European Journal of Molecular & Clinical Medicine

ISSN 2515-8260 Volume 07, Issue 11, 2020

A Hybrid multi-level disease filtering framework using biomedical documents and ICD drug discovery

Konda sreenu,
research scholar, dept of CSE, Acharya Nagarjuna University, Guntur,india.
Dr B.Raja Srinivasa reddy,
Professor, Dept of CSE, Sri vasavi institute of engineering and technology, india.

Abstract

Abstract — Multi-level disease prediction plays a vital role in the drug to disease discovery process. Most of the conventional models use static parameters and filtering approaches in order to filter the high dimensional feature space due to high computational time and memory. Also, these models are having less accuracy and high error rate for the classification models. In order to overcome these issues, a hybrid filtering method is proposed in order to optimize the data preprocessing and feature extraction on the high dimensional dataset. Experimental results proved that the hybrid data filtering and feature extraction models have better efficiency in terms of classification accuracy and runtime(ms) than the conventional models.

Keywords: Biomedical documents, gene-disease entities, ICD codes

1. INTRODUCTION

The volume of information is growing rapidly in different domains with the growth of distributed biomedical repositories. Document preprocessing is a reductive transformation of peer documents to generate summary by selecting an important information in the source documents. However, this has caused the problem of information overload. In order to resolve this drawback, multi-document clustering and feature extraction can be used to minimize the inter-cluster variation. This work considers the feature extraction strategy and the key phrase clustering and pattern discovery approach to eliminate information redundancy resulting from the multiple original documents. Breast cancer, diabetes, liver disease, breast cancer, bowel cancer, obesity, and other heart disorders have all become epidemics in recent years, posing a danger to global health. Cancer [1] is a terrible illness that is sometimes life-altering, life-threatening, and fatal. Most of their signs have a genetic origin [2], and preventing, diagnosing, treating, and curing these diseases poses a slew of challenges. Medical data classification has remained one of the leading research fields in the realms of biomedical informatics, machine learning, and pattern classification since medicine plays such an important role in saving human lives. Medical data [3,4] is the core

of biomedical informatics and straddles clinical and genomic data. Bioinformatics is a modern and increasingly growing area that deals with the collection, processing, analysis, and retrieval of knowledge about the structure and function of biological systems. At the intersection of computer science, applied mathematics, medicine, biology, and healthcare technology, biomedical informatics and bioinformatics come together to provide promising solutions for better clinical decisions. They mutually influence each other on several occasions due to their shared originality. As biomedical databases increase in size, new algorithms to model, handle, and interpret this data must be designed, generated, and implemented as quickly as possible[5]. A more fundamental understanding of biological processes is anticipated with the tools built in data mining, machine learning, and optimization techniques. This knowledge would undoubtedly pave the way for a more comprehensive examination of disease mechanics in order to achieve diagnosis, prognosis, disease screening, and drug discovery. Extracting information from medical data to help medical professionals and medical decision support systems is fascinating. The classification of medical data and genomic data is a common problem in the field of biomedical informatics. The aim of small medical data classification is to predict whether a patient has a disease or not. Furthermore, genomic data classification necessitates dimension reduction, i.e., the extraction of a smaller collection of features using feature extraction or feature selection methods, and then classification using efficient classifiers. For evaluating unknown parameters of neural networks, neurofuzzy hybrid networks, and selecting appropriate genes from genome data, machine learning heavily relies on evolutionary computing techniques. Several thousand to tens of thousands of gene/protein sequences make up a biomedical dataset. The data from each distributed document collection is scanned and converted into continuous, normalized data. The high dimensionality and imbalance existence of distributed document datasets are the main issues. Traditional machine learning classifiers classify and predict disease using a subset of features, with a high true negative rate and error rate[6]. The Lingpipe architecture is thought to be an easy way to integrate document analysis frameworks into other structures, such as the Unstructured Information Management Architecture (UIMA). It has a training mode and a variety of pre-compiled approaches for different domains[7]. Furthermore, these tokens may describe acronyms with recurring meanings in a particular field of science or literature. The majority of biomedical societies have embraced Semantic Web technologies, which include ontology creation, information extraction, and knowledge discovery[8].

2.Related works

Researchers have discovered that hybrid neural fuzzy systems, evolutionary fuzzy systems, evolutionary neural networks, evolutionary neuro fuzzy systems, evolutionary extreme learning machine, and other related systems have shown extraordinary performance in complex real-world applications. Usually, in small medical data, a set of features related to a specific disease are extracted from medical records and given a class mark. The presence or absence of a disease in a patient is demonstrated by the class name. By diagnosing cancer types with greater precision, this new method gives better clinical measurements to cancer patients [9]. The entire classification task is based on correctly identifying those features that contribute to the creation of acceptable

classifiers. The least important or incorrect features are naturally removed, and this method is known as dimension reduction. User-defined subspaces of interest are often used to solve the problem of high dimensionality. In the absence of prior domain information, however, user detection of the subspaces is vulnerable to error. Applying a dimensionality reduction approach to the dataset is another way to fix the curse of dimensionality. In 1901, Karl Pearson suggested the definition of PCA. As a result, PCA successfully reduces the broad dimension of microarray datasets by taking into account coordinates with high variance values and ignoring data with low variance. PCA is a pattern recognition algorithm that can be used to identify similarities and differences in data. The importance of the features is assessed, their accuracy is evaluated, and a predetermined search algorithm with a classifier is used in this process. The classification accuracy of a function subset is used to assess its consistency. The function subset that contributes the highest classification accuracy with the fewest features is thought to be the most optimal. For feature selection, different forms of evolutionary algorithms (EAs) are used. These stochastic search methods work on possible solutions arising from the natural genetics process, and they are modeled after the metaphor of natural biological evolution. At the end, the best solution is found. As a result, the success of ANN as a classifier is largely determined by the right combination of structures and learning algorithms. However, the most important drawback of ANN is its connection with the gradient descent learning algorithm, which slows down model efficiency and raises computational overhead. Due to the initial random choice of parameters in the gradient descent learning algorithm, the convergence rate becomes very sluggish, and it is often stuck in local minima.

Lee , implemented the MapReduce algorithm in order to extract associations among different biomedical concepts in case of large text data [10] . Biomedical text data are considered as very important source of information. In this research paper, they demonstrated the use of MapReduce method which is actually a parallel and distributed programming paradigm. It has the responsibility to mine each and every association various biomedical concepts those are extracted out of different biomedical articles. Initially, biomedical concepts are gathered through a matching text process to unified medical language system. Unified medical language system can be defined as the most commonly used standard biomedical database. In the subsequent step, they introduced a MapReduce method which can be implemented in order to evaluate a specific kind of interestingness measures. The above mentioned method can be divided into two sub-methods. In future, the above proposed method can be extended[11].

Zheng et,al, proposed a new Naïve Bayes classification algorithm in order to automate the linking of gene ontology to MEDLINE documents [12] . They proposed a new and advanced text mining approach and named it as associative naive Bayes (ANB) classifier. This technique is useful to link MEDLINE documents to gene ontology automatically. This technique is actually a non-trivial extension of document classification approach from a specific set of classes acknowledge hierarchy such as gene ontology. As we all know, the complexity of gene ontology is very high, hence, an efficient knowledge representation structure can be implemented here. With the help of the above mentioned structure, they

presented the text mining classifier known as ANB classifier. This classifier has the responsibility to link MEDLINE documents to gene ontology.

The main issues of the conventional models include:

Data Distribution: Distributed data mining algorithm should efficiently work on distributed data from different data sources in the big data environment.

- I/0 Optimization: Even though the massive data are distributed uniformly at each node, Distributed Data Mining (DDM) techniques should try to minimize the dataset input and output operations in the pattern analysis stage.
- Avoid Duplication: Distributed data mining algorithms such as document classification, document pattern mining and feature extraction models should try to minimize the phrase or sentence duplication.

3.Proposed Model

In the proposed model, initially all the biomedical gene/protein entities are extracted from the biomedical repositories such as PubMed/Medline .All the input biomedical documents are pre-processing using tokenization , stemming and stop word removal. Each document is pre-processing and its gene/protein features are extracted using the Abner tagger as shown in fig 1.

1)Tokenization: Tokenization is the first step in the document pre-processing process. All of the words that pattern matching algorithms accept are extracted from various documents in this section. Pattern mining algorithms can consider these groups of terms. Several common terms that have no effect on the pattern extraction method are listed and removed. Stop-words are the most widely used words that have no impact on the pattern mining process. Pronouns, prepositions, conjunctions, and other stop-words are examples. As the number of stop-words is reduced, the pattern mining method is greatly improved.

2) Pruning :Starting with suffix removal and ending with the development of word stems, stemming includes all operations. Since such words are considered equivalent, various stemming methods have been used to translate each word to its source. To put it another way, this tool is used to combine terms with identical conceptual meanings. As a result, all of these types of words are handled as a single word, reducing the total size of the dictionary and necessitating less storage space and processing time.

3) Trimming :The pruning method excludes terms that are either extremely rare or extremely popular. Pattern mining is complicated by terms of extremely low or extremely high frequency.

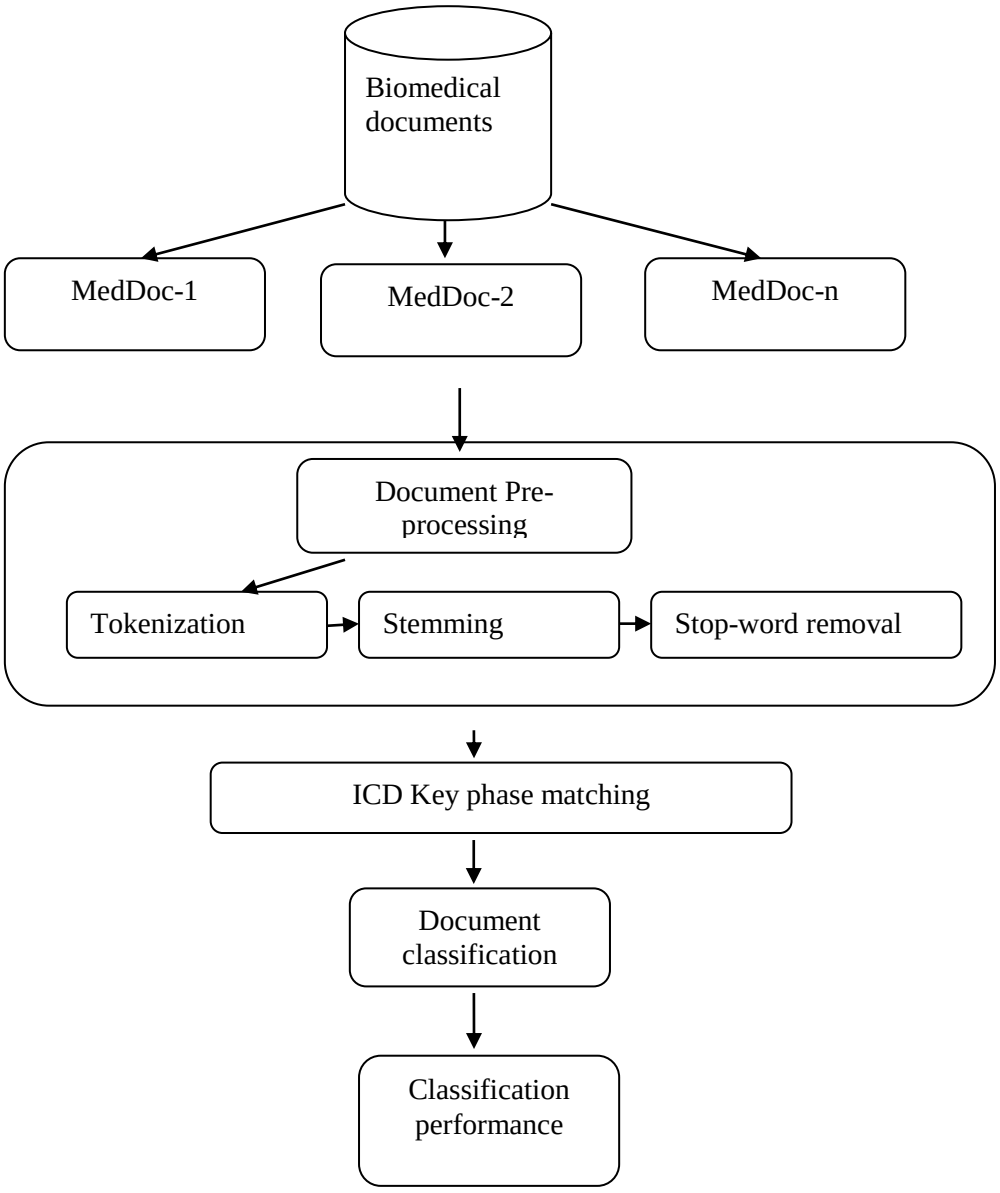

Figure 1: Proposed Framework

A significant number of abstracts from MEDLINE are derived and correlated with gene names. The gene name or synonym in the document's title may be used to identify a gene. There are two stages to the pattern mining process for gene detection. Initially, all words linked to genes are identified and extracted from a MEDLINE database. For the identification of genes related terms, TF-IDF and Z-scores approaches are used. To shape the gene features in the second stage, a document preprocessing method is used. After the preprocessing phase, each document and its corresponding gene/protein tags are extracted to find the highest probability biomedical features

for the graph initialization process. Here, each biomedical token, gene_protein tags and ICD codes are used to find the feature probability ranking.

Algorithm !: Hybrid Multi-level Filtering using biomedical documents and ICD codes

Input : Biomedical XML documents D, λ min threshold;

Output: Training dataset with entitiy and probabilistic scores as features.

Procedure:

1. Read Biomedical PubMed/Medline data in xml format.
2. Extract each document PMID in the xml file.
3. For each document d_i in D[]
4. Do
5. Dt_i=Tokenize(d_i)
6. Sdt_i=stemming(Dt_i)
7. PBD[]=stopwordremoval(Sdt_i)
8. Done
9. DiseaseEntities=DE[]=Abner(PBD);// Extract gene_protein tags using Abner library.
10. DET[]=Tokenize(DiseaseEntities [],D);
11. for each token t in DET
12. Compute ICD_DiseaseEntity probability to each document as
13. $DE\,\text{Prob}(D[i],t) = \text{Max}\{\frac{\text{Prob}(t/\,D[i])}{\text{Prob}(tf(t)/\,D)}\};i = 01,2...N$
14. Find the contextual similarity words in the dataset D.
15. For each ICD code ic in ICD[]
16. Do

    Find the contextual similarity of ICD code in each document as new feature

    $\text{Sim}(ic/D) = \text{Cos}(ic, DET[]) / \sqrt{\chi(ic, DET[])}$

17. Done
18. Extract ICD and its matching feature entities in each document as the training data by using threshold.
19. For each clinical document from D[]

    Do

    For each term in D

    Do

    If($\text{sim}(ic_i, DET(t)) > \lambda$)

    Then

    Add training feature TF.

    End for\

    Add TF to the training dataset TD.
20. End for

In the algorithm1, each document is filtered by using the tokenization, stemming, stopword

removal etc. After tokenization, gene/proteins are extracted using the java library ABNER tagger. These gene/proteins are used to find the probability computation to each document for document entity scoring and ICD mapping process.

Algorithm 2:

1: Read training dataset with multiple features as disease entities.
2: Select class feature in the available features classes list.
3: Classify training data using the ensemble classification learning model.
4: Apply naïve bayes to the training data with the selected class label.
5: Apply bayesnet to the training data with the selected class label.
6: Apply decision tree (random tree) to the training data with the selected class label.
7: Compute majority voting to each classification model for disease prediction
8: Repeat to each disease class for ensemble classification.
9: Compute performance metrics on the ensemble classification process.
10: End classification process.

4.Experimental

Experimental results are evaluated on large collection of clinical document sets taken from the repository. Different biomedical datasets such as Pubmed and medline xml datasets are used for document filtering and classification process. Here, each dataset is pre-processed to remove the uncertain features or noisy content. After the documents pre-processing phase, each document is classified using the ensemble learning algorithm.

Table 1: Comparative analysis of proposed ensemble biomedical disease prediction model to the conventional models on training dataset.

TestSamples	NaiveBayes	SVM	KNN	EnsembleClassifier
Test5%	0.935	0.933	0.936	0.954
Test10%	0.934	0.934	0.923	0.957
Test15%	0.937	0.932	0.926	0.96
Test20%	0.934	0.931	0.937	0.954
Test25%	0.944	0.929	0.924	0.958
Test30%	0.944	0.931	0.941	0.952
Test35%	0.944	0.933	0.942	0.958
Test40%	0.93	0.93	0.923	0.96
Test45%	0.937	0.931	0.919	0.951
Test50%	0.933	0.932	0.931	0.952

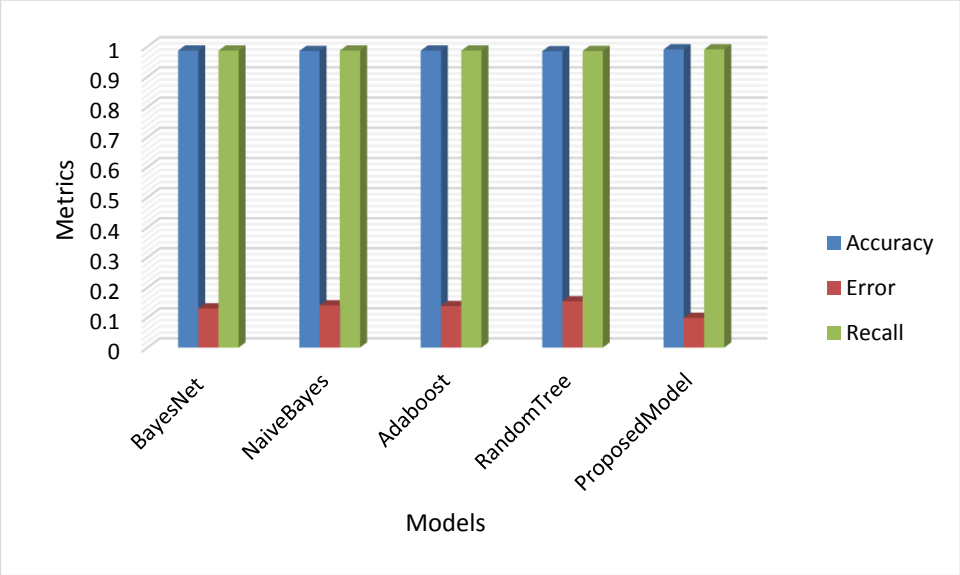

Figure 2: Comparative analysis of proposed accuracy ,recall and error rate of the proposed ensemble learning model to the traditional models on large training datasets.

```
Proposed Filter based Classification Model
------------------

ch_injury = 0
|   ch_diarr = 0
|   |   kindAccidentPoison = 0
|   |   |   ac_injury = 0
|   |   |   |   pe_injury = 0
|   |   |   |   |   kindAccidentRta = 0
|   |   |   |   |   |   treat_compract = 0
|   |   |   |   |   |   |   death_home = 0
|   |   |   |   |   |   |   |   kindAccidentDrowning = 0
|   |   |   |   |   |   |   |   |   death_else = 0
|   |   |   |   |   |   |   |   |   |   born_alive = 0
|   |   |   |   |   |   |   |   |   |   |   acute = 0: 1 (77.0/6.0)
|   |   |   |   |   |   |   |   |   |   |   acute = 1: 0 (40.0)
|   |   |   |   |   |   |   |   |   |   born_alive = 1
|   |   |   |   |   |   |   |   |   |   |   kindAccidentFalls = 0
|   |   |   |   |   |   |   |   |   |   |   |   ch_fever = 0
|   |   |   |   |   |   |   |   |   |   |   |   |   kindAccidentBiteSting = 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   severity_Mild = 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   acute = 0: 1 (280.0/41.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   acute = 1: 0 (848.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   severity_Mild = 1: 0 (29.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   kindAccidentBiteSting = 1: 0 (8.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   ch_fever = 1
|   |   |   |   |   |   |   |   |   |   |   |   |   acute = 0: 1 (154.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   acute = 1: 0 (63.0)
|   |   |   |   |   |   |   |   |   |   |   kindAccidentFalls = 1: 0 (19.0/3.0)
|   |   |   |   |   |   |   |   |   death_else = 1: 0 (64.0)
|   |   |   |   |   |   |   |   kindAccidentDrowning = 1: 0 (66.0)
|   |   |   |   |   |   |   death_home = 1: 0 (124.0/1.0)
|   |   |   |   |   |   treat_compract = 1
|   |   |   |   |   |   |   ch_brl = 0: 0 (28.0/1.0)
|   |   |   |   |   |   |   ch_brl = 1: 1 (2.0)
|   |   |   |   |   kindAccidentRta = 1: 0 (67.0/1.0)
|   |   |   |   pe_injury = 1: 0 (71.0)
|   |   |   ac_injury = 1: 0 (73.0)
|   |   kindAccidentPoison = 1: 0 (19.0/1.0)
|   ch_diarr = 1: 1 (22.0)
ch_injury = 1
|   ab_size = 0
|   |   acute = 0: 1 (6.0)
|   |   acute = 1: 0 (2.0)
|   ab_size = 1: 0 (2.0)
```

```
=== Stratified cross-validation ===

Correctly Classified Instances        2007               97.2384 %
Incorrectly Classified Instances        57                2.7616 %
Kappa statistic                          0.9269
Mean absolute error                      0.0482
Root mean squared error                  0.1553
Relative absolute error                 13.0982 %
Root relative squared error             36.1977 %
Total Number of Instances             2064


=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.969     0.018     0.994       0.969   0.982      0.928    0.990     0.995     0
                 0.982     0.031     0.911       0.982   0.945      0.928    0.990     0.949     1
Weighted Avg.    0.972     0.021     0.974       0.972   0.973      0.928    0.990     0.984
```

PHMRC_IHME_India_Adult_12-69yrsFinal

```
Proposed Filter based Classification Model
------------------

breastSwell = 0: 0 (1221.0/35.0)
breastSwell = 1
|   wgtLoss_Large = 0: 0 (8.0/1.0)
|   wgtLoss_Large = 1: 1 (4.0)

Number of Leaves  :     3

Size of the tree :      5


Time taken to build model: 0.58 seconds
0.03 seconds

=== Stratified cross-validation ===

Correctly Classified Instances        1197               97.0803 %
Incorrectly Classified Instances        36                2.9197 %
Kappa statistic                          0.177
Mean absolute error                      0.0566
Root mean squared error                  0.1682
Relative absolute error                 89.0974 %
Root relative squared error             94.9214 %
Total Number of Instances             1233


=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000     0.900     0.971       1.000   0.985      0.312    0.560     0.971     0
                 0.100     0.000     1.000       0.100   0.182      0.312    0.560     0.139     1
Weighted Avg.    0.971     0.871     0.972       0.971   0.959      0.312    0.560     0.944


=== Confusion Matrix ===

    a    b   <-- classified as
 1193    0 |    a = 0
   36    4 |    b = 1
2.75 seconds
```

Conclusion

In this paper, a hybrid biomedical document filtering and disease classification model is designed and implemented in order to optimize the prediction rate. Since, most of the conventional models are independent of biomedical document mapping with the disease prediction. Also, these conventional models have high error rate or true positive rate. In order to overcome these issues, a hybrid document filtering based disease prediction framework is designed and implemented on the training databases. Experimental results show that the present model has better efficiency in terms of accuracy, recall and error rate.

References

[1] S. Kim and J. Yoon , Link-topic model for biomedical abbreviation disambiguation, Journal of Biomedical Informatics 53 (2015) 367–380.

[2] K. Liu, W. R. Hogan and R. S. Crowley , Natural Language Processing methods and systems for biomedical ontology learning, Natural Language Processing methods and systems for biomedical ontology learning, Journal of Biomedical Informatics 44 (2011) 163–179

[3] T. Theodosiou, L. Angelis, A. Vakali, G. N. Thomopoulos , Gene functional annotation by statistical analysis of biomedical articles, international journal of medical informatics 7 6 (2 0 0 7) 601–613

[4] Z. Wang, S. Xu and L. Zhu , Semantic Relation Extraction Aware of N-Gram Features from Unstructured Biomedical Text, Journal of Biomedical Informatics.

[5] E. Yan and Y. Zhu, Tracking word semantic change in biomedical literature, International Journal of Medical Informatics 109 (2018) 76–86.

[6] N. Zong, S. Lee, J. Ahn and H. Kim , Supporting inter-topic entity search for biomedical Linked Data based on heterogeneous relationships, Computers in Biology and Medicine 87 (2017) 217–229

[7] T. Hofmann "Probabilistic latent semantic indexing" Proc. ACM SIGIR Forum pp. 211-218 2017.

[8] J. O. Wrenn D. M. Stein S. Bakken P. D. Stetson "Quantifying clinical narrative redundancy in an electronic health record" J. Amer. Med. Inform. Assoc. vol. 17 no. 1 pp. 49-53 2010.

[9] R. Cohen M. Elhadad N. Elhadad "Redundancy in electronic health record corpora: Analysis impact on text mining performance and mitigation strategies" BMC Bioinf. vol. 14 pp. 10 Jan. 2013.

[10] C. Lee Z. Luo K. Y. Ngiam M. Zhang K. Zheng G. Chen B. C. Ooi W. L. J. Yip "Big healthcare data analytics: Challenges and applications" in Handbook of Large-Scale Distributed Computing in Smart Healthcare Cham Switzerland:Springer pp. 11-41 2017

[11] T. Zheng W. Xie L. Xu X. He Y. Zhang M. You G. Yang Y. Chen "A machine learning-based framework to identify type 2 diabetes through electronic health records" Int. J. Med. Inform. vol. 97 pp. 120-127 Jan. 2017.

[12] C. Li S. Rana D. Phung S. Venkatesh "Hierarchical Bayesian nonparametric models for knowledge discovery from electronic medical records" Knowl.-Based Syst. vol. 99 pp. 168-182 May 2016.