Reducer phase.

IV. EXPERIMENTAL RESULTS

Experimental results are performed on biomedical datasets such as Medline and PubMed repositories for gene or protein based document ranking. In this experimental study, a filter based document ranking model is proposed using Hadoop framework on biomedical databases such as MEDLINE and PubMed.

SAMPLE BIOMEDICAL GENES OR PROTEINS AND ITS MAPPER INDEXING

R85482 CYSTATIN RESPONSE FACTOR (Homo sapiens)
T61609 LAMININ RECEPTOR (HUMAN);.
T62220 CALPACTIN I LIGHT CHAIN (HUMAN);.
T51574 40S RIBOSOMAL PROTEIN S24 (HUMAN).
T48041 Human RNA for the beta-2 microglobulin.
T96832 INTERFERON-ALPHA RECEPTOR PRECURSOR (Homo sapiens)
H54676 60S RIBOSOMAL_PROTEIN L18A (HUMAN);.
R86975 40S RIBOSOMAL_PROTEIN S28 (HUMAN);.
T63258 ELONGATION_FACTOR 1-ALPHA 1 (HUMAN GENE);.
T57619 40S RIBOSOMAL_PROTEIN_S6 (Nicotiana tabacum)
T88723 UBIQUITIN (HUMAN);.
R36455 NUCLEOLAR TRANSCRIPTION FACTOR 1 (Homo sapiens)
T61602 40S RIBOSOMAL _PROTEIN S11 (HUMAN);.
T58861 60S RIBOSOMAL _PROTEIN L30E (Kluyveromyces lactis)
U21909 "Human cofilin mRNA, partial cds.
T71025 Human (HUMAN);.
T51534 CYSTATIN C PRECURSOR (HUMAN).
T69026 60S RIBOSOMAL PROTEIN L9 (HUMAN);.
H68220 UBIQUITIN-LIKE PROTEIN FUBI (HUMAN);.
T61661 PROFILIN I (HUMAN);.
X57346 H.sapiens mRNA for HS1 protein.
H55758 ALPHA ENOLASE (HUMAN);.
MAPPER U19796 Value is "Human melanoma antigen p15 mRNA, complete cds.
MAPPER H87344 Value is SERUM ALBUMIN PRECURSOR (Homo sapiens)
MAPPER X06614 Value is Human mRNA for receptor of retinoic acid.
MAPPER U07695 Value is "Human tyrosine kinase (HTK) mRNA, complete cds.

For Screenshot of Gene or Protein based Document Ranking, see Appendix.

TABLE I: DOCUMENT PRE-PROCESSING METHODS FOR BIOMEDICAL DATABASES USING MAPREDUCE FRAMEWORK.

Documents size	Bioner	Bayesian Ranking	NMF	IDR	Proposed
#5k	0.8051	0.8345	0.8356	0.9159	0.9876
#10k	0.8131	0.8144	0.8382	0.9312	0.9798
#15k	0.8004	0.828	0.8179	0.9301	0.9862
#20k	0.7844	0.853	0.8519	0.9105	0.9714
#50k	0.7978	0.8443	0.8577	0.9284	0.9783

Table I presents the comparison of the proposed model to the existing models for gene or protein document ranking using the MapReduce framework. From the table, it is evident that the proposed model has high accurate gene or protein based ranking than the existing models. As the size of the document sets increases, proposed model has high computational contextual ranking compared to the existing models.

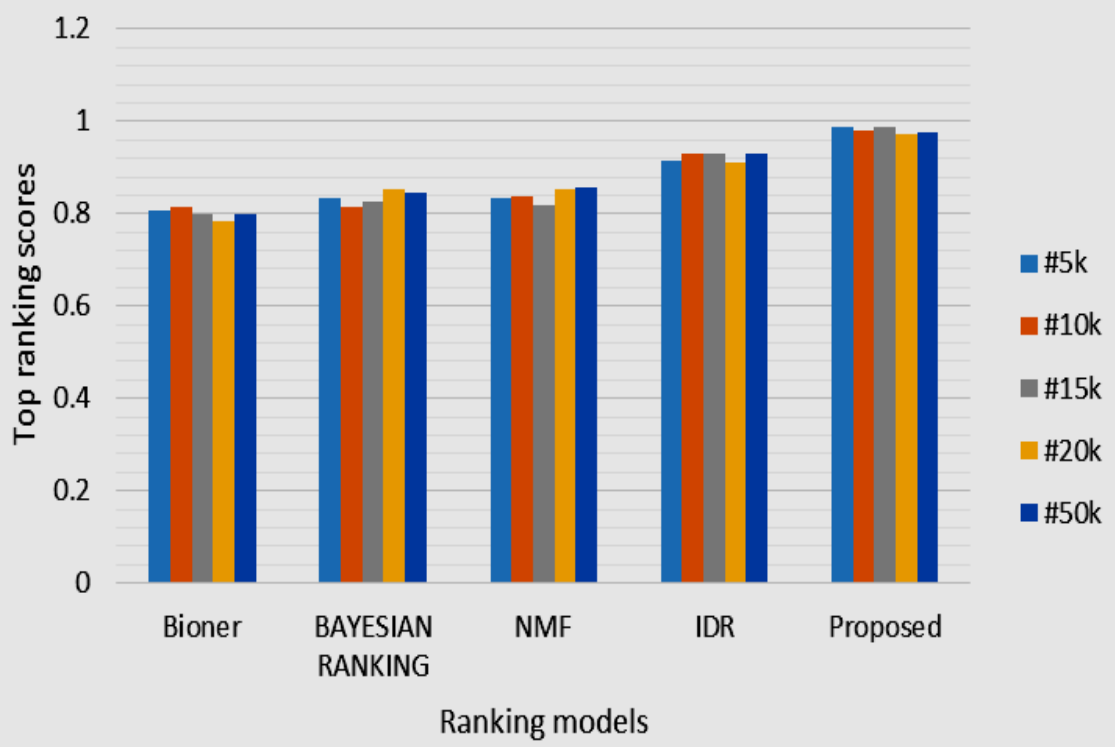

Figure 2: Comparison of the biomedical document ranking models using MapReduce framework.

Figure 2, presents the comparison of the proposed model to the existing models for gene or protein document ranking using the MapReduce framework. From the table, it is evident that the proposed model has high accurate gene or protein based ranking than the existing models. As the size of the document sets increases, proposed model has high computational contextual ranking compared to the existing models.

Table II presents the comparison of the proposed model to the existing models in terms of runtime (secs) for gene or protein document ranking. From the table, it is clear that the proposed model has less runtime (secs) compared to the existing models on large biomedical documents.

TABLE II. COMPARISON OF THE PROPOSED MODEL TO THE EXISTING MODELS IN TERMS OF RUNTIME (SECS)

Documents size	Bioner	Bayesian ranking	NMF	IDR	Proposed
#5k	10.3	9.84	9.63	9.15	8.054
#10k	18.56	17.93	18.58	18.36	16.78
#15k	28.94	27.13	27.43	27.83	26.83
#20k	42.43	38.46	39.56	37.46	35.94
#50k	53.64	52.74	53.64	51.83	45.74

Table III describes the number of relevant document indexing, extraction and ranking of biomedical documents using map-reduce framework. From the table, it is observed that the traditional models have less document extraction process based on gene or protein compared to the proposed model.

TABLE III. PERFORMANCE COMPARISON OF THE PROPOSED MODEL TO THE EXISTING MODELS IN TERMS OF GENE OR PROTEIN RELATED DOCUMENTS COUNT.

Documents size	Bioner	Gene-Protein Related Documents			
		Bayesian Ranking	NMF	IDR	Proposed
#5k	485	574	746	757	1043
#10k	2874	3874	5833	5893	6987
#15k	7973	8944	9763	10723	13878
#20k	8828	10883	13781	15872	18567
#50k	20848	25873	28774	34788	43847

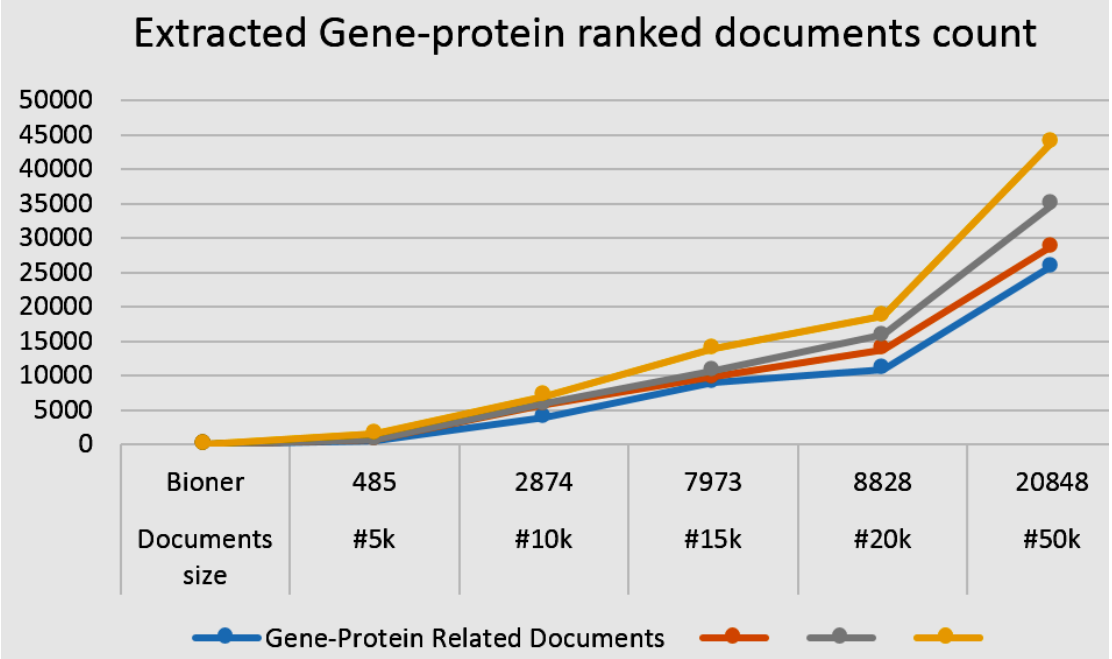

Figure 3. Performance comparison of the proposed model to the existing models in terms of gene or protein related documents count.

Figure 3 describes the number of relevant document indexing, extraction and ranking of biomedical documents using Mapreduce framework. From the table, it is observed that the traditional models have less document extraction process based on gene or protein compared to the proposed model.

V. CONCLUSION

Big data are now rapidly growing in all domain applications and fields. Learning models from these large data are expected to bring essential transformations and opportunities for different domain applications. Similarly, most of the traditional big data mining models and static classification models are not inherently scalable or efficient to find the essential hidden patterns on large distributed databases at high speed, high true positive, low error rate and incompleteness. As the biomedical repositories such as PubMed and MEDLINE are expanding exponentially, an accurate predictive model is required for knowledge discovery in Hadoop environment. However, the traditional text mining approaches are inefficient in extracting useful information from large data. . Also, traditional ranking models are not efficient to rank the biomedical documents using weighted genes or weighted proteins on large biomedical repositories. We focus on the problem of indexing, extracting and ranking biomedical document sets using gene or protein entities on large databases. In the proposed model, a novel MapReduce based natural language processing framework is designed and implemented on large biomedical databases using weighted gene or protein measures and document ranking score. Experimental results show that the proposed model has high contextual ranking accuracy, less search space and time consumption compared to the traditional biomedical document ranking models. In future, this work can be extended to apply multi-gene or protein features for document clustering and summarization process.

REFERENCES

[1] X. Chen, A. E Gururaj, B. Ozyurt, R. Liu, E. Soysal, T. Cohen, F. Tiryaki, Y. Li, N. Zong, M. Jiang, D. Rogith, M. Salimi, H. Kim, P. Rocca-Serra, A. Gonzalez-Beltran, C. Farcas, T. Johnson, R. Margolis, G. Alter, S. Sansone, I. M. Fore, L. Ohno-Machado, J. S. Grethe and H. Xu, "DataMed – an open source discovery index for finding biomedical datasets", Journal of the American Medical Informatics Association, 0(0), 2018, 1–9.

[2] S. Sohn, Y. Wang, C. Wi, E. A. Krusemark, E. Ryu, M. H. Ali, Y. J. Juhn and H. Liu, "Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions", . Journal of the American Medical Informatics Association, 0(0), 2017, 1–7.

[3] J. Cuzzola, E. Bagheri and J. Jovanovic, "UMLS to DBPedia link discovery through circular resolution ",Journal of the American Medical Informatics Association, 0(0), 2018, 1–8.

[4] E. S. Chen, G. Hripcsak, H. Xu, M. Markatou and C. Friedman, "Automated Acquisition of Disease–Drug Knowledge from Biomedical and Clinical Documents: An Initial Study", Journal of the American Medical Informatics Association Volume 15 Number 1 Jan / Feb 2008, pp. 87-100.

[5] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu and H. Xu, "CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines", "Journal of the American Medical Informatics Association, 0(0), 2017, 1–6".

[6] K. Raja, S. Subramani and J. Natarajan, "NLP-MTFLR: Document-Level Prioritization and Identification of Dominant Multi-word Named Products in Customer Reviews".

[7] R. Sivashankari and B. Valarmathi, "Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator".

[8] M. Noll, J. Lete and P. E. Meyer, "A hybrid named entity tagger for tagging human proteins/genes", Int. J. Data Mining and Bioinformatics, Vol. 10, No. 3, 2014, pp. 33-539 27-230 5-329.

[9] Moradi, M. and Ghadiri, N. (2018). Different approaches for identifying important concepts in probabilistic biomedical text summarization. Artificial Intelligence in Medicine, 84, pp.101-116.

[10] Raja, K. and Natarajan, J. (2018). Mining protein phosphorylation information from biomedical literature using NLP parsing and Support Vector Machines. Computer Methods and Programs in Biomedicine, 160, pp.57-64.

.

.

Appendix- Screenshot of Gene or Protein based Document Ranking,


```
        {
       "uid": "PMC2928015",
       "pmcid": "2928015",
       "pmid": "20733057",
       "docSource": "PMC",
       "articleType": "ra",
       "pmc_url": "http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928015",
       "pubMed_url": "http://www.ncbi.nlm.nih.gov/pubmed/20733057",
       "title": "Structure of hibernating ribosomes studied by cryoelectron tomography in vitro and in situ.",
       "fulltext_html_url": "JCB_201005007_thumb.gif",
       "journal_title": "The Journal of cell biology",
       "journal_abbr": "J. Cell Biol.",
       "journal_date":         {
          "day": "23",
          "month": "Aug",
          "year": "2010"
       },
       "authors": "Ortiz JO, Brandt F, Matias VR, Sennels L, Rappsilber J, Scheres SH, Eibauer M, Hartl FU, Baumeister W",
       "affiliate": "Department of Molecular Structural Biology, Max Planck Institute of Biochemistry, Martinsried, Germany.",
       "Outcome":          [
                   {
            "@score": "0.136",
            "#text": "Ribosomes arranged in pairs (100S) have been related with nutritional stress response and are believed to represent a &amp;quot;hibernation state.&amp;quot; Several
proteins have been identified that are associated with 100S ribosomes but their spatial organization has hitherto not been characterized."
          },
                   {
            "@score": "-0.348",
            "#text": "We have used cryoelectron tomography to reveal the three-dimensional configuration of 100S ribosomes isolated from starved Escherichia coli cells and we have
described their mode of interaction."
          },
                   {
            "@score": "0.110",
```

```
   {
 "uid": "PMC2359756",
 "pmcid": "2359756",
 "pmid": "18423048",
 "docSource": "PMC",
 "articleType": "ra",
 "pmc_url": "http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2359756",
 "pubMed_url": "http://www.ncbi.nlm.nih.gov/pubmed/18423048",
 "title": "Expression profiling with RNA from formalin-fixed, paraffin-embedded material.",
 "fulltext_html_url": "http://www.biomedcentral.com/1755-8794/1/9",
 "journal_title": "BMC medical genomics",
 "journal_abbr": "BMC Med Genomics",
 "journal_date":          {
    "day": "19",
    "month": "04",
    "year": "2008"
 },
 "authors": "Oberli A, Popovici V, Delorenzi M, Baltzer A, Antonov J, Matthey S, Aebi S, Altermatt HJ, Jaggi R",
 "affiliate": "Department of Clinical Research, University of Bern, Murtenstrasse 35 CH-3010 Bern, Switzerland. andrea.oberli@dkf.unibe.ch",
 "Outcome":          [
                {
       "@score": "0.165",
       "#text": "Preliminary scores were computed from genes related to the ER response, HER2 signaling and proliferation."
    },
                {
       "@score": "0.186",
       "#text": "Correlation coefficients between intact and partially fragmented RNA from FFPE material were 0.83 to 0.97."
    },
                {
       "@score": "0.191",
       "#text": "We developed a simple and robust method for isolating RNA from FFPE material."
    }
```

```
        {
        "uid": "PMC3484714",
        "pmcid": "3484714",
        "pmid": "23051056",
        "docSource": "PMC",
        "articleType": "ra",
        "pmc_url": "http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3484714",
        "pubMed_url": "http://www.ncbi.nlm.nih.gov/pubmed/23051056",
        "title": "Transient B12-dependent methyltransferase complexes revealed by small-angle X-ray scattering.",
        "fulltext_html_url": "",
        "journal_title": "Journal of the American Chemical Society",
        "journal_abbr": "J. Am. Chem. Soc.",
        "journal_date":         {
          "day": "19",
          "month": "10",
          "year": "2012"
        },
        "authors": "Ando N, Kung Y, Can M, Bender G, Ragsdale SW, Drennan CL",
        "affiliate": "Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.",
        "Outcome":          [
                {
            "@score": "0.113",
            "#text": "In the Wood-Ljungdahl carbon fixation pathway, protein-protein interactions between methyltransferase (MeTr) and corrinoid iron-sulfur protein (CFeSP) are
required for the transfer of a methyl group."
          },
                {
            "@score": "0.123",
            "#text": "While crystal structures have been determined for MeTr and CFeSP both free and in complex, solution structures have not been established."
          },
                {
            "@score": "0.188",
            "#text": "One assembly resembles the CFeSP/MeTr complex observed crystallographically with 2:1 protein stoichiometry, while the other best fits a 1:1 CFeSP/MeTr
arrangement."
          }
    "uid": "PMC2880994",
    "pmcid": "2880994",
    "pmid": "20487543",
    "docSource": "PMC",
    "articleType": "ra",
    "pmc_url": "http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2880994",
    "pubMed_url": "http://www.ncbi.nlm.nih.gov/pubmed/20487543",
    "title": "The number and microlocalization of tumor-associated immune cells are associated with patient's survival time in non-small cell lung cancer.",
    "fulltext_html_url": "http://www.biomedcentral.com/1471-2407/10/220",
    "journal_title": "BMC cancer",
    "journal_abbr": "BMC Cancer",
    "journal_date":          {
      "day": "20",
      "month": "05",
      "year": "2010"
    },
    "authors": "Dai F, Liu L, Che G, Yu N, Pu Q, Zhang S, Ma J, Ma L, You Z",
    "affiliate": "Department of Thoracic and Cardiovascular Surgery, West China Hospital, Sichuan University, Chengdu 610041, China.",
    "Outcome":          [
                {
          "@score": "0.233",
          "#text": "Tumor-associated immune cells may inhibit or promote tumor growth and progression."
      },
                {
          "@score": "0.152",
          "#text": "Correlation of the cell numbers and patient's survival time was analyzed using the Statistical Package for the Social Sciences (version 13.0)."
      },
```