BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(2)

Clustering of PubMed abstracts using nearer terms of the domain

Mary Rajathei David & Selvaraj Samuel*

Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli-620024, India; Selvaraj - Email: selvarajsamuel@gmail.com; *Corresponding author

Received December 26, 2011; Accepted December 28, 2011; Published January 06, 2012

Abstract:
Literature search is a process in which external developers provide alternative representations for efficient data mining of biomedical literature such as ranking search results, displaying summarized knowledge of semantics and clustering results into topics. In clustering search results, prominent vocabularies, such as GO (Gene Ontology), MeSH(Medical Subject Headings) and frequent terms extracted from retrieved PubMed abstracts have been used as topics for grouping. In this study, we have proposed FNeTD (Frequent Nearer Terms of the Domain) method for PubMed abstracts clustering. This is achieved through a two-step process viz; i) identifying frequent words or phrases in the abstracts through the frequent multi-word extraction algorithm and ii) identifying nearer terms of the domain from the extracted frequent phrases using the nearest neighbors search. The efficiency of the clustering of PubMed abstracts using nearer terms of the domain was measured using F-score. The present study suggests that nearer terms of the domain can be used for clustering the search results.

Keywords: domain knowledge, nearer term, clustering, nearest neighbors search, PubMed abstracts;

Background:
The deposition of biological literature into the NCBI's PubMed (http://www.pubmed.gov/) database has increased tremendously in recent years due to fast developments in science and technology. The PubMed is the primary source of abstracts of peer-reviewed biomedical information for researchers in making scientific discoveries and healthcare professionals in managing health-related matters **[1].** The PubMed search engine's rapid responses and integration with other NCBI-hosted databases such as GenBank allow PubMed to provide broad, up-to-date and curated search results. However, a wide variety of users, ranging from those researching results of clinical trials to those examining new scientific discoveries means that PubMed is unable to fulfill the researcher's need while searching and browsing large volumes of literature covering one's specific area of interest. In response to that, the NCBI is continuously making changes in PubMed web services for improvement. In addition to that, the availability of the PubMed database web services opened up the possibility for external developers to provide alternative representations of the biomedical literature for effective knowledge management such as ranking search results **[2, 3, 4, 5]** , displaying summarized knowledge of semantics **[6, 7]** and clustering results into topics **[8, 9, 10, 11, 12].**

Clustering is one feature that groups the search results based on information extracted from the collection. Search Engines such as Textpresso **[8],** XplorMed **[9],** semedico **[12],** novo/seek **[11]** and GoPubMed **[10]** use the controlled vocabularies, such as Gene Ontology (GO), Medical Subject Headings (MeSH) **[13],** Systematized Nomenclature of Medicine (SNOMED), and Unified Medical Language System (UMLS) **[14],** as information resource for topics extraction from search results. However, these vocabularies focus on a particular domain; for example, GO for gene products and MeSH for medical topic and disease. The grouping has been according to the terms in the controlled vocabularies. Informative terms or phrases extracted from the retrieved abstracts are used for grouping the search results

which offer a better understanding about the area of research **[15].** Zamir and Etioni **[16]** have proposed to use a suffix-tree based clustering algorithm (STC) to identify the common phrases shared by the documents. Smith **[17]** has demonstrated the usefulness of suffix tree clustering in browsing events in unstructured text. Readable and unambiguous descriptions of the thematic groups are an important factor of the overall quality of clustering. These provide the users an overview of topics covered in the search results and help them to identify the specific group of documents they were looking for. The LINGO algorithm **[18]** employs suffix arrays and singular value decomposition (SVD) to capture thematic labels in a search result for clustering. A Carrot framework was created to facilitate clustering the search results by including algorithms such as STC and LINGO **[19].**

Domain knowledge could play an important role in knowledge management and discovery. The knowledge of the domain gives an idea of the search results when no prior knowledge about the collection exists **[20, 21].** In a clustering of documents, domain knowledge helps to improve mining efficiency as well as the quality of mined knowledge **[22].** Tsoi et al **[23]** suggested that terms that are frequently occurring with the domain have some meaning in the biomedical literature and provide knowledge of the domain. In the present work, we have proposed FNeTD method that combines frequent multi-word extraction and nearest neighbors search for clustering retrieved documents. To implement this, an algorithm has been introduced to extract frequently occurring multi-word term phrases. Then, the terms that come along with the domain are identified from the extracted multi-word terms by following nearest neighbor's search **[24].** A user-friendly search interface was created to narrow down the search according to nearer terms of the domain. The proposed method was tested by extracting nearer terms of "p53" from the search results which has about 50,000 PubMed abstracts. The efficiency of the method for extracting relevant terms of domain was compared with actual terms of the domain and measured using F-score. The present study suggests that nearer terms of the domain can be used for effective grouping of search results.

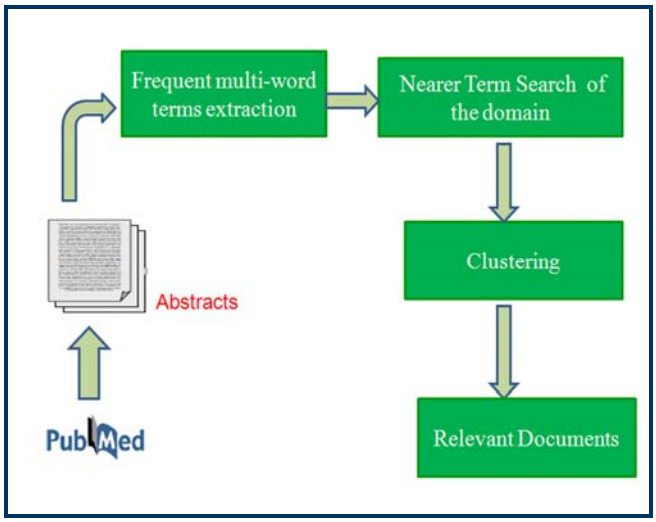

Figure 1: System overview of clustering of PubMed abstracts using nearer terms of the domain

Methodology:

For clustering the search results using domain knowledge, frequently co-occurring nearer terms of the domain have to be extracted. The nearer terms of the domain are identified from the frequently occurring multi-word terms that are present in the PubMed abstracts. The system overview of clustering of PubMed abstracts using nearer terms of the domain is illustrated in **Figure 1.** The entire process was performed using an in-house JAVA program with SUN ULTRA 40M2 workstation.

Preparation of PubMed abstracts

The search results of the given input query were downloaded from NCBI PubMed in XML format. In pre-processing step, the stop words in the each sentence of the PubMed abstracts were removed using rule based approach. Then, the entities such as PubMed Id, title and the processed abstracts were stored in the database.

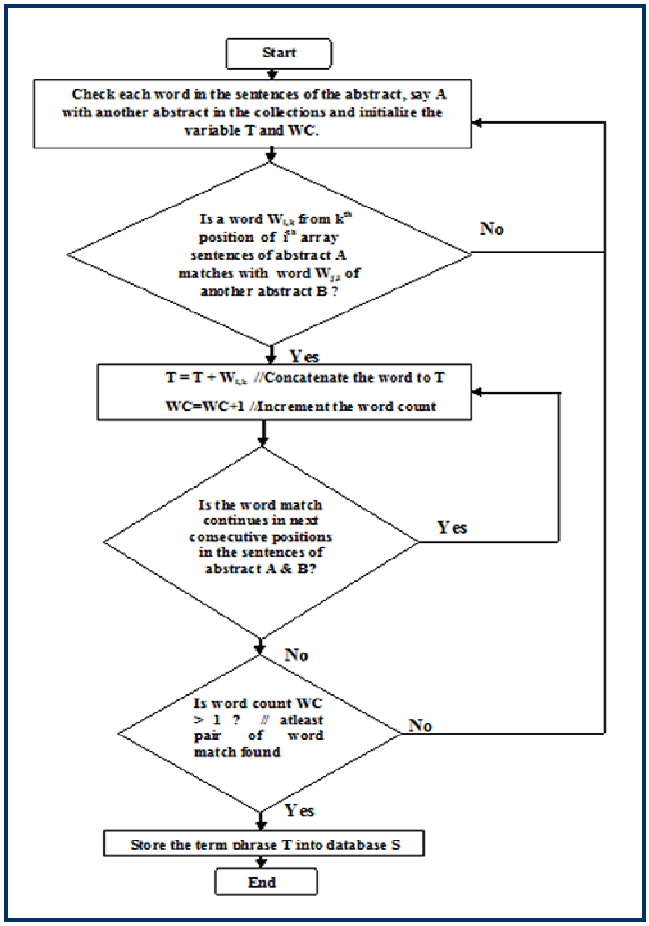

Figure 2: Frequent multi-word term extraction algorithm. The flowchart explains the steps involved in the extraction of multi-word terms from each of the abstract. The computational steps involve comparing two abstracts for the identification of single match, extension of the word match and, storing the commonly occurring multi-word terms into Database S.

Frequent multi-word term extraction

An algorithm **(Figure 2)** was implemented to obtain all frequent terms that are present in more than one abstract. The algorithm

reads one abstract at a time in the collection and splits it into an array of sentences. Then, it tokenizes each sentence into an array of words and initiates the search for exact word match in another abstract in the collection. The steps to be followed for finding frequent multi-word terms are as follows: i) If a word in the sentence of abstract A is found to match in another abstract say B; then tokenizing the word containing sentence of abstract B into array of words for finding maximum word match. The search for word match is extended to the next consecutive position in the word containing sentences of the abstract A and B until the maximum match is found. However, at each consecutive position extensions in the sentence of the abstract, the algorithm checks whether the end of the sentence is reached. If atleast a pair of words match was found in two abstracts then it will be stored in to database S.
ii) If a word in the sentence of abstract A is not found match in abstract B then next abstract in the collection is considered. Steps (i) and (ii) are to be followed for each word in the sentences of the abstracts in the collection.

Multi-level extraction of nearer terms of the domain
The nearer terms of the domain are then identified from the stored multi-word terms using nearest neighbors search. Here, we define nearest neighbors search as one that searches for the input (domain) 't' in a set of stored multi-word terms stored in the database 'S' and find the closest terms in S to t. A JAVA program was developed to extract nearer terms domain from the stored multi-word terms that contained domain in the first level and, co-occurring terms of nearer terms from stored multi-word terms that contained nearer terms in the next level. The extracted terms are then stemmed according to Porter Stemming algorithm **[25].**

Visualization of nearer terms for clustering
In order to cluster PubMed abstracts according to nearer terms of the domain, a web based framework for displaying nearer terms and sub-terms of the domain in the form of hierarchical tree as well as hyper tree view was created using script program Active server page (ASP). The hierarchical tree view is to display all nearer terms and sub-terms of the domain. The hyper tree view is to display the selected starting single character alphabet or two character alphabets of nearer terms of the domain. The web based framework enables the user to cluster the retrieved PubMed abstracts according to the terms selected from the display.

Measurement accuracy of the nearer terms
In document based clustering, the documents are clustered according to a certain similarity measure which usually yields non-overlapped clusters. The clusters quality was measured in terms of intra-cluster similarity and inter-cluster dissimilarity **[26].** However, in label based clustering, the documents are clustered according to informative labels extracted from the related documents and evaluated in terms of precision and recall of the labels **[27].** In this study, nearer terms of the domain are used as labels for PubMed abstracts clustering and hence, the extracted terms are evaluated in terms of precision and recall. The subject index from the book "25 years of p53 research" **[28]** was taken as a reference for the relevant terms of the domain. The precision and recall are defined here in terms of a set of retrieved terms of the domain from the PubMed abstracts and a set of relevant terms of the domain. The F-score measure considers both the precision and the recall to test the accuracy and it was computed using formula:

$$F = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The precision and recall are computed using formula:

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Where tp = number of correctly identified relevant term (true positive); fp = number of incorrectly identified relevant term (false positive); fn = number of relevant terms that are not identified (false negative)

Results:
We have taken the research articles for the query "p53" as input for the experimental study. The number of abstracts downloaded from PubMed as on 1st May 2011 was 53613.

Frequent multi-word terms
The SUN ULTRA 40M2 workstation system took 20 hours to extract all frequent multi-word terms that are present in the 53613 PubMed abstracts of "p53" and 1,24,000 distinct multi-word terms were extracted using our developed algorithm. The computational time required for finding frequent multi-word term in the abstract collections depends on the number of abstracts and number of sentences containing frequently occurring terms. The developed algorithm simply checks each word match in the selected abstract with another abstract in the collection. This simple way of extraction suggests that the algorithm can easily identify frequently occurring multi-word terms that present in the large collections of related documents.

Multi-level extraction of nearer terms of the domain
The terms that are nearer to the domain "p53" and sub-terms that are coming along with the nearer term were extracted from the stored multi-word terms using nearest neighbors search approach. For example, the nearer term "apoptosis" of p53 was identified from the stored multi-word terms contained both "apoptosis" and "p53". The nearer terms of "apoptosis" such as "bax", "DNA damage", "cancer" and "growth arrest" in the next level were also identified from the stored multi-word terms. The distinct multi-word terms that contain both "apoptosis" and "p53" and, multi-word terms that contain "apoptosis" and related terms are shown in Supplementary **Table 1 (see supplementary material)** Likewise all nearer terms and sub terms of the domain "p53" were identified from the stored multi-word terms.

Clustering using nearer terms of the domain
The purpose of extracting nearer terms of domain is to help the user who doesn't have any prior knowledge about the domain to gain the knowledge of commonly co-occurring terms of the domain. This knowledge helps them to understand about the domain and narrow down their search and retrieval. The nearer terms of "p53" are displayed in the form of a structured multi-level hierarchical tree shown in leftmost panel of **Figure 3** and,

the rightmost panel exhibits a screen shot hyper tree overview of all nearer terms starting with "AP". From the leftmost panel display in **Figure 3**, one can understand the terms that come nearer to "p53" and, from right most panel display in **Figure 3**, one gets the knowledge of the nearer terms of "p53" that start with "AP". Using this knowledge, one can narrow down their search and retrieval. For example, the user can easily understand that the term "apoptosis" is relevant to "p53". If the user wants to read the abstracts that discuss both "apoptosis" and "p53" then they can get them by clicking "apoptosis" from the display. This process also helps the user to get the set of clustered abstracts.

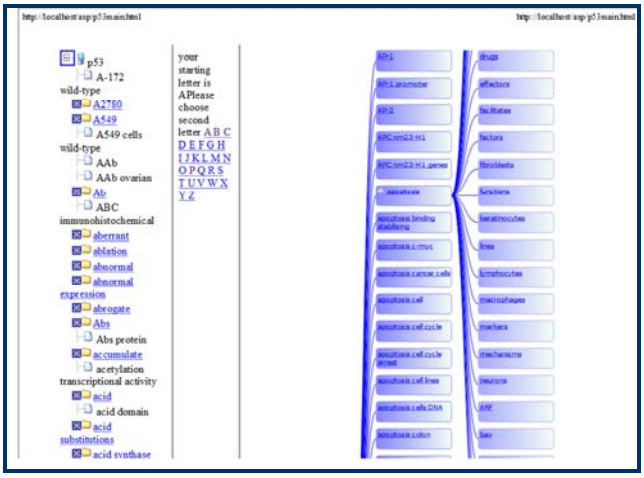

Figure 3: Snapshot overview of nearer terms of the p53 in the form of hierarchical tree as well as hyper tree view.

Evaluation of nearer terms of the domain

Since clustering using technical terms acquired from retrieved results offers a better understanding about the search results **[15],** the number of terms obtained by using FNetD and number of informative terms obtained by using STC and LINGO available under open source Carrot frame work (http://project.carrot2.org) were compared with the actual terms of the domain **[26]** for the 200 PubMed abstracts of "p53". **Table 2 (see supplementary material)** shows the accuracy of the terms obtained for clustering using various methods. The precision rates obtained in different methods have not shown much difference. However, the recall rates have shown that the present FNeTD approach is capable of retrieving more technical terms relevant to search results than STC and LINGO. The F-score obtained using FNeTD (0.36) is significantly greater than F-score of STC (0.10) and LINGO (0.16) which implies that nearer terms of the domain can be used for clustering the search results of PubMed abstracts.

Discussion:

In general, frequently occurring terms that are extracted from the related documents are used as labels for clustering. The methods used for finding frequently occurring terms have required some representations. For example, the STC method uses suffix tree to identify the set of frequently occurring terms in the collections and, LINGO employs suffix arrays to discover frequent phrases and singular value decomposition (SVD) to obtain informative labels. However, the FNeTD method does not require any representation for identifying labels for clustering. The developed algorithm simply checks each word in the sentence of an abstract with another abstract for finding frequent multi-word terms that are present in the related documents collection. The nearer terms of the domain are then identified from the extracted multi-word terms using nearest neighbors search **[24].** When comparing the accuracy of the terms obtained for clustering using STC and LINGO with FNetD, we have found that FNeTD method lists out more terms relevant to search results for clustering than STC and LINGO. The display of frequently co-occurring terms of the domain as labels for clustering enables the user from any unrelated area can able to understand the nearer terms of the domain. The extraction of nearer terms of the domain in more than one level allows the user to get additional information such as, association of p53 with apoptosis and apoptosis with cell cycle. This type of multi-level display of the terms helps to get the connection among terms of the domain which helps to improve the retrieval of relevant documents of one's interested in a specific area quickly.

Conclusion:

In this paper, a new method for clustering the search results was introduced. The novelty of the approach is nearer terms of the domain used as integrating resource for categorizing the retrieved abstracts. The idea behind the frequent nearer terms of the domain extraction is that terms that come nearer to the domain have some meaning in the biological literature and gives knowledge of the domain. We have observed that nearer terms of the domain provide more technical terms related to search results of the domain than frequently occurring terms in the collections. The generated nearer terms of the domain can be used as initial term list for domain ontology development. The present approach is more applicable to scientific related literature, since we obtained higher recall rate while handling the search results of PubMed compared to STC and LINGO. We believe that our methodology for post-processing could be improved by doing more stemming to reduce the size of the indexing terms for clustering. Future studies will aim to train the system to generate high quality resource of domain knowledge comparable to human expert hand-curated one.

References:

[1] Lu Z. *Database* 2011 doi:10.1093/database/baq036.

[2] Yu H *et al. BMC Bioinformatics* 2010 **11**: S6 [PMID: 20406504].

[3] Fontaine JF *et al. Nucleic Acids Research* 2009 **37**: W141

[4] States DJ *et al. Bioinformatics* 2009 **25**: 974 [PMID: 18326507].

[5] Plikus MV *et al. BMC Bioinformatics* 2006 **7**: 424 [PMID: 17014720].

[6] Kim JJ *et al. Bioinformatics* 2008 **24**: 1410 [PMID: 18400773].

[7] Ohta T *et al. Proc COLING/ACL* Australia 2006.

[8] Muller MH *et al. PLoS Biology* 2003 **2**: e.309 [PMID: 15383839].

[9] Perez-Iratxeta C *et al. Nucleic Acids Research.* 2003 **31**: 3866 [PMID: 12824439].

[10] Doms A & Schroeder M. *Nucleic Acid Research* 2005 **33**: W783

[11] Allende RA. *Nat Methods* 2009: 6.

[12] Wermter J *et al. Bioinformatics* 2009 **25**: 815 [PMID: 19188193].

[13] Nelson SJ *et al*. Relationships in Medical Subject Headings (MeSH) in Relationships in the organization Of knowledge Kluwer Acdamic Publisher, Netherlands, 2001 **171**: 184.
[14] Bodenreider O. *Nucleic Acids Res*earch 2004 **32**: D267
[15] Yeh R *et al*. *Proc on Discovery science* Japan 2007 : 291
[16] Zamir O & Etzioni O. *Proc ACM/SIGIR* New York 1998: 46
[17] Smith DA. *Proc ACM/SIGIR New York*. 2002 : 73
[18] Osinski S & Weiss S. *Proc Intelligent Information Processing and Web Mining,* Berlin, Springer-Verlag, 2004: 369
[19] Osinski S & Weiss D. *Proc International Atlantic Web Intelligence Conference,* Berlin, Springer-Verlag, 2005
439: 444.
[20] Bhavani SK. *Extended Abstractions Proc CHI* 2002 610
[21] Jiang X & Tan AH. *Proc IEEE on Data mining USA* 2005.
[22] Tan A.H & Teo C. *Proc on Neural Networks* Alaska 1998: 183
[23] Tsoi LC *et al*. *Journal Biomedical Informatics* 2009 **42**: 824 [PMID: 19318137].
[24] Prasad P *et al*. *IICAI* 2005 : 2092
[25] Porter M. Program 1980 **14**: 130
[26] Carmel D *et al*. ACM SIGIR New York 2009 : 139
[27] Branson S & Greenberg A. *Tech. Rep. CS276A Final Project Stanford University*. 2002.
[28] Hainaut P & Wiman K. 25 Years of p53 Research, Springer-Verlag 2007: 439

Edited by P Kangueane
Citation: David & Samuel, Bioinformation 8(1): 020-025(2012)

BIOINFORMATION open access

Supplementary material:

Table 1: Nearer term "apoptosis" of "p53" and nearer terms of "apoptosis" generated by searching the terms that come nearer to the domain "p53" and then search the terms nearer to "apoptosis".

Terms that come nearer to the domain	Terms extracted under "apoptosis"
P53 apoptosis	apoptosis ARF
apoptosis p53	apoptosis DNA damage
	activation induction apoptosis
	bax apoptosis
	bcl-2 apoptosis
	cancer apoptosis
	cellular apoptosis
	cycle arrest apoptosis
	Fibroblast apoptosis
	growth arrest apoptosis
	Ki67 apoptosis
	mutant apoptosis
	inhibit apoptosis
	intracellular apoptosis
	lymphocyte apoptosis
	MDM2 apoptosis
	mitochondria-induced apoptosis
	oncogene-induced apoptosis
	phosphorylation apoptosis
	protein apoptosis
	PUMA apoptosis
	renal cell apoptosis

Table 2: Comparison between the accuracy of the terms obtained for Clustering using various methods

Domain	Number of PubMed abstracts	Method	Number of terms obtained	Number of correctly predicted terms	Precision	Recall	F-score
P53	200	i)STC	16	6	0.37	0.06	0.10
		ii)LINGO	43	12	0.27	0.12	0.16
		iii)FNeTD approach	135	43	0.31	0.43	0.36