Dynamic Cluster-based Retrieval and Discovery for Biomedical Literature

Michael Segundo Ortiz
Carolina Health Informatics Program, University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
msortiz@unc.edu

Heejun Kim
School of Information and Library Science, University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
heejunk@email.unc.edu

Mengqian Wang
Carolina Health Informatics Program, University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
mengqian@email.unc.edu

Kazuhiro Seki
Department of Intelligence and Informatics, Konan University
Kobe, Hyogo, Japan
seki@konan-u.ac.jp

Javed Mostafa
Carolina Health Informatics Program, School of Information and Library Science, University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
jm@unc.edu

ABSTRACT

Due to increased specialization and experimentation, the volume of biomedical literature is rapidly increasing, where the current modalities of search and retrieval system can no longer support effective and efficient knowledge discovery. Standard information retrieval systems such as PubMed make assumptions as to users' prior knowledge and expect them to formulate a proper query term for the information they are looking for. There exist user feedback mechanisms to help users reformulate their queries, which still assumes that users know how the search results could be effectively narrowed down by way of additional keywords and/or filters. As an alternative, we revisit the Scatter/Gather information retrieval paradigm. Specifically, we explore a real-time dynamic cluster-based document browsing approach in the biomedical domain, discuss the system architecture involving keyword discovery and dynamic clustering, and present a working prototype with a relevant use case in comparison with a standard ranking-based information retrieval system.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; • **Computing methodologies → Cluster analysis**; • **Applied computing → Health informatics**.

KEYWORDS

exploratory search, unsupervised learning, information retrieval, biomedicine

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3342191

ACM Reference Format:
Michael Segundo Ortiz, Heejun Kim, Mengqian Wang, Kazuhiro Seki, and Javed Mostafa. 2019. Dynamic Cluster-based Retrieval and Discovery for Biomedical Literature. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3307339.3342191

1 INTRODUCTION

Automated clustering of scientific publications and spatial encoding of information visualization techniques is of emerging importance for digital libraries [24]. This is especially true for biomedical libraries. The current ranking-based retrieval model assumes that users have a clear understanding of the information need and thus could formulate a proper query that facilitates a ranked list of various information types that are in descending order of relevancy. However, prior understanding of terminology or information space structure may not exist [29, 31]. Moreover, even with a properly formulated query in biomedicine, a flat list of thousands of results does not enable discovery of latent information, logical connections among concepts, or transitive properties within a particular information space; processes of which can be automated and packaged as a real-time generative model to aid researchers in their information exploration and hypothesis generation.

The current mode of access, generally speaking, is a look-up procedure, similar to searching an index in a large textbook. To ultimately improve access, it seems intuitive to provide a table of contents metaphor for users to explore what topical content is in a collection and iteratively reach a more specific information target; a process called exploratory search[12, 30]. For example, this paradigm is also utilized by the 2012 ACM Computing Classification System (CCS)[1] whereby you can generate codes to classify conference documents and also iteratively reach highly relevant literature based on topical content. Such an interactive ontology would be tremendously useful for Medical Subject Headings (MeSH) given that PubMed is arguably the largest biomedical literature base in

[1] https://dl.acm.org/ccs/ccs.cfm

the world, for example. These classification models serve as very powerful filtering tools to narrow the focus of a search. However, there is also the need to build robust tools on top of these paradigms that model spatial semantics and latent information structure in order to enable discovery and generate hypotheses.

In this work we describe early stage development of a prototype system for exploring topical content in the biomedical literature. The organization of this work is as follows; in the related work section we discuss the benefits and limitations of other exploratory information retrieval systems in biomedicine and propose a revival of the Scatter/Gather dynamic clustering paradigm. In the system architecture and description sections we discuss the data flow from server to client, data processing, and visualization. Lastly, we provide a use-case for retrieving information on genomic editing and briefly contrast this with a conventional search in PubMed.

2 RELATED WORK

In this section we will briefly introduce previous work on exploratory search systems and the various modalities they employ. On the user end, classic retrieval optimization involves incremental feedback to search systems in terms of revised queries. However, practice shows that the added step of revising queries is non-desirable [13]. Moreover, some evidence suggests that users spend the least amount of time on queries and focus more on results and facets or filters of retrieved information [10]. Other research indicates that complex search tasks may result in longer queries which implies that more thought or prior knowledge must go into such a query [5]. These findings indicate that exploratory information systems can be a more intuitive model when search intent is unclear. User feedback is essential, however, revising queries appears to be an inferior method.

Routsalo et al. [20, 21] developed a system called SciNet, specifically targeting interactive user intent modeling. User intent is often vague, and proper query terms may not be known or provided by a user. SciNet attempts to solve this problem by an interactive visual interface. The system starts with a user query and returns a wide spectrum of keywords to suggest potential intents on a radar chart-like screen in addition to a standard ranked list of documents. On the radar screen, users can give their feedback by moving any keywords they find relevant to the center of the radar, and vice versa. Given the feedback, the system updates the estimate of the user intent and dynamically updates the results.

Sciascio et al. [22] developed another system called uRank. The system provides a keyword list summarizing a document collection and a document list showing a ranked list of documents with stacked bars of relevance scores for each keyword. The users choose keywords from the keyword list and, for each keyword, its weight can be adjusted by a slider. The rankings of the documents are dynamically changed reflecting the chosen keywords and their weights. The stacked bars of relevance scores help the users to see how much each document is relevant with respect to individual keywords.

Another approach is Scatter/Gather [3, 4, 8]. Scatter/Gather is a browsing-based information exploratory model and presents topically coherent groups (clusters) of documents with descriptive textual summaries as opposed to a ranked list of documents. In other words, documents are "scattered" into topical clusters for browsing. The users then browse the generated clusters and "gather" the ones that are interesting or relevant. Based on the selection, the documents in the selected clusters are re-clustered and presented to the users again. This process is repeated until a user feels they have identified their information target. Figure 1 depicts the Scatter/Gather browsing paradigm.

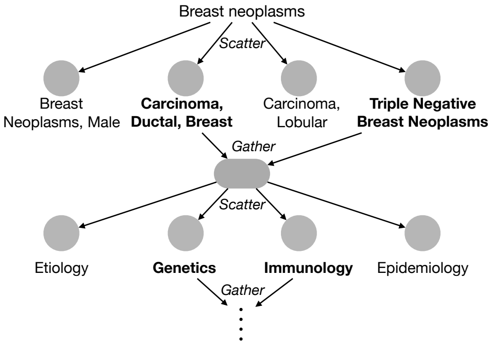


Figure 1: Illustration of Scatter/Gather browsing paradigm [4] for the breast cancer literature.

This particular example starts with a query "Breast Neoplasms" to form a document collection and scatters four topical clusters (i.e., *Breast Neoplasms Male*, *Carcinoma Ductal Breast*, etc.). The user then gathers two clusters of interest: *Carcinoma Ductal, Breast* and *Triple Negative Breast Neoplasms*, then four new and more specific clusters (i.e., *Etiology*, *Epidemiology*, etc.) are identified within the selected clusters and gathered/scattered again on *Immunology* and *Genetics*, indicating the users' interest in immunologic and genetic information in relation to the specified forms of breast neoplasms selected in the first scatter phase. This Scatter/Gather browsing is particularly helpful in cases where a user is unsure about formulating a specific search query because it allows him/her to explore the general content of a collection and iteratively refine their information need based on the relationships among concepts.

Scatter/Gather is built on the cluster hypothesis [17], which states that "closely associated documents tend to be relevant to the same requests". Because multiple documents are clustered into topically related groups, the Scatter/Gather browsing model may reduce user burden by providing a dynamic table of contents metaphor as opposed to querying a collection with vague intent and then scrolling through a potentially large set of individual documents which we analogize to searching a vast textbook-like index.

The effectiveness of Scatter/Gather has been empirically investigated by several studies. Hearst et al. [8] reported that relevant documents for a given query tended to be clustered together and that the users were able to choose the right cluster with the largest number of relevant documents in more than 80% of cases. Gong et al. [6] also reported that this model was found to be particularly helpful for search tasks unfamiliar to users. Ortiz et al. [16] examined a more fundamental question of the effectiveness of cluster-based browsing models and systematically studied various parameters affecting cluster quality.

Despite its potential benefits, a cluster-based browsing search interface has not been extensively studied for biomedical literature primarily due to two challenges. First, clustering must be fast for an arbitrarily large number of documents. The Buckshot clustering algorithm [4] proposed with Scatter/Gather runs in time $O(kn)$ where k is the number of clusters, but linear time is still not fast enough to execute clustering on the fly for a large document collection. A constant time algorithm was also proposed [3], which builds a static hierarchy of clusters in advance in an offline process. However, appropriate grouping of documents will change both for an initial query and for chosen clusters [8], and thus a static cluster structure is often sub-optimal. The second challenge is to develop an intuitive and effective user interface. In the past, much work adopting Scatter/Gather simply used text-based interface, and it is unclear how document clusters and Scatter/Gather mechanisms are best visualized.

3 DYNAMIC CLUSTER-BASED BROWSING

This section describes the design and implementation details of our dynamic cluster-based browsing system, DCB^2. We adopt the Scatter/Gather paradigm for its potential and tackle the open issues concerning real-time clustering, dynamicity, and effective interface design.

3.1 Overview

Figure 2 depicts the architecture and data flow of DCB^2, where the dotted lines indicate iterative processes by a user. The server side system runs on Amazon Web Services Elastic Cloud Compute (AWS EC2)[2] and is implemented by the Flask web framework.[3] The document collection was downloaded from Public Library of Science (PLOS), indexed by Apache Solr,[4] and locally resides in the server for efficiency. The client side is built on a JavaScript visualization library D3.js.[5] For interactive and iterative browsing, the client uses Ajax to asynchronously communicate with the server which eliminates the need to reload the web page as content is dynamically explored. The following sections describe the core components of the system in more detail.

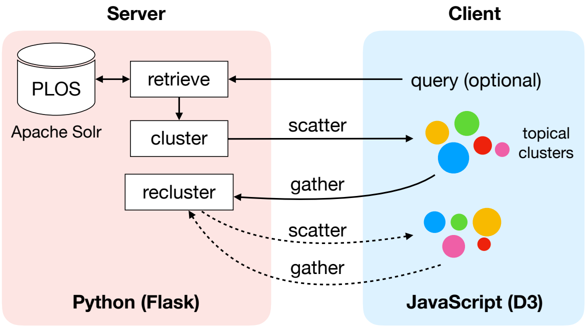


Figure 2: System architecture and data flow.

3.2 Initial Data Retrieval and Clustering

DCB^2 starts with a text box for a user query as with other keyword-based search systems (Figure 3), although our system can initiate information retrieval with or without a query. When a query is given, it retrieves N latest articles satisfying the query. When no query is given, it simply retrieves N latest articles in the collection without considering any particular topic. We fix N to a constant value so that clustering is completed in a constant time regardless of a query, facilitating real-time processing. The rationale behind this design is random sampling in order to deal with the potentially large number of documents. This is similar to the idea of mini-batch k-means [23]. Mini-batch k-means first performs k-means clustering on a random subset of data to compute cluster centroids and then determines the membership of all the data points. Mini-batch k-means is reported to converge to near-optima several orders of magnitude faster than standard k-means. Instead of complete randomness, however, our system favors recency as we are generally interested in more recent information in the biomedical domain. Although limiting the number of documents by N will certainly influence the resulting cluster structure, we assume the effect is limited for a large N. We will investigate the validity of this assumption in Section 4.

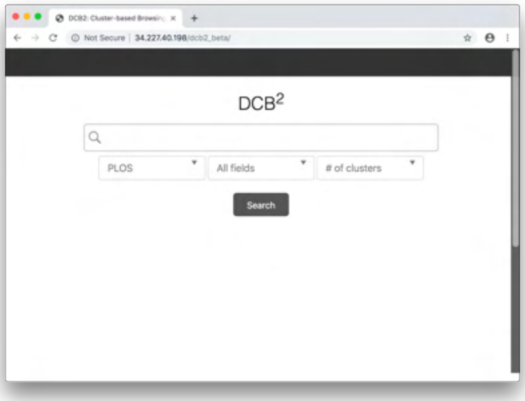

DCB2: Cluster-based Browsi...
Not Secure | 34.227.40.198/dcb2_beta/
DCB²
PLOS
All fields
# of clusters
Search

Figure 3: DCB^2 start page.

As for query language, DCB^2 can understand a wide range of query syntax accepted by Solr, such as Boolean queries, range queries, phrases and wild cards. However, we expect more general queries since our system focuses on exploratory search where users' search intent is not yet clear. Currently, the system retrieves titles, abstracts and body texts and simply concatenates them, but other data including journal names, authors and affiliations are also indexed and readily available for future use. Note that the search field (default is "all fields") and the number of clusters (default is "10") can be specified by the search page for convenience.

After retrieving the article information, the system executes the following processes in this order, which was reported to be beneficial for constructing high quality clusters [16].

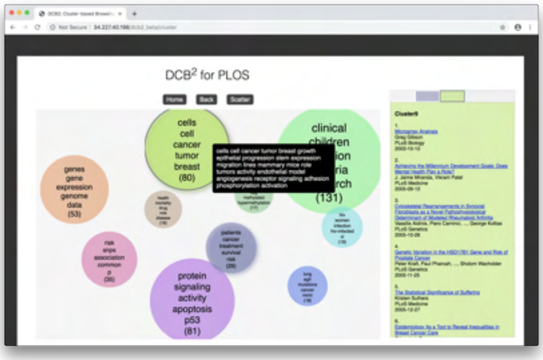
[2] https://aws.amazon.com/ec2/
[3] http://flask.pocoo.org
[4] http://lucene.apache.org/solr/
[5] https://d3js.org

- Keyword discovery: The system first identifies prominent terms to represent the retrieved document set. For this purpose, we adopt a statistical approach called Vocabulary Cluster Generating System (VCGS) [15]. VCGS discovers keywords based on term and document frequencies. Using only the discovered keywords, the document set is represented as a term-document matrix M with tf-idf term weighting [26]. This process greatly reduces the data size.
- Latent semantic analysis (LSA) [11]: To furher reduce the dimensionality of the term-document matrix M and to discover latent associations among keywords, LSA is applied to M. LSA is a matrix decomposition technique that extracts and represents the contextual usage of terms in a collection of documents. The transformation of a full featured matrix to a dimensionally reduced matrix helps reveal implicit semantic associations between documents. The dimensionally reduced matrix can be obtained by first decomposing Mfi into $U\Sigma V^T$, where U and V are orthogonal matrices and Σ is a diagonal matrix with the eigenvalues of the eigenvectors in descending order. The first n rows of matrix V (corresponding to the n largest singular values in Σ) is the n dimensionally reduced matrix. Currently, we empirically use 50 as the number of components (dimensions).
- Clustering: The document set is then topically clustered for presentation. We use the popular k-means++ algorithm [1], where k is set to 10 by considering the trade-off between readability and informativeness. Alternatively, users can choose the value of k from 2 to 10 in the start page.

After these processes, the system generates a set of keywords to describe each cluster for the next visualization stage. More precisely, the centroid of each cluster in the LSA-reduced space is transformed back to the original term-document space and is represented as a term vector. From the vector, keywords with n highest tf-idf values are selected as the description of the cluster. In addition, the centroids in the LSA-reduced space are transformed to coordinate space by t-Distributed Stochastic Neighbor Embedding (t-SNE) [28] and plotted in 2D. Only the descriptions (keyword set) and 2D coordinates for the clusters are sent to the client for efficiency, and other information is retained as session data on the server.

3.3 Visualization

We designed a preliminary Scatter/Gather browsing interface and developed a functional prototype.[6] We relied on Shneiderman's visual information seeking mantra [25]—*overview first, zoom and filter, then details on demand*—for the design process to provide overviews of clusters and to show details according to users' interest. The design incorporates two visualization panels and buttons for Scatter/Gather (Figure). The left panel (hereinafter cluster panel) displays clusters processed on the server-side and the right panel (hereinafter document panel) presents documents that belong to the selected clusters in the cluster panel.


DCB2: Cluster-based Browsi...
Not Secure | 34.227.40.198/dcb2_beta/cluster
DCB² for PLOS
Home
Back
Scatter
cells cell cancer tumor breast (80)
clinical children ...ion ...ria ...rch (131)
genes gene expression genome data (53)
risk snps association common p (35)
protein signaling activity apoptosis p53 (81)
patients cancer treatment survival risk (29)
health mortality drug risk disease (18)
methylated hypermethylation (17)
hiv women infection hiv-infected p (18)
lung egfr mutations cancer nsclc (18)

cells cell cancer tumor breast growth epithelial progression stem expression migration lines mammary mice role tumors activity endothelial model angiogenesis receptor signaling adhesion phosphorylation activation

Cluster9

1.
Microarray Analysis
Greg Gibson
PLoS Biology
2005-10-13

2.
Achieving the Millennium Development Goals: Does Mental Health Play a Role?
J. Jaime Miranda, Vikram Patel
PLoS Medicine
2005-09-13

3.
Cytoskeletal Rearrangements in Synovial Fibroblasts as a Novel Pathophysiological Determinant of Modeled Rheumatoid Arthritis
Vassilis Aidinis, Piero Carninci, ..., George Kollias
PLoS Genetics
2005-10-28

4.
Genetic Variation in the HSD17B1 Gene and Risk of Prostate Cancer
Peter Kraft, Paul Pharoah, ..., Sholom Wacholder
PLoS Genetics
2005-11-25

5.
The Statistical Significance of Suffering
Kristen Suthers
PLoS Medicine
2005-12-27

6.
Epidemiology As a Tool to Reveal Inequalities in Breast Cancer Care

Figure 4: Initial search results for query "breast neoplasms" by DCB^2 presented as topical clusters.

The cluster centroids correspond to the coordinates in semantic space constructed by t-SNE, which reflect their relative semantic relatedness. In the center of each cluster, representative keywords are displayed, as well as the number of documents in parentheses. The area of the cluster (circle) is proportional to the number of documents. If users move their mouse pointer over a cluster, corresponding bibliographies of documents (article titles, author names, journal titles, and publication dates in this order) with hyperlinks to the article registered in PubMed appear temporarily in the document panel. Users can click a cluster(s) of interest for the Gather process and the references will be displayed in the document panel until the cluster is deselected. Another click can deselect the cluster(s). If multiple clusters are selected, users can navigate the corresponding documents for each cluster through the tabs located at the top of the document panel. The circle in the cluster panel and its corresponding tab in the document panel are presented in the same color to facilitate intuitive navigation. Users can zoom or pan the visualization panel as needed. By clicking the "Scatter" button, the chosen clusters of documents are re-clustered (see Section 3.4) and scattered again. After examining the results, users can either try a new Scatter or return to the previous step. Users can iterate through this Scatter/Gather process interactively until they satisfy their information need.

3.4 Re-clustering

Upon receiving a set of selected clusters for the *gather phase*, the system retrieves the document ID data within the clusters from the session data and performs a *scatter phase* that involves a series of processes from keyword discovery to clustering as described in Section 3.2. One may think that these processes are redundant. However, it should be stressed that these processes are crucial for the dynamicity of DCB^2 to identify new keywords, which would be different from the previously identified keywords and, consequently, should yield more relevant topical clusters. The descriptions of the resulting clusters and their 2D coordinates are computed in the same way as previously described and sent to the client. When the total number of articles in the selected clusters becomes smaller than a predefined threshold, their bibliographic data are also sent to the client as well for examination.

[6] http://34.227.40.198/dcb2_beta/

4 EVALUATION

In this section, we first evaluate our sampling-based clustering approach quantitatively and then walk through the prototype system with a possible use case to demonstrate how DCB^2 could be used for information seeking.

4.1 Dynamic Clustering

To realize a cluster-based document browsing system for a large biomedical bibliography database, clustering should ideally be done in a constant time irrespective of the size of the search result. However, existing clustering algorithms running in a constant time typically rely on a pre-computed static hierarchy of categories, which is not suited for iterative, dynamic cluster-based browsing.

To circumvent the problem, DCB^2 uses a simple sampling-based clustering approach, retrieving only the N latest articles for a given query. Although the time complexity of the clustering algorithm itself (k-means) is not constant, the clustering process completes approximately in a constant time for fixed N. The running time could be short enough to perform on the fly if N is small. On the other hand, small N would not produce clusters representative of the entire search results. Therefore, we empirically examined the relation between the sample size (number of documents) and the quality of clusters so as to find appropriate N which could produce clusters with the quality close to those created from the entire search results.

4.1.1 Experimental Setups. For this experiment, we needed a data set in which each document is labeled with a category or class as ground truth. Following the methodology by Ortiz et al. [16], we considered Medical Subject Headings (MeSH) major topics as categories. Specifically, we used the MeSH term “neoplasms by site” to construct our data set as follows:

(1) On the PubMed website, we used a query `"Neoplasms by Site"[MeSH Major Topic]` to retrieve articles on Feb 26, 2019. We restricted the search only to PLOS journals by specifying journal names so that the resulting data set would better reflect the characteristics of the PLOS archive. Note that all the articles annotated with the MeSH terms below “Neoplasms by Site” in the MeSH hierarchy were also retrieved by this query.
(2) All the MeSH terms given to the articles were generalized to the MeSH terms right below “Neoplasms by Site”. Then, six most frequent MeSH terms, “Digestive System Neoplasms” (4,416), “Breast Neoplasms” (2,647), “Urogenital Neoplasms” (2,313), “Thoracic Neoplasms” (1,770), “Endocrine Gland Neoplasms” (1,516), “Head and Neck Neoplasms” (1,413), were identified and treated as topical categories (the numbers in the parentheses show the number of articles annotated with respective MeSH terms). These six categories were chosen such that each category would have at least 1,000 articles. After deleting articles annotated with none of these MeSH terms, 12,530 articles remained.
(3) The same query as above was used to retrieve full-text articles from the PubMed Central database. From the retrieved articles, the body texts of the remaining 12,530 articles were extracted.

There are many criteria for evaluating the quality of clusters. Among them, we used Adjusted Mutual Information (AMI) following a recommendation by Romano et al. [19]. AMI is based on mutual information and is a measure of agreement between true labels and those by a clustering algorithm. It quantifies the amount of information shared between the two assignments and it is defined by term probability distributions and the information-theoretic measure of entropy. AMI is adjusted for chance by using the expected value of mutual information for normalization.

4.1.2 Results. Figure 5 shows the relation between the number of sampled documents and the quality of generated clusters in AMI, where the sample size was gradually increased from 100 to 12,530. We compared three different data, i.e., titles only (denoted as “Title”), titles and abstracts (denoted as “Abstract”), and titles and abstracts and body texts (denoted as “Full text”). One can observe that AMI shaply improved as the sample size increased up to 2,000 for Title and Abstract and then it became more or less stable for the rest. Somewhat unexpectedly, Titles worked comparably with Abstracts, although using abstracts tended to produce more reliable results. On the other hand, using full-text data was not as effective as using titles and/or abstracts, which is consistent with the experiment on a breast cancer subset of PubMed Central data (BRCA-FULL) [16].

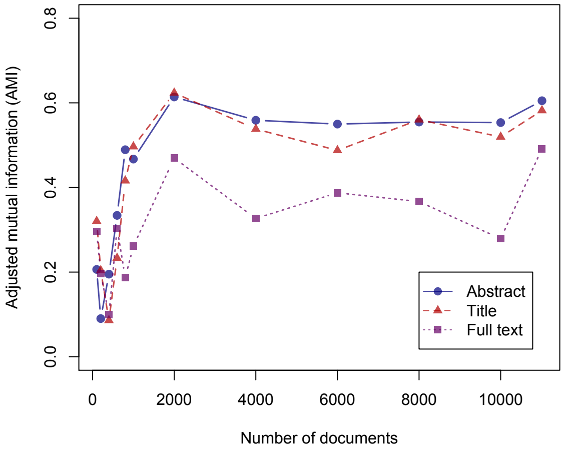


Figure 5: Relation between the number of documents and cluster quality in AMI.

Then, we compared the processing time for Title, Abstract, and Full text. The processing time was measured from loading data to clustering. The results are shown in Figure 6. Naturally, Title was the fastest, followed by Abstract and then Full text, and the processing time grew rapidly as the number of documents increased, especially for Full text. Based on these observations, it is recommended to use 2,000 latest articles (i.e., $N = 2,000$) and to use titles and abstracts for constructing topical clusters. We plan to investigate the validity of these parameters also through a user study in future work.

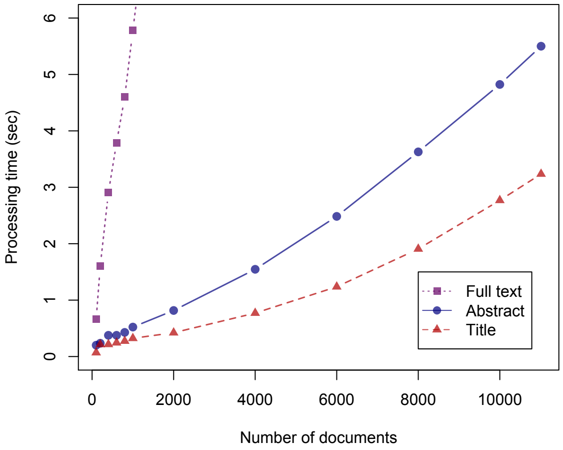


Figure 6: Relation between the number of documents and processing time.

4.2 Use Case

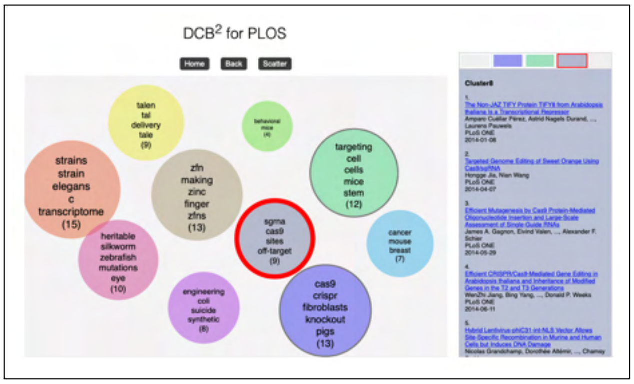
As a use case, we considered a query "genomic editing" and illustrates how a researcher would navigate through the literature using DCB^2. Figure 7 depicts the resulting clusters for this query. The user selects three semantically related clusters based on the close spatial distance and interest in the concepts of "sgrna" (single guide RNA), "sites", "off-target", "targeting", and "crispr" (clustered regularly interspaced short palindromic repeats).

In this instance for example, we imagine a researcher wanting to understand the landscape of genomic editing technologies by exploring the highly-weighted concepts and their latent associations based on spatially encoded information and then mapping the concepts back to their original publications for review when desired. Note that Figure 8 also displays the results for an equivalent query in PubMed Central although an overload of information is presented in which the user must apply filtering tools, re-formulate the query, or sequentially scroll through the ranked results and extrapolate concepts on their own. In contrast, we believe that the Scatter/Gather paradigm may offer the benefit of treating searching as learning [7] in a cognitively less demanding modality.

Figure 9 shows new clusters from the resulting *Scatter* phase. The user now selects one cluster based on the keywords "disruption" and "cas9" (CRISPR associated protein 9) which results in a very small and specific document set on the right-side panel. The researcher quickly learns that genome editing technologies may induce downstream effects based on an error-prone repairing mechanism that leads to mutation and gene disruption and that methods are being developed to improve the fidelity of the technology. For example, much work on improving the technology is already underway[2, 9, 14, 18, 27]

5 CONCLUSION

This paper presented our ongoing work for developing a dynamic cluster-based document browsing system, called DCB^2, for exploring the biomedical literature. Our system adopts the Scatter/Gather paradigm and focuses on three important dimensions, i.e., real-time and dynamic clustering, efficient/accurate representation, and effective presentation. For the first two, we applied on-the-fly keyword


DCB2 for PLOS
Home
Back
Scatter
talen tal delivery tale (9)
behavioral mice (4)
strains strain elegans c transcriptome (15)
zfn making zinc finger zfns (13)
targeting cell cells mice stem (12)
sgrna cas9 sites off-target (9)
heritable silkworm zebrafish mutations eye (10)
cancer mouse breast (7)
engineering coli suicide synthetic (8)
cas9 crispr fibroblasts knockout pigs (13)

Cluster8

1.
The Non-JAZ TIFY Protein TIFY8 from Arabidopsis thaliana Is a Transcriptional Repressor
Amparo Cuéllar Pérez, Astrid Nagels Durand, ..., Laurens Pauwels
PLoS ONE
2014-01-08

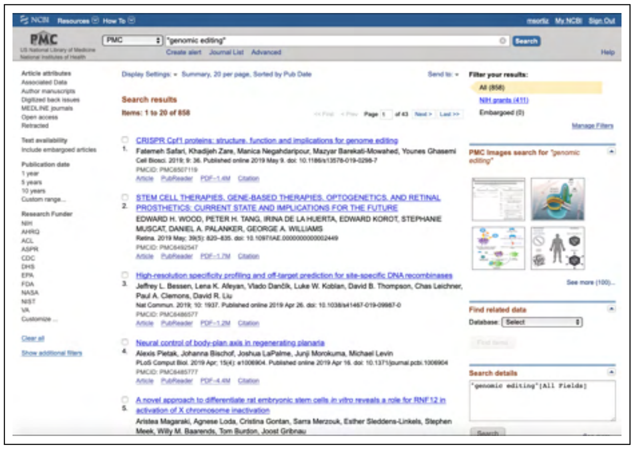
2.
Targeted Genome Editing of Sweet Orange Using Cas9/sgRNA
Hongge Jia, Nian Wang
PLoS ONE
2014-04-07

3.
Efficient Mutagenesis by Cas9 Protein-Mediated Oligonucleotide Insertion and Large-Scale Assessment of Single-Guide RNAs
James A. Gagnon, Eivind Valen, ..., Alexander F. Schier
PLoS ONE
2014-05-29

4.
Efficient CRISPR/Cas9-Mediated Gene Editing in Arabidopsis thaliana and Inheritance of Modified Genes in the T2 and T3 Generations
WenZhi Jiang, Bing Yang, ..., Donald P. Weeks
PLoS ONE
2014-06-11

5.
Hybrid Lentivirus-phiC31-int-NLS Vector Allows Site-Specific Recombination in Murine and Human Cells but Induces DNA Damage
Nicolas Grandchamp, Dorothée Altémir, ..., Charney

Figure 7: DCB^2 gathering clusters for "genomic editing".



"genomic editing"

Display Settings: Summary, 20 per page, Sorted by Pub Date

Search results

Items: 1 to 20 of 858

Page 1 of 43

All (858)
NIH grants (411)
Embargoed (0)

1. CRISPR Cpf1 proteins: structure, function and implications for genome editing
Fatemeh Safari, Khadijeh Zare, Manica Negahdaripour, Mazyar Barekati-Mowahed, Younes Ghasemi
Cell Biosci. 2019; 9: 36. Published online 2019 May 9. doi: 10.1186/s13578-019-0298-7
PMCID: PMC6507119
Article PubReader PDF–1.4M Citation

2. STEM CELL THERAPIES, GENE-BASED THERAPIES, OPTOGENETICS, AND RETINAL PROSTHETICS: CURRENT STATE AND IMPLICATIONS FOR THE FUTURE
EDWARD H. WOOD, PETER H. TANG, IRINA DE LA HUERTA, EDWARD KOROT, STEPHANIE MUSCAT, DANIEL A. PALANKER, GEORGE A. WILLIAMS
Retina. 2019 May; 39(5): 820–835. doi: 10.1097/IAE.0000000000002449
PMCID: PMC6492547
Article PubReader PDF–1.7M Citation

3. High-resolution specificity profiling and off-target prediction for site-specific DNA recombinases
Jeffrey L. Bessen, Lena K. Afeyan, Vlado Dančík, Luke W. Koblan, David B. Thompson, Chas Leichner, Paul A. Clemons, David R. Liu
Nat Commun. 2019; 10: 1937. Published online 2019 Apr 26. doi: 10.1038/s41467-019-09987-0
PMCID: PMC6486577
Article PubReader PDF–1.2M Citation

4. Neural control of body-plan axis in regenerating planaria
Alexis Pietak, Johanna Bischof, Joshua LaPalme, Junji Morokuma, Michael Levin
PLoS Comput Biol. 2019 Apr; 15(4): e1006904. Published online 2019 Apr 16. doi: 10.1371/journal.pcbi.1006904
PMCID: PMC6485777
Article PubReader PDF–4.4M Citation

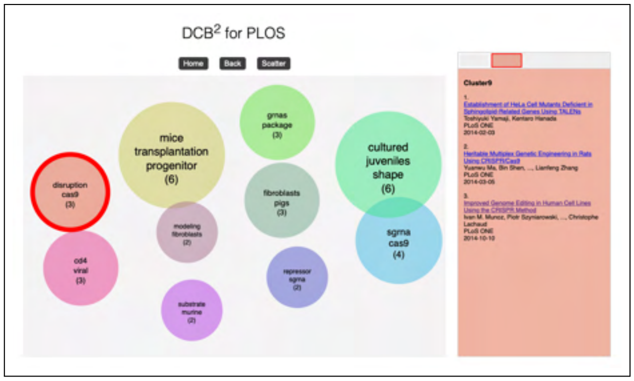
5. A novel approach to differentiate rat embryonic stem cells in vitro reveals a role for RNF12 in activation of X chromosome inactivation
Aristea Magaraki, Agnese Loda, Cristina Gontan, Sarra Merzouk, Esther Sleddens-Linkels, Stephen Meek, Willy M. Baarends, Tom Burdon, Joost Gribnau

PMC Images search for "genomic editing"

Database: Select

Search details
"genomic editing"[All Fields]

Figure 8: PubMed Central default results for "genomic editing".



Cluster9

1.
Establishment of HeLa Cell Mutants Deficient in Sphingolipid-Related Genes Using TALENs
Toshiyuki Yamaji, Kentaro Hanada
PLoS ONE
2014-02-03

2.
Heritable Multiplex Genetic Engineering in Rats Using CRISPR/Cas9
Yuanwu Ma, Bin Shen, ..., Lianfeng Zhang
PLoS ONE
2014-03-05

3.
Improved Genome Editing in Human Cell Lines Using the CRISPR Method
Ivan M. Munoz, Piotr Szyniarowski, ..., Christophe Lachaud
PLoS ONE
2014-10-10

Figure 9: DCB^2 scattered clusters for "comparative genomics".

discovery and proposed sampling-based clustering. For the latter, we designed and built an intuitive user interface. To demonstrate the validity of the approach, we examined the relation between cluster quality and sample size and showed that using around 2,000 documents produced as good clusters as using the entire document collection with much less processing time. Also, a possible use case was provided to illustrate the utility of the system. Future work will examine scalability, building more efficient indices to allow faster iteration, generating interpretable descriptions for clusters, and system evaluation involving prospective users.

ACKNOWLEDGMENTS

The NIH-NLM T15 grant 5T15LM012500-02, United Health Foundation's ENABLE grant, JSPS KAKENHI Grant Number 18K11558 and MEXT, Japan, provided support for this work.

REFERENCES

[1] David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 1027–1035.

[2] Nurit Assia Batzir, Adi Tovin, and Ayal Hendel. 2017. Therapeutic Genome Editing and its Potential Enhancement through CRISPR Guide RNA and Cas9 Modifications. *Pediatric endocrinology reviews: PER* 14, 4 (2017), 353–363.

[3] Douglass R. Cutting, David R. Karger, and Jan O. Pedersen. 1993. Constant Interaction-time Scatter/Gather Browsing of Very Large Document Collections. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*. ACM, New York, NY, USA, 126–134. https://doi.org/10.1145/160688.160706

[4] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. 1992. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '92)*. ACM, New York, NY, USA, 318–329. https://doi.org/10.1145/133160.133214

[5] Souvick Ghosh, Manasa Rath, and Chirag Shah. 2018. Searching as Learning: Exploring Search Behavior and Learning Outcomes in Learning-related Tasks. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, 22–31.

[6] Xuemei Gong, Weimao Ke, Yan Zhang, and Ramona Broussard. 2013. Interactive Search Result Clustering: A Study of User Behavior and Retrieval Effectiveness. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '13)*. ACM, New York, NY, USA, 167–170. https://doi.org/10.1145/2467696.2467726

[7] Preben Hansen and Soo Young Rieh. 2016. Recent advances on searching as learning: An introduction to the special issue.

[8] Marti A. Hearst and Jan O. Pedersen. 1996. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96)*. ACM, New York, NY, USA, 76–84. https://doi.org/10.1145/243199.243216

[9] Melissa L Kelley, Žaklina Strezoska, Kaizhang He, Annaleen Vermeulen, and Anja van Brabant Smith. 2016. Versatility of chemically synthesized guide RNAs for CRISPR-Cas9 genome editing. *Journal of biotechnology* 233 (2016), 74–83.

[10] Bill Kules, Robert Capra, Matthew Banta, and Tito Sierra. 2009. What do exploratory searchers look at in a faceted search interface?. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 313–322.

[11] Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes* 25, 2-3 (1998), 259–284.

[12] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[13] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (April 2006), 41–46. https://doi.org/10.1145/1121949.1121979

[14] Su Bin Moon, Jeong-Heon Ko, Jin-Soo Kim, Yong-Sam Kim, et al. 2019. Improving CRISPR Genome Editing by Engineering Guide RNAs. *Trends in biotechnology* (2019).

[15] J. Mostafa, L. M. Quiroga, and M. Palakal. 1998. Filtering medical documents using automated and human classification methods. *Journal of the American Society for Information Science* 49, 14 (1998), 1304–1318.

[16] Michael Segundo Ortiz, Kazuhiro Seki, and Javed Mostafa. 2018. Toward Exploratory Search in Biomedicine: Evaluating Document Clusters by MeSH as a Semantic Anchor. *CoRR* arXiv:1812.02129 (2018). arXiv:1812.02129 https://arxiv.org/abs/1812.02129

[17] C. J. Van Rijsbergen. 1979. *Information Retrieval* (2nd ed.). Butterworth-Heinemann, Newton, MA, USA.

[18] Khadim Hussain Rimsha Farooq, Shahid Nazir, Muhammad Rizwan Javed, and Nazish Masood. 2018. CRISPR/Cas9; A robust technology for producing genetically engineered plants. *Cell Mol Biol (Noisy le Grand)* 64, 14 (2018).

[19] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. 2016. Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research* 17, 1 (2016), 4635–4666.

[20] Tuukka Ruotsalo, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. 2014. Interactive Intent Modeling: Information Discovery Beyond Search. *Commun. ACM* 58, 1 (Dec. 2014), 86–92. https://doi.org/10.1145/2656334

[21] Tuukka Ruotsalo, Jaakko Peltonen, Manuel J. A. Eugster, Dorota Głowacka, Patrik Floréen, Petri Myllymäki, Giulio Jacucci, and Samuel Kaski. 2018. Interactive Intent Modeling for Exploratory Search. *ACM Trans. Inf. Syst.* 36, 4, Article 44 (Oct. 2018), 46 pages. https://doi.org/10.1145/3231593

[22] Cecilia Di Sciascio, Vedran Sabol, and Eduardo Veas. 2017. Supporting Exploratory Search with a Visual User-Driven Approach. *ACM Trans. Interact. Intell. Syst.* 7, 4, Article 18 (Dec. 2017), 35 pages. https://doi.org/10.1145/3009976

[23] D Sculley. 2010. Web-scale k-means Clustering. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 1177–1178. https://doi.org/10.1145/1772690.1772862

[24] Hui Shi, Wu He, and Guandong Xu. 2018. Workshop Proposal on Knowledge Discovery from Digital Libraries. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*.

[25] Ben Shneiderman. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages (VL '96)*. IEEE Computer Society, Washington, DC, USA, 336–. http://dl.acm.org/citation.cfm?id=832277.834354

[26] Karen Sparck Jones. 1972. Statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28, 1 (1972), 11–20.

[27] Fei Teng, Tongtong Cui, Qingqin Gao, Lu Guo, Qi Zhou, and Wei Li. 2019. Artificial sgRNAs engineered for genome editing with new Cas12b orthologs. *Cell discovery* 5, 1 (2019), 23.

[28] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[29] Ryen W White, Bill Kules, Steven M Drucker, et al. 2006. Supporting exploratory search, introduction, special issue, communications of the ACM. *Commun. ACM* 49, 4 (2006), 36–39.

[30] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.

[31] Max Wilson, Alistair Russell, Daniel A Smith, et al. 2006. mSpace: improving information access to multimedia domains with multimodal exploratory search. *Commun. ACM* 49, 4 (2006), 47–49.