Mining Inter-Relationships in Online Scientific Articles and its Visualization: Natural Language Processing for Systems Biology Modeling

https://doi.org/10.3991/ijoe.v15i02.9432

Nidheesh Melethadathil, Bipin Nair, Shyam Diwakar
Amrita School of Biotechnology, Amrita Vishwa Vidyapeetham, Kerala, India

Jaap Heringa(✉)
Vrije Universiteit, Amsterdam, The Netherlands
heringa@few.vu.nl

Abstract—With the rapid growth in the numbers of scientific publications in domains such as neuroscience and medicine, visually interlinking documents in online databases such as PubMed with the purpose of indicating the context of a query results can improve the multi-disciplinary relevance of the search results. Translational medicine and systems biology rely on studies relating basic sciences to applications, often going through multiple disciplinary domains. This paper focuses on the design and development of a new scientific document visualization platform, which allows inferring translational aspects in biosciences within published articles using machine learning and natural language processing (NLP) methods. From online databases, this software platform effectively extracted relationship connections between multiple sub-domains within neuroscience derived from abstracts related to user query. In our current implementation, the document visualization platform employs two clustering algorithms namely Suffix Tree Clustering (STC) and LINGO. Clustering quality was improved by mapping top-ranked cluster labels derived from an UMLS-Metathesaurus using a scoring function. To avoid non-clustered documents, an iterative scheme, called auto-clustering was developed and this allowed mapping previously uncategorized documents during the initial grouping process to relevant clusters. The efficacy of this document clustering and visualization platform was evaluated by expert-based validation of clustering results obtained with unique search terms. Compared to normal clustering, auto-clustering demonstrated better efficacy by generating larger numbers of unique and relevant cluster labels. Using this implementation, a Parkinson's disease systems theory model was developed and studies based on user queries related to neuroscience and oncology have been showcased as applications.

Keywords—Online scientific databases, natural language processing, systems biology, automated clustering, visualization.

1 Introduction

In recent years, with advancements in scientific research, the numbers of articles in online medical and biological databases have increased substantially requiring automated tools that relate results across disciplines [1][2][3]. Biomedical Natural Language Processing (BioNLP) could help reconstruct and extract knowledge through the visualization of queries [4][5]. Natural Language Processing (NLP) is a sub-domain of artificial intelligence (AI) which employs computational methods to reconstruct spoken or written human language [6]. BioNLP involves information extraction and processing text and literature related to biological sciences [7][8][9] along with organized reconstruction and representation of document information [10], [11].

The study of complex biological systems like the brain and nervous system involves translational sciences with multidisciplinary approaches [12]. Translational sciences, an interdisciplinary branch of the biomedicine combines different disciplines, resources, expertise, and techniques within biomedical technology to promote enhancements in prevention, diagnosis, and therapies [13]. In research domains such as neuroscience, seamless integration of data from different sub-disciplines allow exploring neural and behavioral function and dysfunction. Today's translational sciences rely on complex biological organization and processes that relate genes and molecular constituents to cellular, circuit and behavioral effects. This is further utilized in systems biology as it connects the biological information transfer with different subdomains [14]. Studies on proteomics and genomics related to neurological and oncological conditions have led to literature exploration tools. Clinical researchers often seek relationships between molecular mechanisms involving genes, receptors, cells, tissue functions, organs and behavior in order to connect pathologies to their underlying mechanisms. Such relationship patterns often exist within multiple scientific documents and text-related data mining and analytics allow formulation of useful connections within databases. For example, a study on fibroblast growth factors demonstrated that ataxic or epileptic patterns involving a set of clinical symptoms could be attributed to dysfunctions, including that of sodium channels and intrinsic excitability of certain neurons [15].

Tools like NeuroExtract [16], Blumia [17], Textpresso for Neuroscience [18] and PubMedPortable [19] have been developed to extract information by mining biological data from online databases. Other information retrieval systems like BioIE [20], BioRAT[21], iHOP [22] and Carrot2 [23] allow users to extract information from published biomedical literature. In some NLP tools, retrieved data was organized into different groups by using MeSH class [24], by employing machine learning (ML) algorithms [25][26] or based on gene ontology [27]. Neuroscience Information Framework Literature (NIF-Literature) is another tool [28][29][30][31], which allows refined search of neuroscience literature. Common issues related to BioNLP also present in many software tools include absence of visualization front-end, graphical representations of feature subset relationships among clustered articles, large number of unclustered documents and absence of domain-relevant cluster labels.

To overcome these issues, this paper employs querying, clustering and visualization methods packaged as a BioNLP platform intended to help model relations in systems biology and for mapping multi-disciplinary literature visualization for translational medicine. The software attempts to resolve the issues present in existing platforms, namely, non-clustered documents through a scored auto-clustering approach, reverse mapping scientifically relevant MeSH terms as cluster labels and a weighted graph-based visualization. The paper also highlights the validation of the tool comparing neuroscience and molecular biology queries.

2 Machine Learning and Natural Language Processing

Machine learning has been used in addressing BioNLP aspects including text classification [32], tagging structured models [33], parsing and extraction [34], and unsupervised learning with structure induction and document clustering [35]. In a previous study [36], performance of several machine learning algorithms on BioNLP datasets was evaluated. Eight classifiers-based learning models on 2000-point dataset with MeSH terms as features showed ~78.2% training accuracy, while the root mean square error was <0.26. Among the clustering algorithms tested, k-means and expectation maximization demonstrated the highest accuracy. A study [37] on several clustering algorithms reported that LINGO [38] and STC [39] aggregated 17% more documents than k-means. As a choice for this implementation, based on clustering accuracy (see Table 1), Suffix Tree Clustering (STC) and LINGO suited better for document clustering and were employed. Grouping error for LINGO and STC was significantly less compared to k-means.

3 Methodology

This BioNLP implementation (referred to also as ABioNLP) allows retrieval of research articles from PubMed based on a user query and employs a weighted repetitive clustering approach to resolve a commonly observed NLP issues, including the large number of articles without relevant cluster labels. This web-based literature retrieval platform implements a pipeline of four modules (Fig. 1). First module involves documents retrieval from online literature database and incremental storage of pre-processed data in a local database. Second module performs document clustering based on algorithm selection (LINGO or STC) and mode of clustering (normal or auto clustering). The third module in the pipeline involved validation of the generated cluster labels by mapping with an online service, metathesaurus, and storage of clustering results in a graph database. In the fourth module, visualization of results was facilitated using interactive graphical interface.

Table 1. Comparative analysis of performance of various document-clustering algorithms.

Clustering Algorithm	Search word	Semantics connectivity between search words	Document Density	Outlier Percentage	Grouping Error on a scale of 0 to 10
LINGO	Cancer Oncovirus fibroblast	High	High	21	0.37142
	Cerebellum granule neuron ataxia	Medium	Low	11	0.36111
	Cerebellum pain	Low	Medium	22	0.32352
STC	Cancer Oncovirus fibroblast	High	High	0	0.33333
	Cerebellum granule neuron ataxia	Medium	Low	0	0.33333
	Cerebellum pain	Low	Medium	14	0.26666
K-Means	Cancer Oncovirus fibroblast	High	High	0	0.48
	Cerebellum granule neuron ataxia	Medium	Low	0	0.52
	Cerebellum pain	Low	Medium	1	0.52173

3.1 Document Retrieval and Pre-processing

The software implementation included a simple password-based user authentication procedure [40] facilitating user-specific search history could be recorded. ESearch from Entrez E-Utilities [41] was used to retrieve the PubMed ids corresponding to search word in an XML format. Through the graphical user interface, users could modify parameters like number of documents, number of clusters, selection of algorithms and mode of clustering, while querying PubMed. A database was implemented in MySQL for storing retrieved queries that comprised document information attributes including PubMed_id, author_list, journal_name. MeSH_words, abstract and other details. For every query retrieved using efetch (from E-utilities), the unique identifier for scientific articles was recorded into the local database along with other attributes, but without repetitions to avoid data redundancy.

In the pre-processing pipeline, abstracts were subjected to stop-word removal, stemming and lemmatization [42]. An informative abstract acts as a surrogate for the research article as it describes methods, results and evidence from the study [43]. As a trade-off for performance, instead of full text articles, abstracts were used for clustering. Thirty key terms extracted from an abstract using TF-IDF method [44] and PubMed-retrieved MeSH terms represented a retrieved scientific article. The locally stored database involved extracted features from several scientific articles attributed to queries.

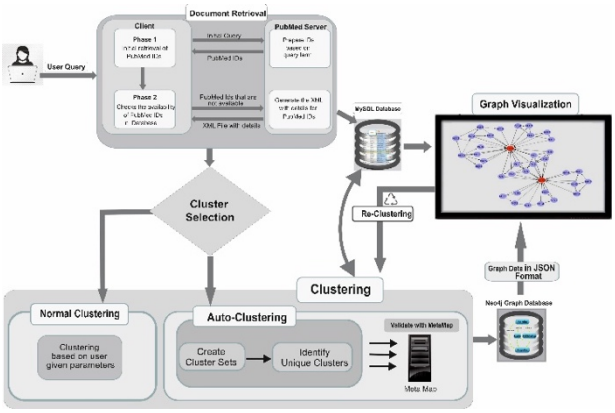

Fig. 1. Architecture of ABioNLP platform. Implementation pipeline included data retrieval, document clustering, cluster validation and visualization. The software tool processes queries, maps cluster labels and allows visualization in text and graph formats

3.2 Clustering

The choice of clustering algorithms was based on estimations of accuracy of the learning models [36]. For LINGO and STC, normal clustering (single-run) and iterative automated clustering (referred as auto-clustering) modalities were implemented. When normal clustering was performed, STC generated many disambiguous cluster labels and LINGO generated several non-clustered documents (assigned to a nondescript "other" cluster). To address these issues, auto-clustering, an iterative scheme that employs a differential evolver [45] to optimize clustering parameters, was introduced. The clustering algorithms, LINGO and STC were implemented based on an open source API [46]. Cluster entropy was reduced by iteratively changing the cluster assignment of those articles within the non-clustered group. Among algorithmic parameters, LINGO's 'term weighing' parameter (Fig. 2. and see Table 2) and 'base-cluster merging threshold' parameter (Fig. 3. and see Table 3)in STC allowed maximally reallocating documents from non-clustered to valid clusters also fine-tuned the clustering accuracy. This iterative process of auto-clustering was terminated when resulting clusters remained unchanged in consecutive iterations. An additional process called "re-clustering" or repetitive clustering was incorporated to allow the generation of sub-clusters. The user could select re-clustering by right clicking on the displayed cluster label. Clustering efficiency was calculated by dividing number of unique clusters formed with time taken and the number of research articles that were not clustered, for both auto and normal clustering methods.

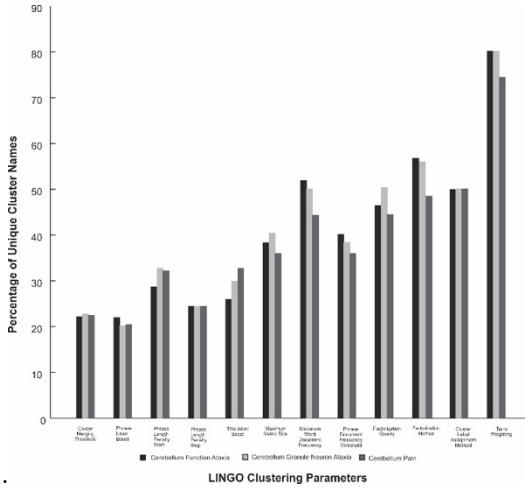

Fig. 2. Effect of Parameters on LINGO. Term weighting and factorization added to maximal unique cluster names.

Table 2. LINGO and the percentage of unique cluster labels

Parameters	Percentage of Unique Cluster Labels
Term weighing parameter	80
Factorization method	59
Maximum word document frequency	54
Factorization quality and cluster label assignment method	50
Maximum metric size	41
Phrase document frequency	40

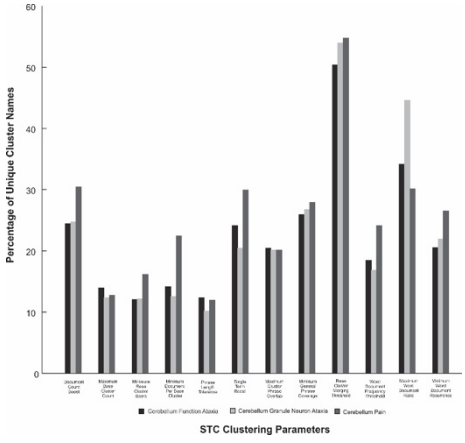

Fig. 3. Effect of Parameters on STC. Fine tuning base cluster merging threshold and maximum word document ratio helped generate maximal unique cluster labels.

Table 3. STC and the percentage of unique cluster labels

Parameters	Percentage of Unique Cluster Labels
Base cluster merging threshold	55
Maximum word document frequency	45
Document count boost and minimum document per base cluster	31
Maximum general phrase coverage	29
Minimum word document recurrence	28

3.3 Cluster Validation and Scoring

A common problem with document clustering was the generation of imprecise and/or scientifically-irrelevant cluster labels [47]. In order to overcome this issue, ranking of cluster labels based on its scientific relevance was performed by mapping them using MetaMap [48]. A cluster score (Cs), computed from generated clusters, involving the number of cluster labels and document count within the cluster indicated relevant cluster labels before mapping with MeSH terms. Post mapping, another score (MM) for all cluster labels generated by MetaMap was included in the computation of a weighted final score (1), that defined top ranked cluster labels representing the query.

$$Final_{Score} = 0.5 * N_C + 0.3 * S_C + 0.1 * \left(M_M * \frac{1}{100}\right) + 0.2 * C_S \quad (1)$$

Where, NC was the number of cluster sets where the given cluster name was present in, SC was the size of cluster, MM was the MetaMap score and CS was the clustering score. Thirty clusters identified by descending order of scores were selected as the final dataset for graph visualization. In order to further distribute articles from large clusters, repeated clustering (re-clustering) was incorporated, which allowed users to repeatedly cluster already formed clusters (Fig. 4.).

3.4 Graph Storage and Visualization

The details of top-ranked thirty clusters including their labels, contents, interconnections and final score were stored in a MySQL table. Since graphically displaying relationships among clustered research articles based on their content similarities using Cypher Query Language [49] could enhance system's theory exploration of biological literature, interactive radial node visualization of cluster results was implemented. A copy of the processed data was stored in Neo4j [50]. Converted as JSON objects from this database, circular graphs with radii proportional to the size of clusters [51] were used for visualization. Documents were represented as nodes and relationships as edges [52] as in a property graph model [53]. The radial node-link graph-based data visualization was implemented using Javascript InfoVis Toolkit (JIT) [54]. Documents within clusters along with their abstract and other details were also made available on the right section of the screen (Fig. 5).

3.5 External Evaluation and Clustering Accuracy

Since tuple-analysis does not provide a difference metric for overall quality of the system [55], manual evaluation of clustering accuracy was performed to validate cluster quality [56]. For all case studies, clustering quality [57] was evaluated using extrinsic measures involving domain experts. Query reformulation related case studies involved normal or auto-clustering by varying number of articles retrieved across a range 50-100-150-200 to highlight the commonness with increase in number of retrieved articles.

The ABioNLP platform was implemented using HTML, Javascript and calling Java APIs with MySQL, Neo4j databases and the source code is available at https://github.com/compneuro/ABioNLP

Fig. 4. Iterative Clustering to handle non-clustered documents: Iterative clustering enhances by re-grouping the articles, which were not previously listed under a cluster; and can be executed by right clicking on the circle area.

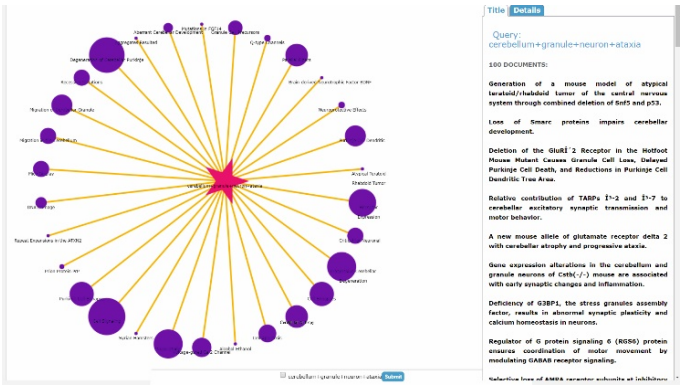

Fig. 5. Auto-clustering in ABioNLP. For a query "cerebellum granule neuron ataxia" represented by a star, auto-clustering generated document clusters represented by circles with diameter representing cluster size.

4 Results

The efficacy of this BioNLP tool was extrinsically evaluated by comparing the clustering results obtained for unique search terms. The aim of using human experts was to satisfy better understanding of document relevance within cluster and the limitations. Towards evaluation and testing, multiple runs of document retrieval were performed by varying the search terms and number of documents to be retrieved. We tested (i) the platform's capability for multilevel data explorations, (ii) the accuracy of query reformulations in document searches and (iii) whether effective visualization of the clustering results for applications in systems theory and biomedicine. In addition to this, querying was conducted to evaluate algorithmic parameters, to compare the performance of normal and auto-clustering algorithms with STC and LINGO and the accuracy of clusters using subject expert's manual verification. Performance indicators were processing time and clustering accuracy. Both normal clustering and auto-clustering with LINGO and STC algorithm were evaluated. The visualization included a node-link graphical representation with radial graphs representing size of clusters with documents grouped together based on the similarity to other research articles. Users could click on labels to see the grouping of the data. Processing time included time taken for query retrieval and document clustering performed locally. As designed, auto-clustering required more runtime than normal clustering due to repetitive processing (see Table 4).

Table 4. A comparison of computational cost (time) for search and post-query processing.

	PubMed	**Article Retrieval Time for Our Implementation**			
No. of Articles to be Retrieved	***Article Retrieval Time (Seconds)***	***Normal Clustering with STC***	***Normal Clustering with LINGO***	***Auto Clustering with STC***	***Auto Clustering with LINGO***
100	3.92 S	5.088 S	5.15 S	31.43 S	34.09 S
200	5.347 S	7.396 S	8.178 S	30.04 S	54.74 S
500	Not available	8.002 S	8.38 S	31.07 S	54.59 S

4.1 Manual verification of document clustering

Since internal evaluation was related to algorithms [58], we evaluated the usefulness of the software tool through its cluster quality evaluated by experts and visualization relating documents to MeSH terms from the metathesaurus. Although labor intensive, generated clusters were evaluated with subject experts manually by counting the relevance of articles retrieved with respect to their assigned cluster labels. For the neuroscience domain, top 100 research articles were extracted from a query (terms: cerebellum, function, ataxia) and normal and auto-clustering using STC and LINGO algorithms were performed. In normal clustering, seven clusters were generated using both algorithms. The cluster labels were different for both algorithms and with LINGO, there was a set of non-clustered documents which were not mapped into any existing cluster labels.

The average clustering accuracy for STC was 96.76% and LINGO was 94.52% see Table 5 & Table 6). The fraction of non-clustered documents produced by LINGO

was 62%. Using auto-clustering, 30 exclusive clusters were generated by both algorithms. The clustering accuracy for both the algorithms (data not shown) was greater than 95%, while there was no un-clustered documents for LINGO.

Table 5. Expert verified cluster results for STC with normal clustering

Cluster Label	Correctly Clustered	Incorrectly Cluster
Cerebellar, Ataxia, Cell	100	0
Cerebellar Granule Neurons	24	1
Spinocerebellar Ataxia	14	0
Parallel Fibers	18	3
System, Central Nervous System	26	0
Voltage gated channels	22	1
Cerebellar Purkinje Cells	14	0

Table 6. Expert verified cluster results for LINGO with normal clustering

Cluster Label	Correctly Clustered	Incorrectly Clustered
Motor coordination	14	2
Cell Migration	11	1
Postnatal CNS	6	1
DNA Damaged Response	5	0
Ataxic Gait	3	0
Syrian Hamsters	2	0
Non-clustered	60	2

Article retrieval and document clustering were also performed for an oncology related search word (query terms included: MMP, metastasis, oncovirus). 100 research articles were extracted and generated 7 clusters for normal clustering (Fig. 6) and 30 clusters for auto-clustering (data not shown) with LINGO and STC. In auto-clustering approach, it was found that the tool indicated clustering results with 95% accuracy and there were no non-clustered articles with the STC algorithm. STC with normal clustering facilitated faster retrieval and LINGO with auto-clustering generated precise clusters.

4.2 Computing Efficacy across Normal and Automated Clustering

Query retrievals were performed four times with normal clustering while varying number of clusters (10, 20, 30 and 40) with number of articles being retrieved set to 200. For an information retrieval system using document-clustering approach, efficiency of the retrieval was proportional to number of unique clusters formed with lesser processing time and minimal number of non-clustered articles. Mean and variance of cluster quality or efficiency were calculated (Fig. 7).

Similarly, with auto-clustering, cluster quality assessments with experts were performed by retrieving 200 research articles and the experiment was repeated four times. For each iteration, LINGO generated 243 unique clusters in 54.65 seconds and STC generated 73 unique clusters in 30.45 seconds (averaged for four iterations). Average efficiency was 4.45 unique clusters/second for LINGO and 2.36 unique

clusters/second for STC (see Table 7). From this, experiment it can be deduced that auto-clustering outperforms normal clustering in retrieving the most relevant research articles. Even though LINGO performs better with Auto-clustering, since the STC algorithm produces better accuracy with Normal clustering method, both STC and LINGO were retained in the platform.

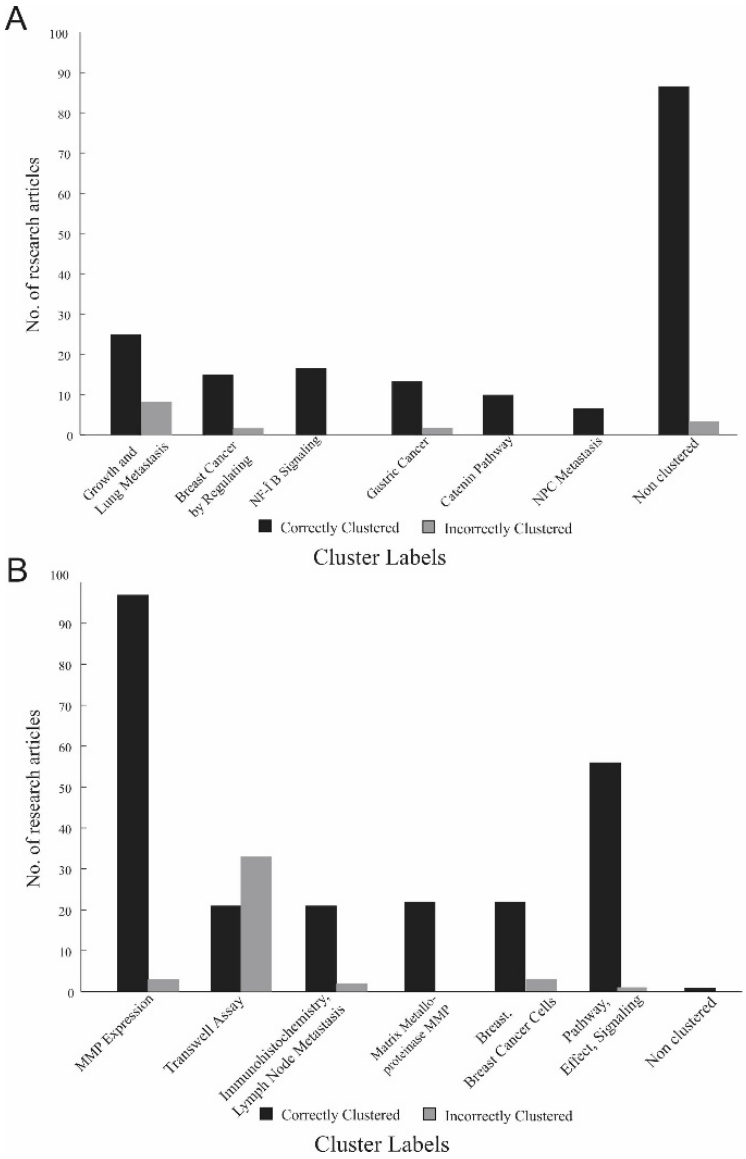

Fig. 6. Accuracy from subject expert verified auto-clustering for the cancer-related query. A) For LINGO, there are more than 85 articles out of 100 which were grouped into "non-clustered" B) For STC algorithm also has generated significant cluster labels.

Table 7. Efficiency comparison of LINGO and STC. LINGO with auto-clustering generated results that were more accurate.

Clustering Algorithms	Avg. Time	Avg. no of Unique concepts	Estimated Efficiency Measure
LINGO	54.65	243	4.44
STC	30.45	72	2.36

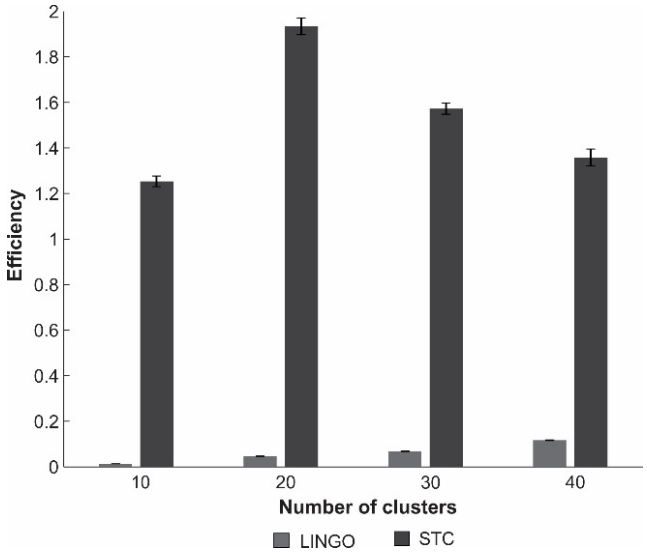

Fig. 7. Comparison of normal clustering between LINGO and STC algorithms. LINGO had a large number of unclustered documents compared to STC. Clustering efficiency was considerably high for STC compared to LINGO with increasing number of articles.

4.3 Search Efficacy: Document Clusters Mapped Inter-Relationships among Research Articles

To quantify usability in establishing inter-connections between different sub-domains of neuroscience through research articles, multiple clustering runs were carried out to explore common characteristics among multiple queries.

Search retrieval and clustering for 50 articles with search terms ("Gene Cerebellum Epilepsy" and "Cerebellum Granule Neuron Epilepsy") relating to two different sub-domains of Neuroscience, namely physiology and genetics, were performed for evaluating the BioNLP platform's functionality.

For the search term related to neuro-genetics, 23 unique cluster labels were formed from 315 available articles. Out of which, 9 cluster labels were exclusively related to genetics whereas 8 cluster labels were related to physiology and 6 cluster labels were related to both neuro-genetics and neurophysiology. Similarly, for a search word related to neurophysiology, 25 unique cluster labels were formed from 68 research articles, of which 20 cluster labels were related to physiology and 5 clusters represented both (Fig. 8).

Two unrelated search terms "ataxia Purkinje neuron" and "multiple sclerosis neural circuits" were used to perceive overlaps between pathology, physiology, molecular biology and genetics sub-domains. Seven cluster labels for the ataxia-Purkinje neuron query and four cluster labels for the multiple sclerosis-neural circuits query among the 30 cluster labels were incorrectly grouped as per human expert evaluation. With smaller numbers of documents, being extracted related to a query, the platform was faster in identifying significant connections between multiple sub-domains of biomedical sciences.

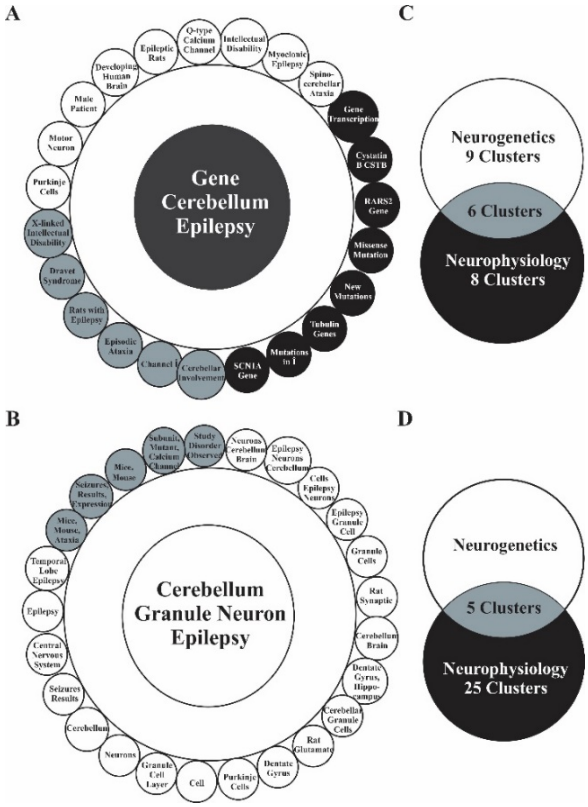

Fig. 8. Extracting inter-domain relationships from queries exploration. For a "gene cerebellum epilepsy" (A) 15 clusters were attributed to neurogenetics and 14 clusters were attributed to neurophysiology with 6 common clusters (C). For a "cerebellum granule neuron epilepsy" query (B), 5 clusters belonged commonly (D) to neurogenetics and neurophysiology while 30 clusters were attributed to neurophysiology.

4.4 Document Clustering as Query Reformulation

To evaluate document clustering for query reformulation, varying numbers of articles were retrieved for neuroscience search terms namely "ataxia" and "cerebellum granule neuron ataxia". While varying number of documents to be retrieved (50, 100,

150 and 200), the relevance of the cluster labels and commonness of research articles across different groups were evaluated. 60% of the cluster labels remained unmodified during both the queries, whereas the other cluster labels evolved according to relevance to specific search terms. (Fig. 9A, 9B and 9C). The average numbers of common articles within the clusters was 43 when total number of articles retrieved was increased from 50 to 100.

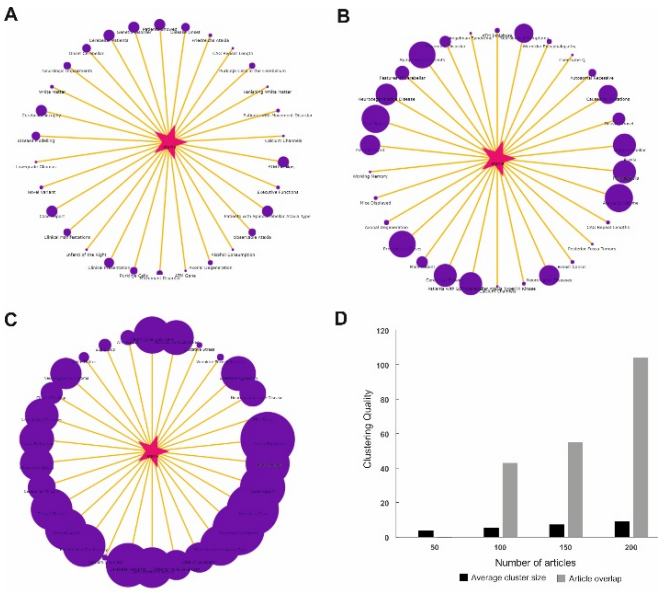

Fig. 9. Query reformulation and visualization. Varying number of retrieved articles generated distinctive cluster sizes and groups. A) With the number of documents to be retrieved set to 50, clusters were sparser although all 30 top cluster labels were ranked. B) With number of articles set to 100, clusters relating critical MeSH terms became prominent. Clusters when the article count was increased was 100. C) With 200 documents, the relevance of cluster labels assigned to a cluster became evidently significant. D) Improvement of cluster results (average cluster size and article overlap) with number of retrieved articles increased.

For an increase in number of articles to be retrieved from 100 to 150, there were 55 common articles. 104 common articles were observed when the number of retrieved documents was increased from 150 to 200 (Fig. 9D). Higher number of documents may improve the relevance of cluster labels assigned to a cluster.

4.5 Document Clustering for Systems Biology Modelling

The application of this software platform in automated curation of molecular signaling pathways was evaluated through a systems biology study [59], aiming the characterizing of neurodegeneration and understanding pathophysiology of Parkinson's Disease (PD). Using multiple queries for reconstruction of dopamine

pathway, α-synuclein aggregation, effect of tau phosphorylation on formation of neurofibrillary tangles and cell death and, role of production and activation of reactive oxygen species that lead to apoptosis, the molecular signaling mechanisms were modelled using biochemical systems theory. For a search term "alpha synuclein in Parkinson's diseases", PubMed search retrieved 15839 articles as a list without any categorization. The search terms when submitted to the BioNLP software provided 30 different clusters of which 9 clusters highlighted the most relevant 18 articles from which the data for this study was collected. For other search terms: "Parkinson's disease", "Dopamine pathways in Parkinson's disease", "role of alpha synuclein in Parkinson's diseases" generated 12 articles (from more than 18000 articles). Similarly, for finding the information related to "α-synuclein aggregation", "parkin" and "ROS", the number of articles recovered through query retrieval (after auto-clustering) was 14 (from >10000),7 (>15,000) and 8 (>11800) respectively. Through four searches, main references to Parkinson's disease and concentration of proteins of related genes establishing the PD pathway were retrieved. Top scored articles were directly a reference and a key citation in the published study [60].

5 Discussion and Conclusion

The design allows to identify the subject level interconnections between different sections of a scientific domain, providing users with a document browsing interface. The software platform employed different learning models for scientific document analysis, combining document clustering with mapping of MeSH terms and generated domain-relevant cluster labels and extended domain-level inter-relationships among various documents queried from PubMed. With the range of ML techniques now being applied to BioNLP, the human expert evaluation of cluster quality suggested the clustering performance was relatively good and a tractable NLP implementation was facilitated by using the mapping process of cluster labels. By a few queries and their visualizations, this reliably allowed building literature dependent systems model of Parkinson's disease [60] that reconstructed predictions dysfunction and kinetic changes of activity-derived pathways of dopamine, α-synuclein aggregation, tau, parkin and ROS in dopaminergic neurons. Such visualizations are crucial for exploiting the community-centered approach in systems biology. From an implementation stand point, increasing sizes of data and literature in PubMed and repeated clustering adding to computational costs, incremental updates in local repository optimized time during the document clustering process.

While comparing to existing tools, the BioNLP platform replaces some of the issues still seen in state-of-the-art tools being used by biomedical and neuroscience researchers. Uncategorized documents generation during a single run clustering as observed in Carrot2 [23] was overcome in this implementation by auto-clustering. ABioNLP platform uses a radial graph visualization for easy to retrieve results unlike in other text mining tools like NeuroExtract [16], Textpresso [18] and iHOP [22] which lacks a visualization component, which was crucial to manually curate experimental data for pathway reconstruction in systems modeling. The platform also

used the concept of using MeSH class as well as categorization of abstracts as in some tools like XplorMed [24] and GoPubMed [27]. Although search and query reformulation in aforementioned BioNLP platforms are reliable, they have issues related to clustering or categorization including listing a significantly large number of unrelated articles as a cluster, absence of meaningful or domain-relevant cluster labels or an appropriate visualization front-end for evaluating relationships among clustered articles and have been by addressed by this BioNLP platform.

Since computational cost was increased due to clustering and mapping, the runtime efficiency was increased by saving query results in a local database and retrieving non-listed documents based on the search results. Mapping cluster labels to MeSH terms helped to avoid irrelevant cluster labels and improved machine learning accuracy [36]. By using different machine learning algorithms, it was found that reverse mapped data showed higher accuracy compared to data extracted directly from retrieved documents. Although, computationally expensive higher number of documents improved the relevance of cluster labels assigned to a cluster. This may be attributed to clustering algorithm's learning model. In addition to a list, visualization as different circles with varying diameter that represented the size of the cluster, allowed users estimate cluster relevance related to the query. The node-link radial graph representation of the search results allowed quick visual inspection of interrelations between sub-domains and alternate clusters.

The online retrieval time was regulated by allowing the user to specify the number of articles to be retrieved. As expected, the number of common articles between these different sub-domains increased with increase in total number of articles retrieved. This may be due to the MeSH-mapping with MetaMap biasing unique domain-specific keywords. As number of retrieved documents increased, new and more search-relevant cluster labels appeared. When manual verification with help of subject experts from various domains was carried out to evaluate the clustering accuracy of both algorithms, STC performed better than LINGO during normal clustering. This may be attributed to the large non-clustered category of "other documents" generated by LINGO. However, if auto-clustering is performed on the LINGO clustering results, performance became clearly superior relative to clustering by STC.

As showcased in the case study on Parkinson's disease, one of the advantages was that it buffers users from metadata and ontologies while providing complex relations between research articles aggregated as clusters with medically relevant labels. Instead of many searches, employing four queries, the pathway studies for prediction of PD signaling related 30 key articles from a PubMed list of more than 10000 research papers. This attributes significantly in reducing literature survey-work time for researchers who study biology and diseased conditions in animals and humans. While focusing on reducing literature data deluge for researchers, the generality of the design allows this search utility to be incorporated in any of the existing Omics and BioNLP platforms and tools.

With management of large diversity of data, linking experimental data to models and integrating translational exploration and comparisons, this software platform facilitates easy extraction of concepts from research articles and facilitates BioNLP, systems theory and modelling for data sciences. Without focusing only on BioNLP

methods, this design sheds a new perspective into architectural modelling of an information retrieval system that can be empowered as a step towards using public data sets and rebuilding ontology-based networks while accessing available databases in the multidisciplinary field of systems biology. Additionally, this BioNLP platform saves reading time for researchers and clinicians since retrieved queries can cluster primary references limiting the many thousands to a fewer number as in the case of the systems biology case study. Although the current scope of the platform was restricted to medicine and biology, with any other normally inter-operable document archiving databases and data streams, this can be enhanced for several applications. With capability to scale for big data analytics and streaming data analysis, this can be effectively re-implemented to include scaling deep learning models.

6 Acknowledgement

This project derives direction and ideas from the Chancellor of Amrita Vishwa Vidyapeetham, Sri Mata Amritanandamayi Devi. Authors would like to thank Priya Chellaiah for the simulation runs, Asha Vijayan for help with editing and proof reading the paper, Rohit Joseph Sebastian, Unnikrishnan Suresh for their design contributions, Sanu Shaji for his help in evaluating the cluster results and Hemalatha Sasidharakurup for her test case evaluation. The work was partially supported by Young Faculty Research Fellowship from Ministry of Electronics and IT, Government of India and by Embracing The World.

7 References

[1] T. Hey and A. Trefethen, "The Data Deluge: An e-Science Perspective," *Grid Comput.*, no. January 2003, pp. 809–824, 2003. https://doi.org/10.1002/0470867167.ch36

[2] C. C. Lapish, N. Tirupattur, and S. Mukhopadhyay, "Text Mining for Neuroscience: A Co-morbidity Case Study," in *Knowledge-Based Systems in Biomedicine and Computational Life Science*, vol. 450, 2013, pp. 117–136.

[3] L. Marenco, R. Wang, G. M. Shepherd, and P. L. Miller, "The NIF DISCO Framework: Facilitating automated integration of neuroscience content on the web.," *Neuroinformatics*, vol. 8, no. 2, pp. 101–12, 2010. https://doi.org/10.1007/s12021-010-9068-8

[4] F. Carrara, A. Esuli, T. Fagni, F. Falchi, and A. Moreo Fernández, "Picture it in your mind: generating high level visual representations from textual descriptions," *Inf. Retr. J.*, vol. 21, no. 2–3, pp. 208–229, Jun. 2018. https://doi.org/10.1007/s10791-017-9318-6

[5] K. Tulipano, Y. Tao, W. Millar, P. Zanzonico, K. Kolbert, H. Xu, H. Yu, L. Chen, Y. A. Lussier, and C. Friedman, "Natural language processing and visualization in the molecular imaging domain," *J. Biomed. Inform.*, vol. 40, no. 3, pp. 270–281, Jun. 2007. https://doi.org/10.1016/j.jbi.2006.08.002

[6] K. Liu, W. R. Hogan, and R. S. Crowley, "Natural Language Processing methods and systems for biomedical ontology learning," *J. Biomed. Inform.*, vol. 44, pp. 163–179, 2011. https://doi.org/10.1016/j.jbi.2010.07.006

[7] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," *Commun. ACM*, vol. 49, no. 9, pp. 76–82, 2006. https://doi.org/10.1145/1151030.1151032

[8] I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar, "Text mining and ontologies in biomedicine: Making sense of raw text," *Brief. Bioinform.*, vol. 6, no. 3, pp. 239–251, 2005. https://doi.org/10.1093/bib/6.3.239

[9] A. Stavrianou, P. Andritsos, and N. Nicoloyannis, "Overview and semantic issues of text mining," *ACM SIGMOD Rec.*, vol. 36, no. 3, p. 23, Sep. 2007. https://doi.org/10.1145/1324185.1324190

[10] A. Naud and S. Usui, "Exploration of a collection of documents in neuroscience and extraction of topics by clustering.," *Neural Netw.*, vol. 21, no. 8, pp. 1205–11, Oct. 2008. https://doi.org/10.1016/j.neunet.2008.05.009

[11] S. Usui, A. Naud, N. Ueda, and T. Taniguchi, "3D-SE Viewer: A Text Mining Tool based on Bipartite Graph Visualization," in *2007 International Joint Conference on Neural Networks*, 2007, pp. 1103–1108. https://doi.org/10.1109/IJCNN.2007.4371112

[12] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems.," *Nat. Rev. Neurosci.*, vol. 10, no. 3, pp. 186–98, Mar. 2009. https://doi.org/10.1038/nrn2575

[13] R. J. Cohrs, T. Martin, P. Ghahramani, L. Bidaut, P. J. Higgins, and A. Shahzad, "Translational Medicine definition by the European Society for Translational Medicine," Elsevier, Mar. 2015.

[14] M. W. Kirschner, "The meaning of systems biology.," Elsevier, May 2005.

[15] M. Goldfarb, J. Schoorlemmer, A. Williams, S. Diwakar, Q. Wang, X. Huang, J. Giza, D. Tchetchik, K. Kelley, A. Vega, G. Matthews, P. Rossi, D. M. Ornitz, E. D'Angelo, S. Page, and E. D. Angelo, "Fibroblast Growth Factor Homologous Factors Control Neuronal Excitability through Modulation of Voltage-Gated Sodium Channels.," *Neuron*, vol. 55, no. 3, pp. 449–463, Aug. 2007. https://doi.org/10.1016/j.neuron.2007.07.006

[16] C. J. Crasto, P. Masiar, and P. L. Miller, "NeuroExtract: facilitating neuroscience-oriented retrieval from broadly-focused bioscience databases using text-based query mediation.," *J. Am. Med. Inform. Assoc.*, vol. 14, no. 3, pp. 355–60, 2007. https://doi.org/10.1197/jamia.M2321

[17] R. Richardet, J. C. Chappelier, and M. Telefont, "Bluima: a UIMA-based NLP Toolkit for Neuroscience," in *UIMA@ GSCL*, 2013, pp. 34–41.

[18] H.-M. Muller, A. Rangarajan, T. K. Teal, and P. W. Sternberg, "Textpresso for Neuroscience: Searching the Full Text of Thousands of Neuroscience Research Papers," *Neuroinformatics*, vol. 6, no. 3, pp. 195–204, 2008. https://doi.org/10.1007/s12021-008-9031-0

[19] K. Döring, B. A. Grüning, K. K. Telukunta, P. Thomas, and S. Günther, "PubMedPortable: A Framework for Supporting the Development of Text Mining Applications," *PLoS One*, vol. 11, no. 10, p. e0163794, Oct. 2016. https://doi.org/10.1371/journal.pone.0163794

[20] A. Divoli and T. K. Attwood, "BioIE: extracting informative sentences from the biomedical literature.," *Bioinformatics*, vol. 21, no. 9, pp. 2138–9, 2005. https://doi.org/10.1093/bioinformatics/bti296

[21] D. P. A. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones, "BioRAT: Extracting biological information from full-length papers," *Bioinformatics*, vol. 20, no. 17, pp. 3206–3213, 2004. https://doi.org/10.1093/bioinformatics/bth386

[22] R. Hoffmann and A. Valencia, "Implementing the iHOP concept for navigation of biomedical literature," *Bioinformatics*, vol. 21, no. SUPPL. 2, pp. ii252-8, 2005.

[23] S. Osiński and D. Weiss, "Carrot 2 : Design of a Flexible and Efficient Web Information Retrieval Framework," in *Advances in Web Intelligence*, 2005, pp. 439–444.

[24] C. Perez-Iratxeta, P. Bork, and M. a Andrade, "XplorMed: a tool for exploring MEDLING abstracts," *TRENDS Biochem. Sci.*, vol. 26, no. 9, pp. 573–575, 2001. https://doi.org/10.1016/S0968-0004(01)01926-0

[25] N. R. Smalheiser, W. Zhou, and V. I. Torvik, "Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results.," *J. Biomed. Discov. Collab.*, vol. 3, p. 2, 2008. https://doi.org/10.1186/1747-5333-3-2

[26] Y. Yamamoto and T. Takagi, "Biomedical knowledge navigation by literature clustering," *J. Biomed. Inform.*, vol. 40, pp. 114–130, 2007. https://doi.org/10.1016/j.jbi.2006.07.004

[27] A. Doms and M. Schroeder, "GoPubMed: exploring PubMed with the Gene Ontology.," *Nucleic Acids Res.*, vol. 33, no. Web Server issue, pp. W783-6, Jul. 2005.

[28] J. Cachat, A. Bandrowski, J. S. Grethe, A. Gupta, V. Astakhov, F. Imam, S. D. Larson, and M. E. Martone, "A Survey of the Neuroscience Resource Landscape," in *International review of neurobiology*, vol. 103, 2012, pp. 39–68. https://doi.org/10.1016/B978-0-12-388408-4.00003-4

[29] A. Eshghishargh, K. Gray, S. K. Milton, and S. C. Kolbe, "A semantic system for answering questions in neuroinformatics," in *Proceedings of the Australasian Computer Science Week Multiconference on - ACSW '18*, 2018, pp. 1–5. https://doi.org/10.1145/3167918.3167960

[30] D. Gardner, H. Akil, G. a Ascoli, D. M. Bowden, W. Bug, D. E. Donohue, D. H. Goldberg, B. Grafstein, J. S. Grethe, A. Gupta, M. Halavi, D. N. Kennedy, L. Marenco, M. E. Martone, P. L. Miller, H.-M. Müller, A. Robert, G. M. Shepherd, P. W. Sternberg, D. C. Van Essen, and R. W. Williams, "The neuroscience information framework: a data and knowledge environment for neuroscience.," *Neuroinformatics*, vol. 6, no. 3, pp. 149–60, 2008. https://doi.org/10.1007/s12021-008-9024-z

[31] J. D. Tenenbaum and C. Blach, "Best practices and lessons learned from reuse of 4 patient-derived metabolomics datasets in Alzheimer's disease," *Pac Symp Biocomput*, vol. 23, pp. 280–291, 2018. https://doi.org/10.1142/9789813235533_0026

[32] T. H. Nguyen and K. Shirai, "Text Classification of Technical Papers Based on Text Segmentation," Springer, Berlin, Heidelberg, 2013, pp. 278–284. https://doi.org/10.1007/978-3-642-38824-8_25

[33] A. Subramanya, S. Petrov, and F. Pereira, "Efficient Graph-Based Semi-Supervised Learning of Structured Tagging Models," pp. 167–176, 2010.

[34] A. B. Clegg and A. J. Shepherd, "Benchmarking natural-language parsers for biological applications using dependency graphs," *BMC Bioinformatics*, vol. 8, no. 1, p. 24, Jan. 2007. https://doi.org/10.1186/1471-2105-8-24

[35] M. Sokolova and S. Szpakowicz, "Machine Learning in Natural Language Processing," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, IGI Global, 2009, pp. 302–324.

[36] N. Melethadathil, P. Chellaiah, B. Nair, and S. Diwakar, "Classification and clustering for neuroinformatics: Assessing the efficacy on reverse-mapped NeuroNLP data using standard ML techniques," in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, 2015, pp. 1065–1070. https://doi.org/10.1109/ICACCI.2015.7275751

[37] R. Mahalakshmi and V. L. Praba, "A Relative Study on Search Results Clustering Algorithms - K-means , Suffix Tree and LINGO," *Int. J. Eng. Adv. Technol.*, no. 6, pp. 31–35, 2013.

[38] S. Osiński, J. Stefanowski, and D. Weiss, "Lingo : Search Results Clustering Algorithm Based on Singular Value Decomposition," *Adv. Soft Comput. Intell. Inf. Process. Web*

Mining, Proc. Int. IIS IIPWM '04 Conf., pp. 359–368, 2004. https://doi.org/10.1007/978-3-540-39985-8_37

[39] O. Zamir and O. Etzioni, "Web document clustering: a feasibility demonstration," in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98, 1998, pp. 46–54. https://doi.org/10.1145/290941.290956

[40] A. Conklin, G. Dietrich, and D. Walz, "Password-based authentication: a system perspective," in *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, 2004, p. 10 pp. https://doi.org/10.1109/HICSS.2004.1265412

[41] E. Sayers, "The E-utilities In-Depth: Parameters, Syntax and More," Nov. 2017.

[42] A. G. Jivani, "A Comparative Study of Stemming Algorithms," *Int. J. Comp. Tech. Appl.*, vol. 2, no. 6, pp. 1930–1938, 2011.

[43] S. S. Fuller, D. Revere, P. F. Bugni, and G. M. Martin, "A knowledgebase system to enhance scientific discovery: Telemakus," *Biomed. Digit. Libr.*, vol. 1, no. 1, p. 2, Dec. 2004. https://doi.org/10.1186/1742-5581-1-2

[44] P. K. Shah, C. Perez-Iratxeta, P. Bork, and M. A. Andrade, "Information extraction from full text scientific articles: where are the keywords?," *BMC Bioinformatics*, vol. 4, p. 20, 2003. https://doi.org/10.1186/1471-2105-4-20

[45] S. Das, A. Abraham, and A. Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm," *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 38, no. 1, pp. 218–237, Jan. 2008. https://doi.org/10.1109/TSMCA.2007.909595

[46] A. Gonzales-Aguilar and M. Ramírez-Posada, "Carrot2: Búsqueda y visualización de la información," *El Prof. la Inf.*, vol. 21, no. 1, pp. 105–112, 2012. https://doi.org/10.3145/epi.2012.ene.14

[47] N. Shah and S. Mahajan, "Semantic based Document Clustering: A Detailed Review," *Int. J. Comput. Appl.*, vol. 52, no. 5, pp. 0975-8887, 2012.

[48] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: historical perspective and recent advances," *J. Am. Med. Informatics Assoc.*, vol. 17, no. 3, pp. 229–236, 2010. https://doi.org/10.1136/jamia.2009.002733

[49] S. Batra and C. Tyagi, "Comparative Analysis of Relational And Graph Databases," no. 2, pp. 509–512, 2012.

[50] M. Buerli and C. Obispo, "The Current State of Graph Databases," *Dep. Comput. Sci. Cal Poly San Luis Obispo*, 2012.

[51] E. Kandogan, "Visualizing multi-dimensional clusters, trends, and outliers using star coordinates," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*, 2001, pp. 107–116. https://doi.org/10.1145/502512.502530

[52] C. T. Have and L. J. Jensen, "Are graph databases ready for bioinformatics?," *Bioinformatics*, vol. 29, no. 24, pp. 3107–8, Dec. 2013. https://doi.org/10.1093/bioinformatics/btt549

[53] W. Schnotz, "Towards an Integrated View of Learning From Text and Visual Displays," vol. 14, no. 1, pp. 101–120, 2002.

[54] J. D. Fekete, "The InfoVis Toolkit," in Proceedings - IEEE Symposium on Information Visualization, INFO VIS, 2004, pp. 167–174. https://doi.org/10.1109/INFVIS.2004.64

[55] P. Resnik and J. Lin, "11. Evaluation of NLP Systems," in *The Handbook of Computational Linguistics and Natural Language Processing*, 2010, pp. 271–295.

[56] A. V Leouski and W. B. Croft, "An Evaluation of Techniques for Clustering Search Results," 1996.

[57] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," in *KDD-2000*, 2000.
[58] S. Xu, X. Qiao, L. Zhu, Y. Zhang, C. Xue, and L. Li, "Reviews on Determining the Number of Clusters," *Appl. Math. Inf. Sci*, vol. 10, no. 4, pp. 1493–1512, 2016. https://doi.org/10.18576/amis/100428
[59] H. Sasidharakurup, N. Melethadathil, B. Nair, and S. Diwakar, "A Systems Model of Parkinson's Disease Using Biochemical Systems Theory," *Omi. A J. Integr. Biol.*, vol. 21, no. 8, pp. 454–464, 2017.
[60] H. Sasidharakurup, N. Melethadathil, B. Nair, and S. Diwakar, "A Systems Model of Parkinson's Disease Using Biochemical Systems Theory," *Omi. A J. Integr. Biol.*, vol. 21, no. 8, 2017.

8 Authors

Nidheesh Melethadathil is currently working as Assistant Professor at Amrita Vishwa Vidyapeetham, India. (nidheesh@am.amrita.edu)

Jaap Heringa is the Professor at the Centre for Integrative Bioinformatics at VU (IBIVU), Vrije Universiteitt, Amsterdam, The Netherlands. (heringa@few.vu.nl).

Bipin Nair is the Professor and Dean of Faculty of Sciences, Amrita Vishwa Vidyapeetham, India. (e-mail: bipin@amrita.edu).

Shyam Diwakar is the Associate Professor and Lab Director of Computational Neuroscience and Neurophysiology Laboratory at the School of Biotechnology. Amrita Vishwa Vidyapeetham, India. (e-mail: shyam@amrita.edu).

Submitted, 29 August 2018. Resubmitted 29 October 2018. Final acceptance 30 October 2018. Final version published as submitted by the authors.