SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM

TMLAI TRANSACTIONS ON
MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Volume 5 No. 3
ISSN 2054-7390

An Emperical Study of Clustering Algorithms to extract Knowledge from PubMed Articles

[1]Deepak Agnihotri, [2]Kesari Verma, [3]Priyanka Tripathi
[1,2]Dept. of Computer Applications, National Institute of Technology Raipur, CG-492010, INDIA;
[3]Dept. of Computer Engineering and Applications, National Institute of Technical Teachers Training and Research Bhopal, MP, INDIA;
dagnihotri.phd2012.mca@nitrr.ac.in;kverma.mca@nitrr.ac.in; ptripathi@nitttrbpl.ac.in

ABSTRACT

Extraction of useful information from biomedical literature is one of the thrust for the world nowadays due to availability of almost articles on the web in electronic form. Information retrieval (IR) from biomedical literature is finding useful patterns from the unstructured text corpus that satisfies information. In this paper intelligent text analysis is carried out on PubMed articles related to influenza virus. In this context, various algorithms are discussed to reveal the information from PubMed articles, like year wise count of articles containing influenza virus related terms (viz. H1N1, H5N1, and H7N1 etc.), countries with their publication count, which tells about the outbreaks of the diseases in these countries. The articles may be grouped by searching the keyword "influenza virus strain" pattern with the help of regular expressions. Automatic text categorization is another challenging issue for text mining. We applied k-means, fuzzy C-means, and fuzzy C-shell algorithm for automatic categorization of text articles. The association between words based on their cooccurrence is computed which further helps to categorize the documents based on their cooccurrences. The basic k-means clustering algorithm is first applied to cluster the documents, and then to handle the fuzzy nature of words which may belong to more than one cluster, fuzzy c-means clustering is applied to form more accurate clusters. As Fuzzy c-means method clusters the documents which are in linear spaces but not in the circle, spherical, or ellipsoidal spaces. A new method is proposed here, which considers the clusters of documents in the radius of the circle.

Keywords: Information Retrieval, Text Mining, Fuzzy Clustering, Influenza Virus, PubMed.

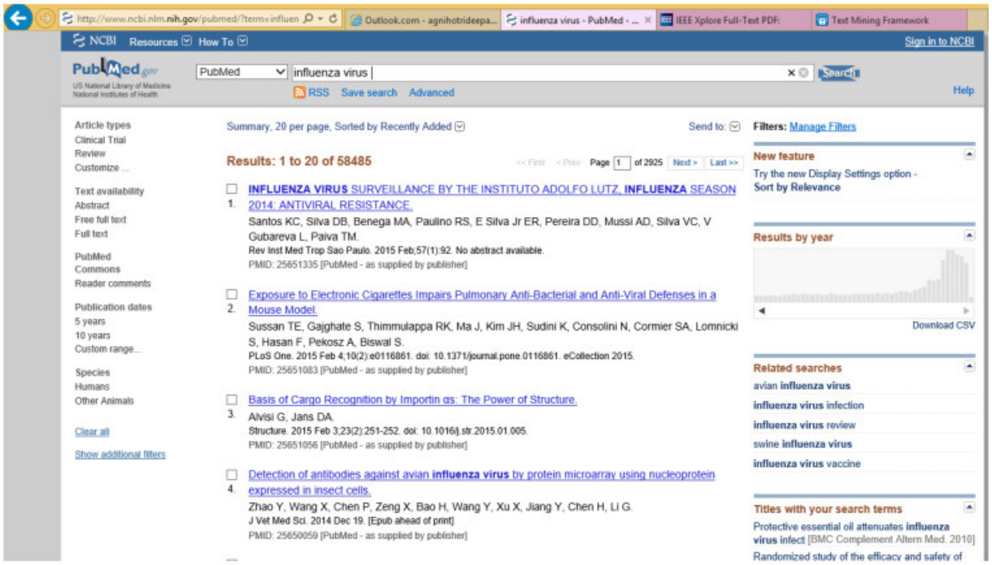
1 Introduction

It is not news that science produces an enormous literature, presently more than 23 million citations in MEDLINE alone and that needs computational means such as text mining (TM) to extract meaningful knowledge from it. The biological literature largely focused on describing relationships between entities (e.g. Genes, proteins and complexes), including how such entities interact and affect each other [10]. The increasing number of electronically available publications stored in databases such as PubMed, which increases interest in text mining and information extraction strategies applied to the biomedical and molecular biology literature. PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The United States National Library of Medicine (NLM) at the National Institutes of Health maintains the database as part of the Entrez system

DOI: 10.14738/tmlai.53.3106
Publication Date: 5th June, 2017
URL: http://dx.doi.org/10.14738/tmlai.53.3106

of information retrieval [11]. Thus, biological TM research has focused extensively on the automatic recognition, categorization [2] andnormalization of variant forms [3, 4] and mapping of these entities to unique identifiers in databases. This can facilitate entity-based searching of documents, which can be far more effective than simple keyword-based searches.

PubMed database is a large collection of bio-medical research papers available on the web and provides free access to these articles for further research and study. In this study, we applied text mining techniques for Influenza Virus related articles. As per World Health Organization (WHO) Influenza Fact Sheet November, 2014 report, Influenza Virus is spreading around the world in a yearly outbreak, resulting in about three to five million cases of severe illness and about 250,000 to 500,000 deaths. The death occurs mostly in the young, the old and those with other health problems. In this paper a novel algorithm is proposed which automatically extracts the PubMed articles along with its PubMed id, title, authors, country, year, issn, journal, publisher and some other important features of concerned articles. All these information are in the form of XML tags. Extraction of information from XML tags is also a big challenge which requires a method to parse these XML tags. After XML parsing all useful information is stored in Mysql5.6 relational database management system (see Fig. (1)).



Results: 1 to 20 of 58485

□ 1. **INFLUENZA VIRUS** SURVEILLANCE BY THE INSTITUTO ADOLFO LUTZ, **INFLUENZA** SEASON 2014: ANTIVIRAL RESISTANCE.
Santos KC, Silva DB, Benega MA, Paulino RS, E Silva Jr ER, Pereira DD, Mussi AD, Silva VC, V Gubareva L, Paiva TM.
Rev Inst Med Trop Sao Paulo. 2015 Feb;57(1):92. No abstract available.
PMID: 25651335 [PubMed - as supplied by publisher]

□ 2. Exposure to Electronic Cigarettes Impairs Pulmonary Anti-Bacterial and Anti-Viral Defenses in a Mouse Model.
Sussan TE, Gajghate S, Thimmulappa RK, Ma J, Kim JH, Sudini K, Consolini N, Cormier SA, Lomnicki S, Hasan F, Pekosz A, Biswal S.
PLoS One. 2015 Feb 4;10(2):e0116861. doi: 10.1371/journal.pone.0116861. eCollection 2015.
PMID: 25651083 [PubMed - as supplied by publisher]

□ 3. Basis of Cargo Recognition by Importin αs: The Power of Structure.
Alvisi G, Jans DA.
Structure. 2015 Feb 3;23(2):251-252. doi: 10.1016/j.str.2015.01.005.
PMID: 25651056 [PubMed - as supplied by publisher]

□ 4. Detection of antibodies against avian **influenza virus** by protein microarray using nucleoprotein expressed in insect cells.
Zhao Y, Wang X, Chen P, Zeng X, Bao H, Wang Y, Xu X, Jiang Y, Chen H, Li G.
J Vet Med Sci. 2014 Dec 19. [Epub ahead of print]
PMID: 25650059 [PubMed - as supplied by publisher]

Protective essential oil attenuates **influenza virus infect** [BMC Complement Altern Med. 2010]
Randomized study of the efficacy and safety of

Figure-1 illustrates the results of search query "influenza virus" from NCBI web page for PubMed articles

Automatic term recognition (ATR) refers to the automatic extraction of technical terms from domain-specific texts. It additionally encompasses the assignment of semantic categories of the extracted terms and mapping of them to concepts contained within terminological or ontological resources [1]. The Need for Term Recognition is due to an overwhelming number of textual resources becoming available in biomedicine, there is an increased interest in text mining techniques that can identify, extract, manage, and exploit biomedical knowledge hidden in the literature [1-2].

In systems biology, the terms are the backbone of such knowledge, since they constitute the linguistic realization of specialized concepts in the biology domain. As the main purpose of terms is to classify scientific knowledge, they are used as a means of scientific communication to convey biological concepts

[2]. Term disambiguation is the process of determining the sense of a term in a particular context and constitutes an important part of the process of automatic term recognition. It is often the case that a particular term can have different meanings in different contexts. For example, in systems biology texts, acronyms are very frequent: the acronym ER has about 80 possible definitions (expansions), including estrogen receptor, emergency room, enhancement ration, etc. Only by examining the context of the acronym in a text its correct meaning can be determined. A further example relevant to systems biology concerns species. Terms that are concerned with a number of different species are commonly referred to using the same name. This can cause problems for text mining applications that aim to link terms to a particular species [6]. Document clustering involves the use of descriptors and descriptor extraction. The Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users. The application of document clustering can be categorized into two types, online and offline. Online applications are usually constrained by efficiency problems when compared with offline applications.

The rest of the paper is organized as follows; second section gives the brief introduction about literature review related to this paper. The research problem discussed in the third section, research methodology used to solve the defined research problem is explained in the fourth section. Finally, the paper concludes with various research findings and their needs and how it may be helpful to the research society.

2 Literature Review

The main developments in this area have been related to the identification of biological entities (named entity recognition), such as protein and gene names [18] as well as chemical compounds and drugs in free text, the association of gene clusters obtained by microarray experiments with the biological context provided by the corresponding literature, automatic extraction of protein interactions and associations of proteins to functional concepts (e.g. Gene ontology terms). Even the extraction of kinetic parameters from the text or the sub cellular location of proteins has been addressed by information extraction and text mining technology. Information extraction and text mining methods have been explored to extract information related to biological processes and diseases [1, 2, 6, 8, 9, 10, 12, 21, 22, 23, 24, 25].

In the recent years, the edification of automated methods to recognize relationships between entities in biomedical texts has increased considerably, moving from the calculation of simple co-occurrence to the detection of pairwise semantic relations between interacting protein and to the extraction of sophisticated entities and event structures, involving multiple and categorized participants[12]. Fuzzy clustering approaches have been reviewed since 1993[17]. Goswami et. al. proposed a fuzzy based approach to Text Mining and document clustering. Rajesh N. Dave et al. proposed Adaptive Fuzzy c-Shells Clustering and Detection of Ellipses which is one of the main inspired works of this paper [7]. There are various useful biomedical text mining based articles published in Plos One Journal at [21, 22, 23, 24, 25] which may be helpful to start working in the field of biomedical text mining.

2.1 Research Problem

PubMed articles database published by various journals related with influenza virus is used as corpus in this paper. This database contains various information about the articles like unique PubMed id, article id, title, authors, country, abstract, year of publication, publication date, issn, page number, source or journal name of the published articles, publisher etc. The information is available in the form of xml tags.

So for automatic information retrieval from these PubMed articles, there is a need to parse these xml tags. These articles contain various words which may occur in almost all the documents so there is a need to normalize the weight of these words so that we can put them in same level with the words occurred less frequently. The documents may belong to more than one cluster due to the fuzzy nature of words for appearing in more than one document, so there is a need to group these documents on a fuzzy basis with membership value in the interval of 0-1. If the clusters of documents is formed in linear space, then no problem, but in case of text documents the more fuzzy nature of inside words may form the cluster in a nonlinear space like a circle, ellipse, and spherical curves. So there is a need of a method which is able to manage the clusters in a nonlinear space as well as linear space.

3 Research Methodology

By taking care of the points raised in the research problem section we define the steps and prepare the model, which may be helpful to solve the issues in the field of text mining. In this paper following steps are applied to carry out the research- (a).Data Collection, (b).Pre-processing, (c).Normalization, (d) Document Clustering Methods, (e).Experimental study, (f).Results and Discussions. The research methodology of this paper is shown in Fig. (3)

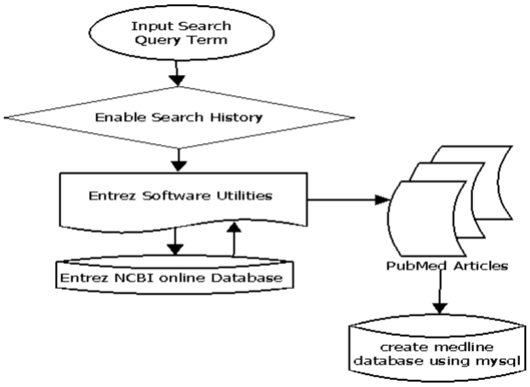

Input Search Query Term
Enable Search History
Entrez Software Utilities
Entrez NCBI online Database
PubMed Articles
create medline database using mysql

Figure 2- illustrates the extraction of PubMed articles using E-utilities.

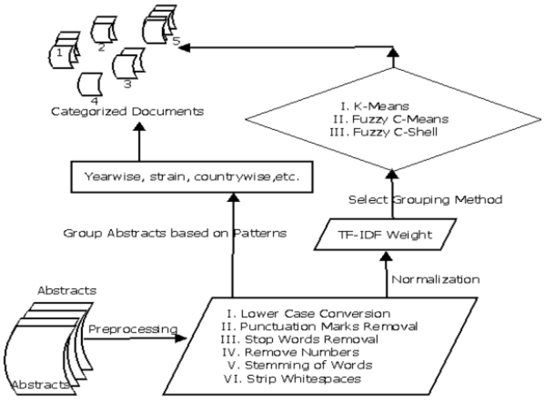


Figure 3- illustrates Research Methodology of the paper.

3.1 Data Collection

The Entrez Programming Utilities are used in this paper to automatically extract the influenza virus related PubMed articles. The Entrez Programming Utilities (E-utilities) consists of a set of nine server-side programs. It provides a stable interface into the Entrez query and database system at the National Centre for Biotechnology Information (NCBI). The E-utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve the requested data. The E-utilities are therefore the structured interface to the Entrez system, which currently includes 38 databases covering a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature. The E-utilities, access the core search and retrieval engine of the Entrez system and, therefore, it is only capable of retrieving data that are already in Entrez. Although the majority of data at NCBI is in Entrez but there are several datasets that exist outside of the Entrez system. Before beginning a project with the E-utilities, check that the desired data can be found within an Entrez database [3-4]. The entire process of data collection is shown in Fig. (2).

3.2 Information on Avian Influenza Virus

There are various diseases and topics that can be searched from PubMed articles, but in this study Influenza Virus related PubMed articles have been selected, because according to WHO Influenza Fact Sheet Nov 2014, Influenza Virus spreads around the world in a yearly outbreak, resulting in about three to five million cases of severe illness and about 250,000 to 500,000 deaths. Death occurs mostly in the young, the old and those with other health problems. Avian influenza refers to the disease caused by infection with avian (bird) influenza (flu) Type A viruses. These viruses occur naturally among wild aquatic birds worldwide and can infect domestic poultry and other bird and animal species. Avian flu viruses do not normally infect humans. However, sporadic human infections with avian flu viruses have occurred. The links below offer information about avian influenza for different audiences. Human infections with avian influenza viruses are rare and most often occur after people are in contact with an infected bird. However, non-sustained person-toperson spread of other avian influenza viruses was thought to have occurred in the past, most notably with avian influenza A (H5N1) viruses [13]. The nomenclature of influenza virus and strain designation of virus types can be found in memoranda [16].

A database named "medline" is created to analyze the information of PubMed articles with the help of Mysql5.6. The relation schema of one of the table of this medline database is shown in Table-1. This table contains 12 attributes in which Doc_no is used as primary key and pmid is used as unique key.

Table 1- illustrates PUBMED relation Schema of medline Database

	Field	Type	Null	Key	Default	Extra
1	Doc_no	Int (11)	No	PRI	(NA)	auto_increment
2	Id	Varchar (20)	Yes		(NA)	
3	Pmid	Varchar (10)	Yes	UNI	(NA)	
4	Title	Varchar (500)	Yes		(NA)	
5	Authors	Varchar (1000)	Yes		(NA)	
6	Country	Varchar (50)	Yes		(NA)	
7	Abstract	Varchar (10000)	Yes		(NA)	
8	Years	Varchar (20)	Yes		(NA)	
9	Pubdate	Varchar (20)	Yes		(NA)	
10	Issn	Varchar (20)	Yes		(NA)	
11	Pgno	Varchar (10)	Yes		(NA)	
12	Journal	Varchar (50)	Yes		(NA)	

3.3 Pre-processing

The automatically retrieved PubMed articles contain some character like “ ' ” mysql is not able to store these characters in is database tables. So we have to remove these characters as a first preprocessing step. The generalized regular expression which can be used to remove all types of punctuation marks is as- "punctuation = '[]\\?!\"\'#$%&(){}+*/:;,._`|~\\[<=>@\\^-]' ". Then we convert all the letters of words in lower case. The next step is to remove the stop words like articles(a, an, the), this, that, those, there, which are used as an identifier of nouns, pronouns, verb, adjectives in the grammatical sentences. These words don’t make any sense in the categorization of text documents. Stemming of words is performed to change the words and adjectives to their root form. So the words like 'virus' and 'viruses' should be considered as one word 'virus', similarly word 'vaccination', 'vaccinated', 'vaccines' are changed to one word 'vaccine'. The next step is to remove some sparse terms which have not occurred in the most of the documents.

3.4 Normalization

Term normalization is the process of mapping variants of a term to a single, standardized form and constitutes an important part of the process of automatic term recognition. As an example of term variation, consider the terms IL2, IL-2, and interleukin-2, which all represent the same biological concept. Term variation can present a major obstacle to the effective use of dictionaries by text mining systems – if term variants are not present in the dictionary, then the system may fail to recognize important terms occurring in the text, such as gene and protein names. In order to address this problem, term normalization methods can be used to map previously unseen variants of a term to known concepts. Simple normalization methods include converting capital letters to lower case letters, and deleting hyphens and spaces. A further approach is the use of soft string matching, which calculates similarity scores between strings [2]. Another approach for term normalization is TF-IDF weight score of each term [5, 14]. This Scheme uses Bag of Words Model for term vectors representation and doesn’t consider the ordering of words in a document [25-27]. In this scheme- "John is quicker than Mary" and "Mary is quicker than John” have the same vectors {"John", "is", "quicker", "than", "Mary"}. Each document is then represented as a vector in this space containing the frequency of all the words in that document. If the dimension are ('and', ‘the’, 'machine', 'learning'), then a document may be represented as (2, 3, 2, 0) the number indicates the frequency of the corresponding words in that document and using Boolean model (1,1,1,0) where the number indicates the presence(1) or absence(0) of a word in the document. In Vector-Space model the dimensions of vector-space is a count of all the documents in the corpus [14].

3.5 TF-IDF Weight

Since every document may be different in length, it is possible that a term would appear much more times in long documents than shorter ones. Term frequency (TF) is the count of repeated words which are strongly related to the contents of documents. Inverse document frequency (IDF) is the count of repeated words occurred most frequently in as many documents. So the TF-IDF increases the weight of uncommon term which occurred less frequently in other documents are more important than most frequent terms occurred in as many documents. Normalization by document length is required because long documents contain many distinct words and contain some same words as many times. So the term-weights for long documents should be reduced by dividing their TF with document length. The Eq. (1)-(3) are used to find TF, IDF, and TF-IDF weight of any term for the documents [5,14,25-27].

$$tf_{t,d} = 1 + log_{10}\left(\frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}\right) \quad (1)$$

$$idf(t) = log_{10}\left(\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}\right) \quad (2)$$

$$w_{t,d} = tf_{t,d} \times idf(t) \quad (3)$$

3.6 Words Association

If two terms co-occur within the same paragraph, they constitute an association, e.g. <term1, term2, associative frequency>. For terms (t_1, t_2 … . . t_k) the measure for the association {(t_1, t_2 … . . t_k)} is as – k (t_1,t_2 … . . t_k) = Sentence Frequency (t_1, t_2 … . . t_k) X idf (t_j) where j = 1, 2 … k. The most frequently occurring terms in different-different stories or documents, these words are associated with other words to a specified threshold value, e.g. word “vaccine” is associated with two words “virus” and “influenza” at 0.21 threshold value [14].

3.7 Text Document Clustering

Document clustering (or text clustering) is the application of cluster analysis to textual documents. It has applications in the automatic document organization, topic extraction and fast information retrieval or filtering. In general, there are two common algorithms. The first one is the hierarchical based algorithm, which includes single link, complete linkage, group average and Ward's method. By aggregating or dividing, documents can be clustered into a hierarchical structure, which is suitable for browsing. However, such an algorithm usually suffers from efficiency problems. The other algorithm is developed using the K-means algorithm and its variants. These algorithms can further be classified as hard or soft clustering algorithms. Hard clustering computes a hard assignment – each document is a member of exactly one cluster. The assignment of soft clustering algorithms is soft – a document’s assignment is a distribution over all clusters. In a soft assignment, a document has fractional membership in several clusters. Dimensionality reduction methods can be considered a subtype of soft clustering; for documents, these include latent semantic indexing (truncated singular value decomposition on term histograms) and topic models. Other algorithms involve graph based clustering, ontology supported clustering and order sensitive clustering. Clustering can be beneficial to automatically derive human-readable labels for the clusters. The cosine similarity between two documents can be measured by Eq. (4) and Eq.(5). The Jaccard coefficient shown in Eq. (6) may be helpful in case of the Boolean BOW model to find the similarity between two documents. These measures are the base for further clustering of documents and their mathematical expressions are as follows,

Cosine Similarity,

$$\mathrm{Sim}(d_1, d_2) = \frac{\vec{v}(d_1).\vec{v}(d_2)}{|\vec{v}(d_1)|\,|\vec{v}(d_2)|}\,|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{M}(\,x_i - y_i\,)^2} \tag{4}$$

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q}\,.\,\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\cdot\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}} \tag{5}$$

Jaccard Coefficient,

$$Jaccard\ (A, B)\ =\ |A\ \cap\ B|\ /\ |A\ \cup B| \tag{6}$$

Algorithm 1: K-means algorithm

```
K-Means($\{\vec{x}_1, \vec{x}_2,\ldots, \vec{x}_N\}$,k)
  $(\vec{s}_1, \vec{s}_2,\ldots, \vec{s}_k) \leftarrow$ SELECT RANDOM SEEDS ($\{\vec{x}_1, \vec{x}_2,\ldots, \vec{x}_N\}$,k)
  for k← 1 to k
  do $\vec{\mu}_k \leftarrow \vec{s}_k$
  while stopping criterion has not been met
   do for k← 1 to k
   do $\omega_k \leftarrow \{\}$
   for n←1 to N
     do j← $arg\ min_{j'}\ |\vec{\mu}_{j'} - \vec{x}_n|$
      $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$   (reassignment of vectors)
     for k← 1 to k
     do $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|}\sum_{\vec{x}\in\omega_k}\vec{x}$   (recomputation of centroids)   (7)
     return $\{\vec{\mu}_1, \vec{\mu}_2, \ldots, \vec{\mu}_k\}$
```

- For most IR applications, the vectors

Algorithm 2: Fuzzy C-means Clustering Algorithm

1. Input Data Matrix X, whose columns represent Variables and rows represent Values of the variables.

2. Fix the number of clusters c, c < n, where n is the number of data points. Fix $m = 1$ for hard memberships or $0 < m < 1$ for fuzzy memberships.

3. Set iteration counter $j = 0$. Initialize the fuzzy c-partition $U=[u_{ij}]$ *matrix*, $U^{(0)}$.

4. Calculate the cluster center vectors $C^{(k)} = [C_i]$ with $U^{(k)}$

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}{}^{m} x_j}{\sum_{j=1}^{n} u_{ij}{}^{m}} \tag{8}$$

5. Calculate the distance measure $d_{ij}{}^2$ from data point X_j to cluster center C_i

$$d_{ij}{}^2 = (X_j - C_i)^{\mathrm{T}} A_i (X_j - C_i) \tag{9}$$

6. Update the membership $U^{(k+1)}$, *and* $U^{(k)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left[\frac{\|x_j - c_i\|}{\|x_j - c_k\|}\right]^{\frac{2}{m-1}}} \tag{10}$$

7. If $\|\mathrm{U}^{(k+1)} - \mathrm{U}^{(k)}\| < \epsilon$, then converged, stop; otherwise set i = i + 1 and go to step 4

Algorithm 3: Fuzzy C-shell Clustering Algorithm

1. Input Data Matrix X, whose columns represent Variables and rows represent Values of the variables.

2. Fix the number of clusters c, c < n, where n is the number of data points. Fix $m = 1$ for hard memberships or $0 < m < 1$ for fuzzy memberships.

3. Fix the radius r of the circle whose center is C_i

4. Set iteration counter $j = 0$. Initialize the fuzzy c-partition $U=[u_{ij}]$ *matrix*, $U^{(0)}$.

5. Calculate the cluster center vectors $C^{(k)} = [C_i]$ with $U^{(k)}$

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}{}^{m} X_j}{\sum_{j=1}^{n} u_{ij}{}^{m}}$$

6. Calculate the distance measure $d_{ij}{}^2$ from data point X_j to cluster center (C_i, r)

$$d_{ij}{}^2 = ((X_j - C_i) - r)^{\mathrm{T}} \quad A_i((X_j - C_i) - r) \tag{11}$$

7. Update the membership $U^{(k+1)}$, *and* $U^{(k)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left[\frac{\|X_j - C_i\|}{\|X_j - C_k\|}\right]^{\frac{2}{m-1}}}$$

8. If $\|U^{(k+1)} - U^{(k)}\| < \epsilon$, *then* converged, stop; otherwise set $i = i + 1$ and go *to* step 5.

3.8 Experimental Study

All the experiments have been carried out with the ASUS Laptop with specification as core i7, 8GB RAM, 2.4GHz Processor in UBUNTU 14.04 64-bit OS. The author has used Mysql5.6, Python2.7 and R-3.1.2 to store the articles with their concerned information, process the biological words, and for statistical analysis respectively. The Entrez software utilities are used to fetch influenza virus related PubMed articles from NCBI web page. The search history should be enabled true, because only 500 articles can be extracted in one request. The Algorithm 4, extracts the PubMed articles based on a search query term from the NCBI web page and to store the retrieved data in a table PUBMED of relational database medline.

Algorithm 4: An automatic extraction of PubMed articles from NCBI web page

```
1. input= Email Id
2. input query="influenza[mesh]+virus[mesh]"
3. Enable search history
3. Use Entrez software utilities to extract all PubMed articles information as per query term
4. do for each PubMed article
5. Store id, pmid, title, authors, country, abstract, year, date, issn, journal, and page number in
separate variables;
6. Open Mysql Database connection with medline database
7. CREATE TABLE PUBMED to store all above variables.
8. Insert records into PUBMED table
9. return ($X_1$ $X_2$,... $Xn$) PubMed articles
```

Algorithm 5: Counting of journals that published articles related with term 'influenza virus' from NCBI web page.

```
START
1. input search query term=term.
2. Open Mysql Database connection with medline database
3. do for each PubMed article in PUBMED table
4. select journal, Count(*) from PUBMED where abstract like '%term%' and years between 2010
   and 2015
5. return ( $J_1$, $J_2$ ... $Jm$) Journals related to search term.
6. Plot the result using Bar plot.
END
```

3.9 Results and Discussions

The plot for a number of PubMed articles containing the terms H7N9, H1N1, and H5N1 is shown in Figs. (4)- (6) of this section. The count of Journals in which influenza virus related PubMed articles have been published between years 2010 and 2015 can be found using Algorithm 5. This result is shown in Fig. (7) of this section using bar plot between counts of journals which published influenza virus related articles and journal names. The abstracts are categorized based on terms, country, and year of publications, etc., then basic K-means clustering algorithm is used to categorize the abstracts. The details of the algorithm are shown Algorithms 1-3, as k-means is not able to deal with the fuzzy nature of words, and fuzzy C-means solves this issue and forms the cluster of abstracts which are in linear space. However, the fuzzy C-means is not able to form the cluster of abstracts which are in a nonlinear space. The detail of this algorithm is shown in Algorithm 2. To solve this issue, a new fuzzy approach is proposed named fuzzy C-Shell clustering shown in Algorithm 3, which is able to form the clusters in linear as well as nonlinear space. In this paper fuzzy C-Shell clustering algorithm is able to form the cluster of abstracts which are in a nonlinear space in the form of circle only.

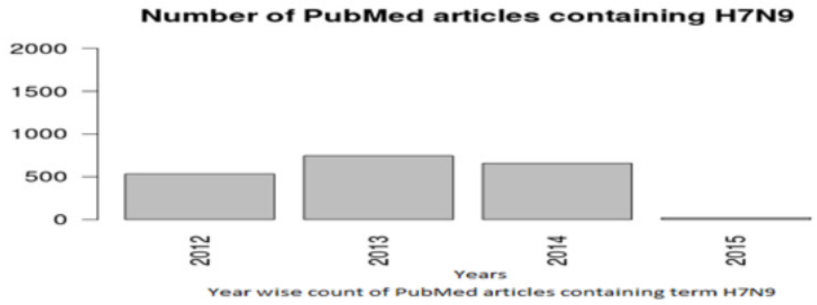


Figure-4: illustrates year wise count of PubMed articles containing term H7N9

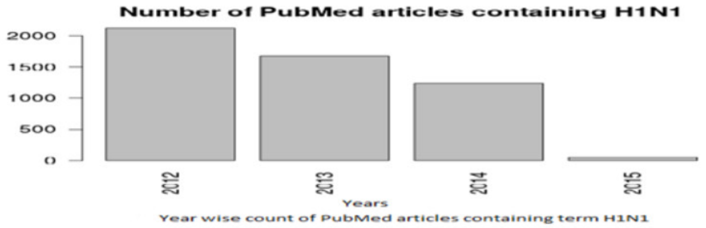


Figure-5: illustrates year wise count of PubMed articles containing term H1N1

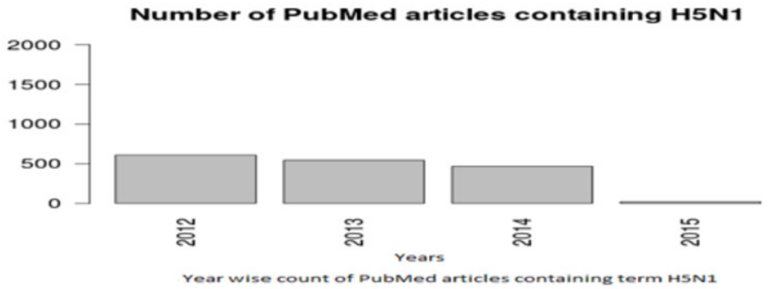


Figure-6: illustrates year wise count of PubMed articles containing term H5N1

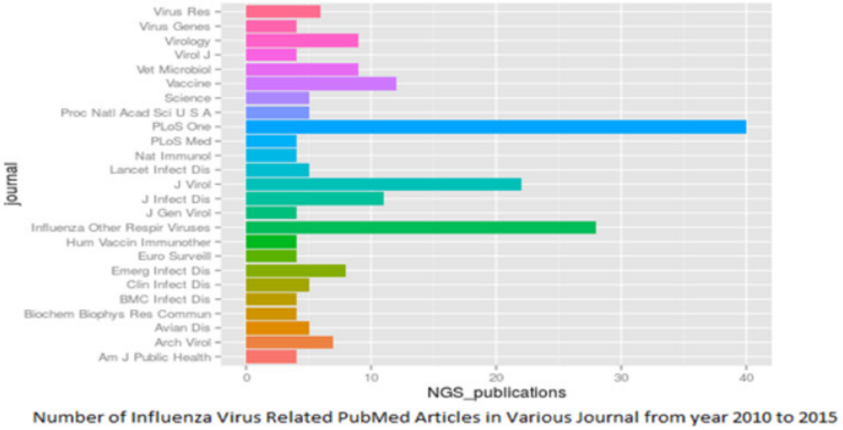


Figure-7: illustrates influenza virus related pubmed articles published in various journals from year 2010-2015

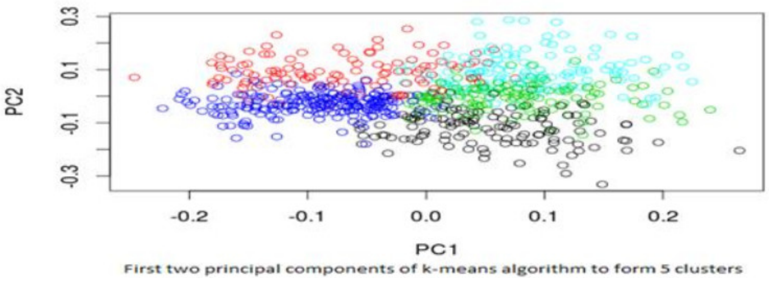


Figure-8: illustrates first two principal components of k-means clustering to form 5 clusters

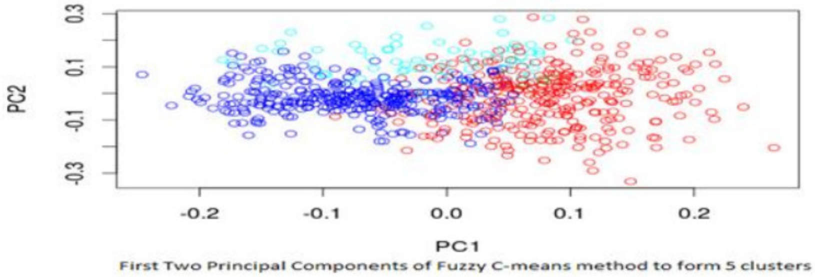


Figure-9: illustrates first two principal components of Fuzzy C-means clustering to form 5 clusters

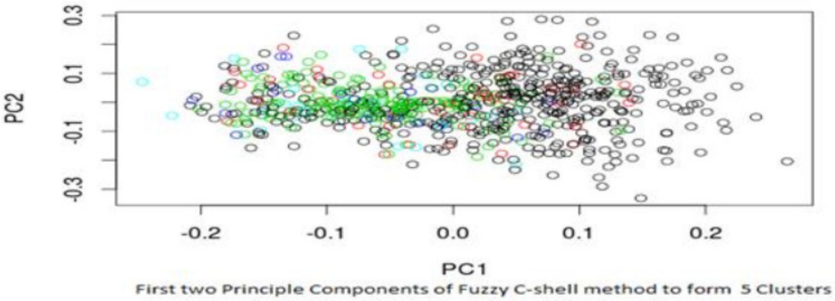


Figure-10: illustrates first two principal components of Fuzzy C-shell clustering to form 5 clusters

There are total 1000 abstracts downloaded automatically in this paper. The dimension of this corpus is extremely large and the results cannot be visualized in two dimensions. The principal component analysis method is used in this situation to visualize the results of clustering algorithms used in this paper. As the principal components show the direction of maximum variances in the dataset and first two principal components covers the almost maximum (96%) variance [15] in the data, only the first two principal components PC1 and PC2 are used here to show the 5 number of clusters in the corpus of 1000 abstracts.

It is very much clear from the results shown in Fig 6, Fig 7, and Fig 8 that the clustered abstracts are very much overlapped in K-Means clustering, while less overlapped in fuzzy C-Means clustering method. The overlapping of clustered abstracts is very less in proposed fuzzy C-Shell clustering method.

4 Conclusion

Complex Biomedical terms extraction systems can help researchers in a number of ways. The semantic information extraction due to the enormous volume of biological literature demands computational methods to allow pertinent information to be found and analyzed efficiently. Text Mining facilitates the retrieval of semantic information such as entities (proteins, genes, diseases, etc.) and events (binding, distillation, regulation, etc.) in which the entities and events participate to make the sentence meaningful in biomedical documents. The Text Mining based information retrieval systems are now sufficiently accurate to support the development of various user-oriented applications, including sophisticated semantic search which allow for more efficient retrieval of relevant information than traditional keyword-based searching methods, provide means for linking biological words pathways with supporting evidence in the literature, helpful in tasks such as the semi-automatic curation of biomedical databases and ontologies.

The proposed database Text Mining system is helpful to analyze the online records of offline period, reduces the overheads involved with the online text retrieval and analysis. As the biomedical terms are very much fuzzy in the nature such as more than one term may represent the same concept and may be found in as many documents. Thus the proposed fuzzy clustering algorithms which is useful in clustering of text documents in a linear as well as nonlinear spaces, would be helpful for Text Mining based information retrieval systems.

REFERENCES

[1] Sophia, Ananiadou, "Automated Term Recognition", Encyclopedia of Systems Biology, Springer, 2013, pp 57-59.

[2] Sophia, Ananiadou, "Term Normalization, Text Mining", Encyclopedia of Systems Biology, Springer, 2013, p 2155.

[3] Eric,Sayers, A General Introduction to the E-utilities, http://www.ncbi.nlm.nih.gov/books/NBK25501/, NCBI, August 9, 2013.

[4] luwening,Accessing NCBI’s Entrez databases, http://nbviewer.ipython.org/github/gumption/Using_Biopython_Entrez/blob/master/Biopython_Tutorial_and_Cookbook_Chapter_9.ipynb, BioPython Documentation, 2011.

[5] Michal Toman, Roman Tesar, Karel Jezek,"Influence of Word Normalization on Text Classification", http://textmining.zcu.cz/publications/inscit20060710.pdf, 2007.

[6] Sophia, Ananiadou, "Term Disambiguation, Text Mining", Encyclopedia of Systems Biology, Springer, 2013, pp 2154-2155.

[7] Rajesh, N. Dave, and Kurra, Bhaswan. "Adaptive Fuzzy c-Shells Clustering and Detection of Ellipses", IEEE Transactions On Neural Networks, VOL. 3, NO. 5, September 1992, 643-662.

[8] Kevin, W. Boyack, David, Newman, Russell, J. Duhon, Richard, Klavans, Michael, Patek, Joseph, R. Biberstine, Bob, Schijvenaars, Andre, Skupin, Nianli Ma, Katy, Borner. “Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches”, PLOS ONE, 17 Mar 2011, doi/10.1371/journal.pone.0018029.

[9] Senay, Kafkas, Jee-Hyub Kim, Johanna R. McEntyre. “Database Citation in Full Text Biomedical Articles”, PLOS ONE, 29 May 2013, doi/10.1371/journal.pone.0063184.

[10] Sophia, Ananiadou, Paul, Thompson, Raheel, Nawaz, John, McNaught, and Douglas, B. Kell, "Eventbased text mining for biology and functional genomics", Briefings in Functional Genomics, June 6, 2014, pp 1-18.

[11] YangYan, Lihui Chen, William-Chandra Tjhi, Fuzzy semi-supervised co-clustering for text documents, Fuzzy Sets and Systems, Elsevier, 215 (2013), pp 74–89.

[12] Jensen K, Panagiotou G, Kouskoumvekaki I (2014), "Integrated Text Mining and Chemoinformatics Analysis Associates Diet to Health Benefit at Molecular Level", PLoS Comput. Biol. 10(1): e1003432. doi:10.1371/journal.pcbi.1003432 [13] http://www.cdc.gov/flu/avianflu/index.htm.

[13] Agnihotri, Deepak, Verma, Keshri, and Tripathi, Priyanka, "Pattern and Cluster Mining on Text Data", 2014 Fourth International Conference on Communication Systems and Network Technologies, IEEE Computer Society, 978-1-4799-3070-8/, DOI 10.1109/CSNT.2014.92, p 428-432.

[14] Sunghae, Jun, Sang-Sung, Park, Dong-Sik Jang, "Document clustering method using dimension reduction and support vector clustering to overcome sparseness", Expert Systems with Applications, Elsevier, 41(2014), pp 3204–3212.

[15] Memoranda, A revision of the system of nomenclature for influenza viruses: a WHO Memorandum, Bulletin of the World Health Organization, 58 (4): 585-591 (1980).

[16] Edvard, G Randell, "Influenza Virus Types, Subtypes, and Strains", PA Pandemic influenza virus, planning summit 2006.

[17] Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, Jun’ichi Tsujii, " GENIA Corpus Manual Encoding schemes for the corpus and annotation ", http://www-tsujii.is.s.u-tokyo.ac.jp/TR/, 2006.

[18] Yang, M.S., A Survey of Fuzzy Clustering, Mathl. Comput. Modelling ,Vol. 18, No. 11, 1993, pp. 1-16

[19] Sumit Goswami, and Mayank Singh Shishodia, " A Fuzzy Based Approach To Text Mining And Document Clustering ", Cornell University Library, CoRR arXiv:1306.4633, june 2013.

[20] T. Theodosiou, N. Darzentas, L. Angelis and C. A. Ouzounis, "PuReD-MCL: a graph-based PubMed document clustering methodology", Bioinformatics Data and Text Mining, Oxford University Press, Vol. 24 no. 17 2008, pages 1935–1941,doi:10.1093/bioinformatics/btn318.

[21] Fei Zhua, Preecha Patumcharoenpol, Cheng Zhanga, Yang Yanga, Jonathan Chanc, Asawin Meechaie, Wanwipa Vongsangnaka, Bairong Shena, "Biomedical text mining and its applications in cancer research", Journal of Biomedical Informatics, Volume 46, Issue 2, April 2013, Pages 200–211, doi:10.1016/j.jbi.2012.10.007.

[22] Rodriguez-Esteban R (2009) Biomedical Text Mining and Its Applications. PLoS Comput. Biol. 5(12): e1000597. doi:10.1371/journal.pcbi.1000597.

[23] Cohen KB, Hunter L (2008) Getting Started in Text Mining. PLoS Comput. Biol. 4(1): e20. doi:10.1371/journal.pcbi.0040020.

[24] Rzhetsky A, Seringhaus M, Gerstein MB (2009) Getting Started in Text Mining: Part Two. PLoS Comput. Biol. 5(7): e1000411. doi:10.1371/journal.pcbi.1000411.

[25] D. Agnihotri, K. Verma, and P. Tripathi, "Computing symmetrical strength of ngrams:a two pass filtering approach in automatic classification of text documents,"*SPRINGERPLUS*, vol. 5, no. 942, pp. 1–29, 2016.

[26] D. Agnihotri, K. Verma, and P. Tripathi, "Computing correlative association of terms for automatic classification of text documents," in Proceedings of the Third International Symposium on Computer Vision and the Internet. ACM, 2016, pp. 71–80.

[27] Deepak Agnihotri, Kesari Verma, Priyanka Tripathi, Variable Global Feature Selection Scheme for automatic classification of text documents, Expert Systems with Applications, Volume 81, 15 September 2017, Pages 268-281, ISSN 0957-4174, http://doi.org/10.1016/j.eswa.2017.03.057.