An Efficient Extraction of Data in Biomedical Using Stemming Algorithm

[1]A.Srinivas, [2]V.Sudheer Goud, [3]J.Stalin Babu, [4]T.Charan Singh

***Abstract*-Now days Text Mining is a term which is currently being used to mean various things by various people. In my paper it has been used to refer to any process of Extracting information using Successor Variety Stemmer Algorithm to extract in patterns in Biomedical Text Mining and it is useful in research areas such as information extraction, information retrieval, natural language processing. Natural language processing uses some level of underlying linguistic representation of text, to make sure that the generated text is grammatically correct and fluent. Information retrieval involves directly with text mining process by extracting useful information from the texts. Information Extraction deals with the extraction of specified entities, events and relationships from unrestricted text sources. The primary resource for biomedical text mining is obviously based on text given by doctors and this may be used to introduce some widely-used text collections in the biomedical domain fields. The original conception of literature-based discovery was facilitated by the use of Medical Subject Headings to extract biomedical information, which are controlled vocabulary terms added to bibliographic citations during the process of med liner – indexing.**

***Index Terms*— Text Mining, Successor Variety Stemmer, Medical Subject Headings biomedical, Information Retrieval, Natural Language Processing.**

I. INTRODUCTION

A few years ago, a rapid different types of techniques have been implemented Biomedical Text Mining. We had used different types of software tools to extract data from the data base by many scientists, but they might not be not implemented in proper and simple and in a similar way. This work makes a part to close this scenario, proposing a framework to ease the development of naïve users and practicable applications in this field, discussed based on a set of modular units. These modules can be connected in various ways to create applications that fit different user roles. And also, developers could develop of new algorithms have a framework that allows them to simple and easily integrate their executing with software tools for related tasks using knowledge discovery from biomedical databases.

In present scenario research work for Information Extraction has improved very fast in a bioinformatics field, where these biomedical journal have become an very important and simple application area. This type of work comes from biologists and for them it will be very useful in the field. In the bioinformatics domain, the biomedical research area has been a target for text mining. The main goal of the text mining in this related area is to allow biomedical researchers to extract knowledge from the biomedical data warehouse literature in helping new uncovering in a more fast ,efficient and simple manner. Many scientist of the text mining research in this domain has been done in the context of Medline. It is open source software available in the market. Medline records consist of a title, an abstract, a set of manually allotted data about data terms. We had a various types of text mining in mining Information Extraction based data about data profiling have been proposed for Medline data. In evaluating biomedical text mining, most of the s researchers claimed that, most research scholars in this field they focuses on the development of only on functions.. Although some text mining systems have been developed but some people argued that none is real procedure used by end-users. Most of the text-mining tools in biomedical domain they had focused on test patterns and text collections which may developed by individual research groups. This field has been used to extract the data using text mining tool in biomedical field. The main source for biomedical text mining is obviously it is text, and this text introduces some generally collections of text in the biomedical field. The original creation of survey based on Medical Subject Headings, which are checked vocabulary terms added to bibliographic citations during the process of medliner indexing. We had used biomedical text mining tools and frameworks, and it provides links to text patterns. In this paper Biomedical text mining computers systems may have to utilize a variety of text processing techniques, but in this paper it is used one type of algorithm to extract simple information retrieval to advance from the biomedical data ware house and the use of algorithms developed by machine learning tools. Biomedical text is a rule governed by the software that extracts information or sentences relating to DNA, Medicines, protein families, their structures, functions and diseases from the biomedical data warehouse. The query based on user or biologists terms given by biologist of guide and rules, it may be to extract the data at accurate term extraction rather than wide recall. Person uploading biologist's main text or Medline data we are retrieving terms

ISSN: 2278 – 1323

from data ware houses, the persons or Doctors can extract sentences based on pre-determined guide. It also provides a responsibility insight into the set of rules, syntax and semantic context of the main source text by looking at words like medicines, syrups, surgical items, diseases, proteins etc. Biologists can extract data from dataware houses by recognition of entities like genes or proteins, functions and diseases from the biomedical literature shown in figure1

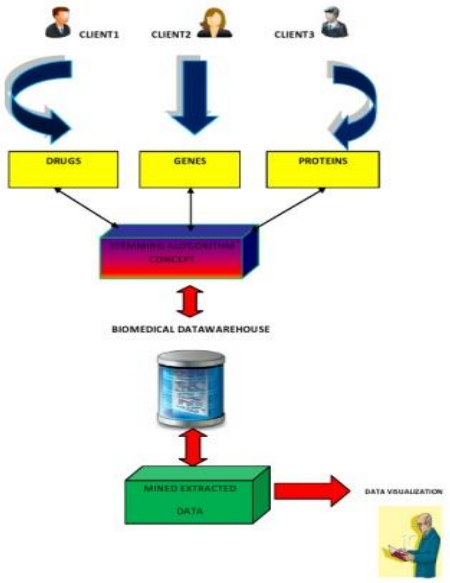

Figure 1: Architecture of Biomedical Mined Data

This paper has been presented one of the techniques and applications in text mining. The focus has been given on fundamental methods for conducting text mining. The methods include one of the stemming algorithm processing and information extractions. The purpose of this part is to give a brief summary to a reader on how text mining systems can be used in real time. The paper also addressed the most challenging issue in developing text mining systems. We have different types of stemming algorithms. These are classified into four types. In my paper am using Successor Variety stemming algorithms.

II. TYPES OF STEMMING ALGORITHMS

a) Table lookup approach
b) Successor Variety
c) N-gram stemmers
d) Affix Removal Stemmers

III. WORKING PROCEDURE

To determine the searching techniques in any whish book list or given data, the person may be he or she wants to find the term or queries from the given database or whish book list. For example a person wants to look or search the term from the database, i.e. "MARKET" The person may be the naïve user so he will enter the some number based on the number, the number of occurrences of term will be displayed as shown:

THE TERM "MARKET":

1) For the term MAR i.e. the number of occurrences=10.

2) For the term MARK i.e. the number of occurrences=7

3) For the term MARKS i.e. the number of occurrences=5

4) For the term MARKET i.e. the number of occurrences=2

5) For the term "empty" the number of occurrences=0

This type of method of using a stemmer algorithm in a searching provides a very difficult for the new users or naïve users they might be getting false matches.

Successor Variety algorithm:

To find out the word or string using successor variety algorithm ,the number of different characters follows in body of the text, for Example if you take the words

If you want to find out the successor varieties of words:

Color, Common, Consider, Compromise, cofactor.

To find out successor varieties of the word "CUBE". The first letter is "C" this is followed in the text body by four characters' O", and The Second letter is "CO" this is followed in the text body by five characters "L",M,N,M,F. and so on.

Consider the example below where the task is to determine the stem of the word DRUGS
Test Word: DRUGS
Corpus: RUG, RUGS, BUGS, JUGS, MUGS, LUGS, SHRUGS, LUGS.

Prefix	Successor Variety	Letters
D	4	R, U, G, S
DR	3	U, G, S
DRU	2	G, S
DRUG	1	S
DRUGS	1	BLANK

If you use the complete word the word may be partitioned or segmentation into in to "D" and "RUGS," since RUGS appears as a word in the corpus. So it gives the result. If the segment found in 5 words then it is in corpus, otherwise go for second stem.

IV. APPLICATIONS OF TEXT MINING IN BIOMEDICAL FEILD

a) DNA-expression arrays
b) Molecular medicine
c) Functional annotation
d) Protein interactions

e) Cellular location

V. IMPLEMENTATION

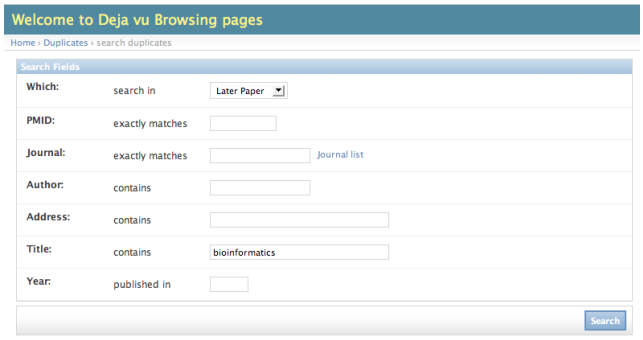

Figure 2: searching pages.

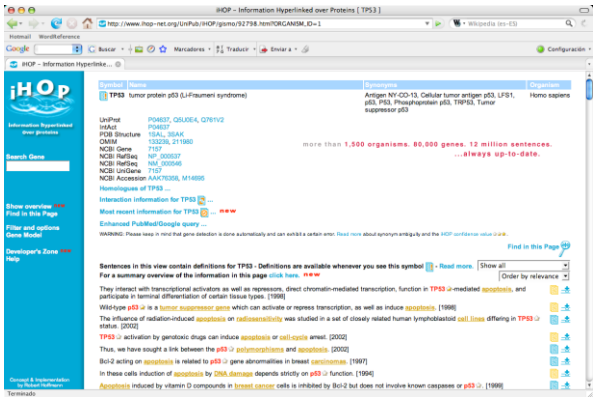

Figure 3: Searching for Proteins

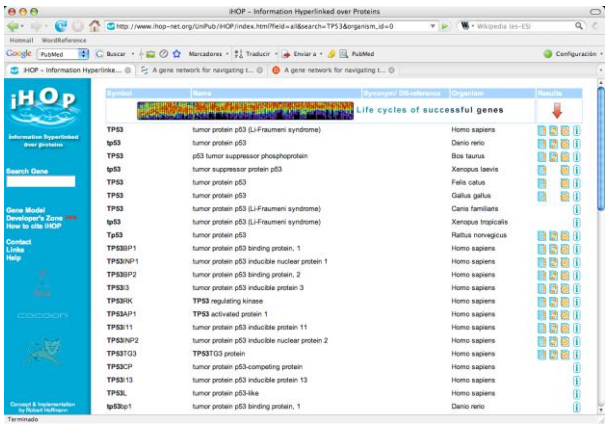

Figure 4: searching for genes

VI. CONCLUSION

Text mining technology is becoming an emergence technology for national security defense and Government agencies are investing considerable resources in the surveillance of all kinds of communication, such as email, chats in chat rooms. Extraction of text from biomedical literature is an essential operation. Given that there have been many text extraction methods developed; this paper presents a novel technique that employs keyword based article based on stemming algorithm to further enhance the text extraction process. The proposed algorithm, using data mining algorithm, seems to extract the text with contextual completeness in overall, individual and collective forms, making it able to significantly enhance the text extraction process from biomedical literature. The development of the proposed algorithm is of practical significance, but we can apply different techniques of stemming algorithm and it is challenging to text extraction that retrieves the relevant biomedical text more efficiently.

REFERENCES

[1] Cohen AM, Hersh WR, A survey of current work in biomedical text mining, Briefings in Bioinformatics, 2005, 6: 57-71
[2] Han, J., & Kamber, M., Data Mining Concepts and Techniques. CA : Morgan Kaufmann, 2001
[3] Airola, A., et al., All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinformatics, 2008. 9(Suppl 11): p. S2.
[4] Marcotte, E.M. et al. (2001) Mining literature for protein–proteininteractions. Bioinformatics17, 359–363
[5] Chun, H.W.et al. (2006) Extraction of gene disease relations from MEDLINE using domain dictionaries and machine learning. InPacific Symposium on Biocomputing 2006(Altman, A.B.et al., eds), pp. 4–15,World Scientific Publishing Co.
[6] Spasic, I.et al. (2005) Text mining and ontologies in biomedicine:making sense of raw text.Brief Bioinform.6, 239–251
[7] B. Stapley and G. Benoit,"Biobibliometrics: Information Retrieval and Visualization from Co-occurrences of Gene Names in MEDLINE Abstracts,"Proceedings of the Pacific Symposium on Biocomputing00 (PSB00), Honolulu, HI (2000), pp. 529–540
[8] MedicalSubjectHeadings(MeSH), http://www.nlm.nih.gov/mesh/meshhome.html

AUTHORS

A.Srinivas, Post Graduated in Computer Science & Engineering (**M.Tech**) From JNT **University**, Hyderabad in 2009 and Graduated in Computer Science & Information Technology (B.Tech) from **JNTU**, Hyderabad in 2004.He is currently working as an Assistant Professor, Department of Computer Science & Engineering in **Sri Indu College of Engineering & Technology (SICET),** (V) Sheriguda, (M) Ibrahimpatnam, R.R.Dist, and Telangana, India. He has 9+ years of Teaching Experience. His research interests include Cloud Computing, Data Mining, Information Security, Software Testing, Wireless Networks and Software Quality.

V.Sudheer Goud, Post Graduated in Master of Computer Application (**MCA**) from OU,1994 , Post Graduated in Master of Business Administration (**MBA**) from OU,2006, Post Graduated in Master of Computer Science & Engineering (**M.Tech**) from IETE , Hyderabad in 2013and Pursuing Phd in Computer Science in ANU. He is currently working as an Associate Professor, Department of Computer Science in **Holy Mary Institute of Technology and Science** (HITS), (V) Bogaram, (M) Keesara, R.R.Dist, Telangana, India. He has 21 years of Teaching

ISSN: 2278 – 1323

Experience. His research interests include, Text Mining, Cloud Computing and Information Security.

J.Stalin Babu Post Graduated in Computer Science & Engineering (**M.Tech**) From JNT **University**, Kakinada in 2011 and Graduated in Computer Science & Engineering (B.Tech) from Andhra **University,** Visakhapatnam in 2005.He is currently working as an Assistant Professor, Department of Computer Science & Engineering in Nalla Mallareddy Engieering college**,** Divyanagar,Near Narapally, (M) Ghatkesar, R.R.Dist, and Telangana, India. He has 7+ years of Teaching Experience. His research interests include Data Mining, Information Security, Software Engineering, and Cloud Computing & Computer Networks.

T.Charan Singh, Post Graduated in Computer Science & Engineering (**M.Tech**) From JNT **University**, Hyderabad in 2010 and Graduated in Computer Science & Engineering (B.Tech) from **JNTU**, Hyderabad in 2006.He is currently working as an Assistant Professor, Department of Computer Science & Engineering in **Sri Indu College of Engineering & Technology (SICET),** (V) Sheriguda, (M) Ibrahimpatnam, R.R.Dist, and Telangana, India. He has 6+ years of Teaching Experience. His research interests include Data Mining, Information Security, Software Engineering and Cloud Computing.

