A Study of Intelligent Text Miner Using GPU

[I]Kanchan Pawar, [II]Prof. A. D. Thakare, [III]Pratiksha Patil, [IV]Poonam Kshirsagar, [V]Rupa Solapure

Abstract

Text mining is most important technology that can be applied to numerous tasks in the biomedical domain, E-commerce and other fields where the massive amount of data is present. In this proposed work we are using Medline and PubMed repositories that contains huge amount of medical data. The rapid growth of these online repositories where large volumes of documents are available which has motivated the search for the hidden knowledge from these resources. Thus, an automated system which could correctly extract information in parallel with efficient speed from PubMed is required. Text mining is able to retrieve useful information from PubMed and Medline. It can be considered as an extension of data mining or knowledge discovery from databases.

The goal of Intelligent Text Miner Using GPU is to extract knowledge from online repositories and put it into practical use in the forms diagnosis, prevention and treatment of diseases. So, the novel text mining algorithm to improve the performance of the information retrieval system is proposed.

Keywords

Bio-Medical Text mining, Named Entity Recognition (NER), Relation extraction

I. Introduction

The huge numbers of biomedical text provide a rich source of knowledge for biomedical research. Text mining can help to mine information and knowledge from a large amount of text and it is now widely used in biomedical research. Text mining focuses on many computational technologies, such as machine learning, natural language processing, biostatistics and pattern recognition, to find new, exciting and actionable knowledge in unstructured biomedical text. Using text mining approach not only knowledge given in texts could be identified and extracted but also new or implicit knowledge may also be exposed. In recent years, text mining is applied in many areas, such as finding concepts from text and their relations. End users do not have to read huge amount of text and depends only on text mining systems to extract important information from these textual resources. In the biomedical text mining, corpus mainly consists of annotated journal and conference articles that are extracted from MEDLINE which is the most important databases in the biomedical domain.

In the following, we introduce the basic concepts and techniques of text mining by using GPU (Graphical Processing Unit). We address some algorithms for each major task in text mining and we also discuss at great lengths how far these algorithms can be utilized in biomedical text mining.

II. Related work

The Biomedical literature contains large amount of integrated but unstructured data and it is almost impossible for people to read all of these biomedical text and discover new knowledge. Text mining is able to help researchers to complete this difficult task. For this reason, biomedical communities have adopted Semantic Web technologies, including ontology building, information retrieval, and knowledge discovery. To provide intelligent Web resources for clinical decision- making Semantic Web concepts are also employed.[1]

Useful work is already done on symptom-disease and disease-gene relationships. Finding of relationships between symptom-gene from bibliographic literature will be useful. Environmental factors are mainly correlated with TCM diagnosis or genes in conventional medicine. Safety issues of Chinese herbs are mainly considering the side-effects of herbs.[2]

Syndrome and disease association was extracted from TCM literature. Term co-occurrence was used for the identification of disease gene associations from MEDLINE. Relationships between syndromes and genes were identified in common diseases.[3]

TCM clinical records contain huge information including the traditional disease, their symptom, environmental, social factors, diagnoses and treatments. Manually data collected for 20,000 inpatients and over 20,000 outpatients. To analyze the clinical data, a system to support the medical knowledge discovery and the Clinical decision-making was developed [4].

Unified TCM Language System (UTCMLS) is the largest TCM Semantic Web ontology including 5,000 concepts and 20,000 instances, serving as a common knowledge representation scheme to improve the quality of semantic search and query, and to infer semantic suggestions such as synonyms and associated concepts [5].

III. Data and Areas of Study

The biomedical Abstracts are taken as training dataset from PubMed. The PubMed Central (PMC) is a free digital repository that archives publically available full text scholarly articles of biomedical and life sciences journal literature. Hence, in this proposed work we are applying different text mining on PubMed abstracts to extract novel knowledge such as gene, protein names and gene associations by using GPU.

IV. Evaluation of related work

The text mining is a technique useful in different areas where large amount of data is available and to process this integrated but unstructured data, high performance and parallel system is needed. While considering currently available systems, they are capable of handling such a large amount of data but the drawback is it takes more time than the proposed system.

The time required to process such a large amount of data can be reduce by using parallel computation which can be easily achieve by using GPU. As proposed system takes abstracts of various diseases as an input. Diseases like cancer, malaria, MTB are one of the most important study areas for biomedical researchers. There are many publications on this diseases and keeps increasing every year.

It is almost impossible for the people to read all of these publications and discover new knowledge. Intelligent Text miner using GPU is able to help researchers to complete this difficult task and helps in diagnostics, treatments and prevention.

V. Proposed Work

The proposed work contains five main modules as shown in figure, the major modules are text pre-processing, dictionary construction, and text classification are performed in parallel using GPU. Basically GPU, the graphical processing unit is a specialized electronic circuit designed for rapid accessing and manipulating computer graphics by using its highly parallel structures. A GPU computing is a new computing approach where by hundreds of streaming processors (SP) on a GPU chip simultaneously communicates and cooperates to solve complex computing problems.

The first module is Text Gathering, in these module disease abstracts are collected from PubMed. In this work, we are using abstracts of three different diseases like Cancer, M.T.B, and Malaria.

Second Module is Text Pre-processing. In this module tokenization and then part-of-speech tagging or bag of word approach word stemming and the application of stop word list. Tokenization is a process in which text is divided into words or terms. Part of Speech (PoS) tagging tags words according to the grammatical context of the Word in the sentence, hence dividing up the words into nouns, verbs, etc. This is important for the exact analysis of relations between words, as it is needed in the extraction of relations between proteins [6].

Third module is Relation Extractor. The Relation Extractor extracts biomedical relations from titles and abstract texts in PubMed articles. We have defined the relation using Gene, protein names dictionary. For extracting relations, we use some natural language processing (NLP) techniques like part-of-speech (POS) tagging and syntactic parsing; and machine learning techniques like Maximum Entropy models. After extracting gene, proteins, we score associations between proteins and diseases using the scoring scheme as described [7], which is also the basis for the co-occurrence-based text-mining scores in STRING v9.1.

An important feature of the scoring scheme is that it simultaneously takes into account co-occurrences at the level of abstracts as well as individual sentences. To this end, we first calculate a weighted count (C(G, D)) for each pair of a gene (G) and a disease (D) over the n abstracts in the text corpus:

$$C(G,D) = \sum_{k=1}^{n} w_s \delta_{sk}(G,D) + w_a \delta_{ak}(G,D)$$

Where, w a = 3 and w s = 0.2 are the weights for co-occurrence within the same abstract and the same sentence, respectively, and the delta functions δ_{sk}(G, D) and δ_{ak}(G, D) signify whether or not G and D co-occur in abstract k or a sentence within it. A co-occurrence score (S (G, D)) is calculated from the weighted counts as:

$$S(G,D) = C(G,D)^{\alpha} \left(\frac{C(G,D)C(\cdot,\cdot)}{C(G,\cdot)C(\cdot,D)} \right)^{1-\alpha}$$

Where, C (G, .) is the sum over all diseases paired with gene G, C(., D) is the sum over all genes paired with disease D, the normalizing factor C(., .) is the sum over all pairs of genes and diseases, and the weighting factor a = 0.6. All parameters (w a , w s , and a) have in earlier work been optimized to give the best possible performance on finding functionally associated genes. An important property of this function is that it not only rewards for the gene and disease being mentioned together, but also penalizes for them being frequently mentioned together with other diseases or genes, respectively.

We next convert the co-occurrence scores (S(G, D)) to z-scores (Z(G, D)), which are easier to interpret and are robust to changes in the size of the text corpus. Finally, we calculate the confidence score (stars) as Z(G, D)/2, limited to a maximum of four stars to account for automatic text mining never being as reliable as manually curated annotations.

The fourth module is Relation Visualizer accumulates extracted relations and shows the relationship network into graphs and tables.

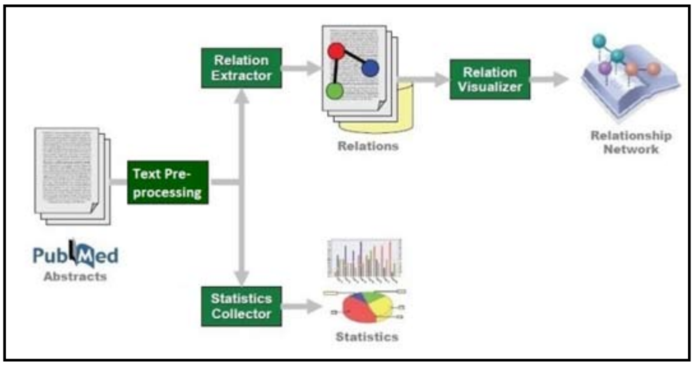

Fig. 1: System Overview

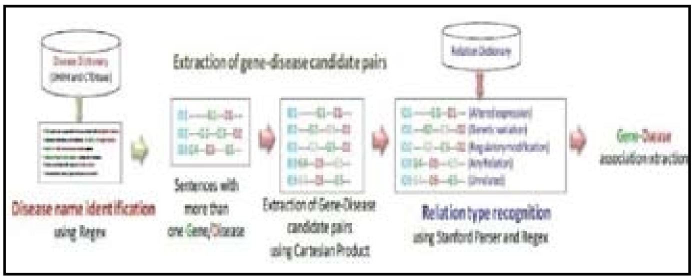
Extraction of gene-disease candidate pairs
Relation Dictionary
Disease name identification using Regex
Sentences with more than one Gene/Disease
Extraction of Gene-Disease candidate pairs using Cartesian Product
Relation type recognition using Stanford Parser and Regex
Gene-Disease association extraction

Fig. 2: Steps in Gene-Disease association and Extraction

When users select a relation in a network to see detailed information, sentences or abstract texts, where the relation is extracted, are displayed with the highlighted subject entity, object entity and event string.

The final module is Statistics Analyzer. In this module, The Relation Extractor mines information from the content of the articles, but the Statistics Analyzer gathers information from the outside of the content, that is, meta data etc. After collecting statistics information, the system shows statistics charts to users, and user can also export the statistics information as a Microsoft Excel file format.

VI. Conclusion

We introduced a system that discovers relations between various types of biomedical entities and gathers statistics information from metadata. An input to our system is the abstracts collected from PubMed and it visualizes relationship network from the collection after information extraction and shows statistics information after gathering metadata of the collection.

We think that our system will be useful for literature based validity survey before biomedical experiments and for biomedical research trends survey with statistical analysis and our relation extraction method will be also useful for large scale literature processing.

VII. Acknowledgement

This work was supported by the CSIR-National Chemical Laboratory, Research, development and consultancy organisation Pune. The authors thank Dr. M. Karthikeyan for help in basic

idea for the paper.

References

[1] *Zeeshan Javed1, Dr. Hammad Afzaf, "Biomedical Text Mining for Concept Identification from Traditional Medicine Literature", ICOSST, 2014 IEEE.*

[2] *Zhou X, Liu B, Wu Z, Feng Y.lntegrative mining of traditional Chinese medicine literature and MEDLlNE for functional gene networks.*

[3] *Zhou X, Chen S, Liu B, Zhang. "Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support".*

[4] *Zhaohui Wu, Tong Yu, Huajun Chen. "Information Retrieval and Knowledge Discovery on the Semantic Web of Traditional Chinese Medicine".*

[5] *Xuezhong Zhou, Yonghong Peng, Baoyan Liu. "Text mining for traditional Chinese medical knowledge discovery": A survey*

[6] *N. Daraselia, A. Yuryev, S. Egorov , S. Novichkova, A. Nikitin, I. Mazo, "Extracting human protein interactions from MEDLINE using a full ¬sentence parser", Bioinformatics, 20(5), 604-611, 2004*

[7] *S. Mørk, S. Pletscher-Frankild, A. Palleja, Bioinformatics (2013), http:// dx.doi.org/10.1093/bioinformatics/btt677.*

[8] *A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, et al., Nucleic Acids Res. 41 (2013) D808–D815, http://dx.doi.org/10.1093/nar/ gks1094.*

Authors Profile

Kanchan L. Pawar completed Diploma in Computer Engg. from Maharashtra State Board of Technical Education, Pune in 2013. Currently pursuing BE, Bachelor's in Computer Engg. from Savitribai Phule Pune University, Pune.

Pratiksha Y. Patil completed Diploma in Computer Engg. From Maharashtra State Board of Technical Education, Aurangabad in 2013. Currently pursuing BE, Bachelor's in Computer Engg. from Savitribai Phule Pune University, Pune

Poonam D. Kshirsagar completed Diploma in Computer Engg. From Maharashtra State Board of Technical Education, Pune in 2013. Currently pursuing BE, Bachelor's in Computer Engg. from Savitribai Phule Pune University, Pune

Rupa T. Solapure completed Diploma in Computer Engg. from Maharashtra State Board of Technical Education, Solapur in 2013. Currently persuing BE, Bachelor's in Computer Engg. from Savitribai Phule Pune University, Pune.