Author manuscript, published in Proceedings of the IEEE Pacific Visualization Symposium, pp. 81-85, 2021.
DOI: 10.1109/PacificVis52677.2021.00019

Know-What and Know-Who: Document Searching and Exploration using Topic-Based Two-Mode Networks

Jian Zhao* University of Waterloo
Maoyuan Sun Northern Illinois University
Patrick Chiu FXPAL
Francine Chen Toyota Research Institute
Bee Liew FXPAL

ABSTRACT

This paper proposes a novel approach for analyzing search results of a document collection, which can help support *know-what* and *know-who* information seeking questions. Search results are grouped by topics, and each topic is represented by a two-mode network composed of related documents and authors (i.e., biclusters). We visualize these biclusters in a 2D layout to support interactive visual exploration of the analyzed search results, which highlights a novel way of organizing entities of biclusters. We evaluated our approach using a large academic publication corpus, by testing the distribution of the relevant documents and of lead and prolific authors. The results indicate the effectiveness of our approach compared to traditional 1D ranked lists. Moreover, a user study with 12 participants was conducted to compare our proposed visualization, a simplified variation without topics, and a text-based interface. We report on participants' task performance, their preference of the three interfaces, and the different strategies used in information seeking.

Index Terms: Human-centered computing—Visual analytics; Applied computing—Document searching

1 INTRODUCTION

Common information searching tasks about a document collection are of the types *know-what* and *know-who* [17]. Popular search engines such as Google retrieve a list of documents but not a list of authors (who are associated closely with the documents) in response to a query. One way to provide both documents and authors is to employ two-mode networks to analyze and visualize the data [20,34]. A *two-mode network* is a special kind of network which consists of two types of entities (nodes) and relations (links) between the two entity types, which is often the output of data biclustering [29]. In our case, the entity types are documents and authors, and the relations between entities are formed by authorship. This document-and-author based information seeking can provide many benefits in different applications, for example, exploring academic publication corpora or digital libraries [2, 9].

One scenario is browsing and exploring conference publications. For example, NIPS is a highly influential venue in machine learning and computer science. A researcher or student may wish to find important papers and authors who are experts on a particular subject. Moreover, some of the seminal papers are of historical interest and can help understand the development of an area of research.

However, sensemaking of the retrieved documents and authors is challenging. First, prior studies indicate that displaying general search results in a ranked list is not sufficient for users to best utilize the results for new discoveries [22, 33]. Making things worse, there are two types of objects, documents and authors, in the results; a ranked list according to either of these two cannot adequately reveal their complicated many-to-many relationships, thus hindering users from understanding and making use of the retrieved information.

To address these challenges, we present a novel *analysis pipeline* that employs topic modeling [3] to cluster the search results into maximal two-mode sub-networks (i.e., biclusters). It includes topic analysis of both documents and authors and computing a ranked list of *topic biclusters* (Fig. 1-a). These biclusters can then be visualized to support the interactive exploration of the search results. To demonstrate the usage of our analysis pipeline, we design a visual layout inspired by BiDots [34] (Fig. 1-b), which has shown effectiveness in exploring many biclusters simultaneously. Different from BiDots, we compute and display topic keywords in the middle of each topic bicluster in a row. The topic biclusters are ordered vertically based on their rank. In a row, entities of a bicluster are ordered based on their similarity to the topic keywords. Moreover, we generate thumbnail images to represent documents/authors. In contrast to BiDots that focuses on the visualization, we focus on the analysis of document search results with a novel approach based on the concepts of two-mode networks and topic modeling.

For evaluation, we examined the spatial distribution of the relevant documents of sample queries on the NIPS publication dataset [24] in our 2D layout, showing its effectiveness compared to a classic 1D list layout. We also conducted an initial user study with 12 participants to compare our proposed interface, a simplified variation without topics, and a text-based interface. The results show that participants found the topic words useful and the proposed interface attractive, were more satisfied with the 2D layout, and made decisions with lower cognitive load.

In summary, this paper highlights the following key contributions:

- A analysis pipeline, which enables integrating the use of topic analysis to organize documents and authors into two-mode networks for supporting sensemaking of search results.
- A visual method enhancing BiDots for meaningfully displaying and organizing computed biclusters, which enriches the design space of bicluster visualizations.
- Evaluations to illustrate how our proposed topic-based two-mode networks support exploring document search results.

2 RELATED WORK

Many methods have been developed to compute two-mode networks. A majority of methods is based on mining biclusters in data [29, 32]; in our case, a closed bicluster is a set of documents with the same authors. Co-clustering [1, 8] performs clustering on two aspects simultaneously, but it is less flexible than soft groupings [26] that allow a document/author to be in multiple biclusters. Finally, topic modeling [3], employed in our approach, has been validated for corpus exploration and information retrieval [4, 10]. It finds soft biclusters along with topic words describing the biclusters.

For visualizing two-mode networks, entity-driven designs (e.g., Jigsaw [25] and parallel node-link bands [12]) allow users to select nodes and show links connecting nodes. They do not explicitly visualize the biclusters but rely on user interactions. Edge bundling [27] or ordering can somehow address this issue, but finding an optimal order for reducing visual clutter is not easy [28]. Relationship-driven designs (e.g., BiVoc [15] and Bicluster viewer [16]) display data in matrices with relationships shown as matrix cells. One limitation is

*Author e-mails: jianzhao@uwaterloo.ca (corresponding author), smaoyuan@niu.edu, patrick_chiu@acm.org, francine@acm.org, beeyliew@gmail.com.

Figure 1: (a) Analysis pipeline for document search results using topic-based two-mode networks. (b) Visualization of two-mode networks between documents and authors. Each row corresponds to a topic, sorted by relevancy; the most relevant authors and documents are closest to the center.

that all biclusters cannot be revealed completely without constantly reordering rows and columns or duplicating some of them. Cluster-driven designs aim to be more scalable for showing many biclusters. For example, Bixplorer [11] represents each bicluster as a matrix and connects the same entities with links. BiDots [34] displays the nodes of each bicluster in a row with the links in the middle. While these techniques reveal biclusters with different strategies, none of them consider the organization of entities within biclusters. Moreover, they require user checking entities to understand biclusters without offering any higher-level summary. Due to the two drawbacks, for sensemaking of biclusters-encoded search results between authors and documents, they can only provide limited support.

Our work is also related to visualization of topic models and documents. PaperLens [18] visualizes topic trends in research papers using bar charts. Termite [7] provides a table visualization of terms and topics to help assess topic model quality. Node-link diagram is another major visualization for presenting topics, documents, and their relationships, such as TopicNets [14] and Literature Explorer [30]. Collaboration Map [5] uses a node-link graph where topics are represented by persistent fixed nodes and authors are represented by transient movable nodes. Some tools also integrate citation networks with topics to support literature search [21]. Moreover, ParallelTopics [10] employs parallel coordinates to present the topical probabilities of documents. To support faceted exploration, PivotPaths [9] uses a tri-partite graph to navigate authors, papers, and keywords, exposing faceted relations as visual paths.

In document searching, an energy based layout is applied for displaying search results with text snippets based on content proximity [13]. Topic-Relevance Map [22] uses a radial layout to visualize a topical overview of the search result space as keywords with respect to relevance (radius) and topical similarity (angle).

In contrast to these approaches, we use topic modeling to cluster search results and compute two-mode networks between authors and documents based on topics. The proposed approach mines relationships of topics, authors, and documents based on biclusters, which supports answering the know-what and know-who questions. Inspired by BiDots [34], computed biclusters are visualized in a row, and their entities are horizontally arranged based on their similarity to computed topics that are placed in the center.

3 Analyzing and Visualizing Search Results

We use two corpora for preparing document datasets used for development and evaluation. One is the NIPS conference papers from 1987-1999, with about 1700 papers and 2200 authors [24]. The other contains publications from an industry research lab, with 628 documents and 424 authors from 1995–2019. A document dataset must include metadata for each document's title and authors, as well as the text for topic analysis. We can use the title along with either the abstract, or, if available, the document's content text. We extracted the text by using software tools on the PDF files, or by scanning and optical character recognition (OCR). Preprocessing of the text include removing stop words, converting plural to singular, and filtering out words that are infrequent (in less than 5 documents) and too frequent (in more than 50% of the documents). We also created a thumbnail image for each document by applying a picture detection method [6] on page images.

3.1 Analyzing Search Results

We propose an approach for analyzing search results of documents by computing two-mode networks based on topic analysis, as shown in Fig. 1-a. The parts (in blue) on the left side can be pre-computed or periodically computed as the dataset is updated. Alternatively, the topic analysis can be performed after each query on the retrieved results. While this requires more computation, the cohesiveness of the documents to each other and the topics may be higher, resulting in more intuitive document groupings.

Topic analysis (Fig. 1-a1) is performed using LDA topic modeling [3], where each topic (Fig. 1-a2) is represented by a set of terms and their associated probabilities. In this paper, we set the number of topics to 20. For each topic t_i of a document d_j, we compute a topic similarity score (Fig. 1-a3) based on matching the topic terms W_j against the document text:

$$sim(d_j,t_i) = \sum_{w \in W_j} p(w|d_j)p(w|t_i). \quad (1)$$

All authors associated with a topic are ranked by their similarity to and prolificity of documents on the topic, as well as authorship order. Specifically, for each author a_k, we compute a topic similarity score (Fig. 1-a4) for each topic t_i by taking a weighted sum of $sim(d_j,t_i)$ over the set of documents by that author. The weights should factor in the author's position and the number of documents by the author on that topic, with greater weight given to authors with an earlier position and authors with more documents on the topic. We use discounted cumulative gain to compute the weights: $w_d = \log_2(n+1)$, where n is the author's position in document d. Thus, the author ordering is computed by ranking the author score:

$$score(a_k) = \sum_{d \in D(a_k)} w_d \cdot sim(d,t_i), \quad (2)$$

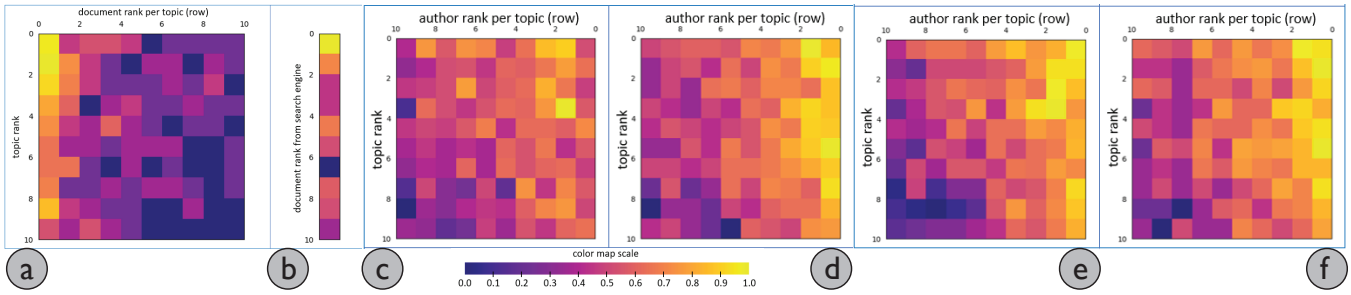

Figure 2: (a) 2D heatmap of the relevant documents in the layout for our visualization (right side of Fig. 1-b with left-to-right ordering). (b) 1D heatmap of the relevant documents in a layout for a ranked list view. (c-f) 2D heatmaps of the authors of interest in the layout for our visualization (left side of Fig. 1-b with right-to-left ordering): (c) lead authors with simple topic scores, (d) prolific authors with simple topic scores, (e) lead authors with DCG topic scores, and (f) prolific authors with DCG scores. Heatmaps are in logarithmic scale.

where $D(a_k)$ are documents by author a_k. For domains where the authors are listed alphabetically, the weights can be set to 1.0.

To generate the ranked list of relevant topics, we compute a topic rank score (Fig. 1-a7): for each topic t_i, we take the maximum search rank score of the relevant documents $\{d_j\}$. Other options for the topic rank score are taking the average of the search rank scores of d_j, or taking the average of topic similarity scores s_{ji} with respect to d_j. Then, the topics are sorted by these topic rank scores. For each topic t_i, a two-mode network (Fig. 1-a8) is formed by taking the most relevant documents D_i and authors A_i with respect to that topic. Specifically, from the top-k (e.g., $k = 100$) documents returned by the search engine, we take D_i to be a subset of these with topic similarity scores s_{ji} greater than epsilon (e.g., 0.0001) for each topic t_i, and take A_i to be the authors of D_i. An optional step is to prune these sets so that there are no isolated nodes on each row.

We employ Lucene [19] to handle the search query (Fig. 1-a5), which returns a ranked list of relevant document metadata items and their search rank scores (Fig. 1-a6). A threshold (set to 100) gives the number of top-ranked documents retrieved. The search result is processed as described above to obtain topic-based two-mode networks for the documents and for the authors.

3.2 Visualizing Topic-Based Two-Mode Networks

As shown in Fig. 1-b, each topic's two-mode network is displayed as a row, along with the top three words of the topic in the center column. Based on computed topic rank scores, two-mode networks are ordered in a top-down manner, with highest score on the top. Moreover, documents and authors are ordered by their topic similarity scores, with the higher scores nearer to the center column.

This design is inspired by BiDots [34], since it employs a cluster-driven design that promotes the topic-based two-mode networks in visualization, which is the key structure we want to present. This is unlike the traditional entity-driven (e.g., node-link diagrams, Jigsaw [25]) or relationship-driven approaches (e.g., adjacency matrices, BiVoc [15]), where nodes or links are the primary targets to visualize.

Moreover, this design fulfill the gap of existing bicluster visualizations from two key aspects. First, the center-placed topic words offers semantic summary of biclusters, which enables user quickly getting key information from biclusters without digging into entities. Second, it considers entity arrangement inside biclusters and uses spatial organization to reveal important entities. Specifically, the ordered representation of topics, documents and authors offers users an overview of the documents returned from the query with the most relevant topics and authors at the top and center of the visualization. Thus, they help to enrich the design space of bicluster visualizations.

To make the visualization more informative, we use thumbnail image tiles to represent the network nodes (i.e., documents and authors). A unique color is chosen for each author. The size of the thumbnail images decreases away from the center to convey that they are less relevant to the topic in the center. This allows us to place more items, however, it makes the text or image smaller on both sides. Thus, it is an open question to increase the scalability of the visualization on both aspects. Following BiDots [34], a number of user interactions are supported, such as showing a description of the document or the author when hovered over. See [34] for more advanced features.

4 Evaluation of Layout

We performed a quantitative evaluation to understand how the relevant documents and authors of interest are distributed in the 2D layout using our topic-based two-mode networks.

To check whether a document is actually relevant, we need a test set of queries and for each query the subset of documents labeled as relevant. One way to create a test set is to have humans label the relevant documents for each query; however, this requires substantial resources to produce. For the NIPS dataset, another way to check the relevance is using the subject index, which is included in the metadata. The subject index contains phrases and the page numbers of the associated papers. These phrases do not appear in the papers as metadata; the subject index is separated from the papers and not used for indexing in the search engine.

By using a phrase from the subject as a query, we assess which papers retrieved by the search engine are actually relevant (i.e., true positives): a paper is counted as *relevant* if it is associated to the query phrase from the subject index. We performed 100 queries sampled from the subject index phrases. For testing, we used a 10×10 2D layout, whereas in practice, the layout dimensions can change depending on the browser window size.

From the 100 queries, out of a total of 256 relevant papers, 172 were retrieved by the Lucene search engine (cutoff at top 100 items per query) and 67% of these 172 appear in the 2D layout: 243 papers with 127 duplicates that appear in multiple topic rows. The 2D heatmap of these relevant papers is shown in Fig. 2-a. As in Fig. 1-b, the topics computed to be most relevant should be at the top and the documents on a topic computed to be most relevant should be on the left. The heatmaps show that, indeed, the relevant items appear more frequently at the top-left and less frequently at the bottom-right, as one would expect. Additionally, there are relevant items in other columns (not the first column) across the layout, which indicates that that the 2D layout helps to access the various items.

We also created a heatmap for a 1D layout corresponding to a ranked list view (Fig. 2-b), which shows that the relevant items appear mostly as the top ranked item. However, only 39% of the 172 relevant documents appear in the 1D layout's top 10 items, whereas 80% appears in the top 50. This means that on a single web page, the 2D layout shows many more relevant documents than a ranked list view (67% vs. 39%), and the user would have to navigate through more than three pages to see 67% of the items.

For the authors, there are two kinds that are important: *lead authors* and *prolific authors*. A prolific author is defined to be an author with the most papers on the row and with more than one paper. Using the same 100 queries from above, we compared two methods

of computing the author's topic score: (1) a simple baseline method with weights set to 1.0, and (2) using discounted cumulative gain (DCG) described above. From the lead authors heatmaps (Fig. 2-c,e), we can see that using DCG pushes the lead authors toward the right edge (the center in the visualization) where the more important items should be placed. Using DCG does not degrade the heatmaps for the prolific authors as they stay near the right edge (Fig. 2-d,f).

5 Initial User Study

In addition to the quantitative evaluation, we conducted a controlled study to understand how users performed browsing tasks by comparing three interfaces *MM*, *MO*, and *TT*: MM is the proposed design showing *M*ultiple authors and *M*ultiple documents; MO is a simplified variation with each row showing *M*ultiple authors and *O*ne document, similar to a basic ranked list but with author information; and *TT* is a *T*ext-based list view of documents ordered by the search rank, akin to the Google Scholar webpage. See our supplementary materials for details of the study interfaces and settings.

5.1 Study Design

We used the dataset of the research lab's publications in this study. Unlike the NIPS dataset, it contains a broad range of subjects from many conferences and journals in computer science. From these publications, we generated content with 6 pre-defined queries "audio", "face", "video", "security", "web", and "sensor". The topic analysis was computed on the individual query results.

Participants and Apparatus We recruited 12 participants (9 males and 3 females) from the same research lab that produces the study dataset, with the criteria that they have little familiarity with the publications dataset. Participants completed the study in a quiet room. The study was performed on a desktop computer connected to two 24-inch monitors: one for the tool interface, and one for the study task workspace (e.g., note taking).

Design, Tasks, and Procedure We employed a within-subjects design. Each participant used all the interfaces in one sitting, and the order was counterbalanced across the participants. For each interface, participants were given an explanation and were allowed to try it out, and then were asked to perform two sets of tasks, each associated with one pre-generated query. Thus, each participant completed six sets of tasks in total for the three interfaces. The queries were randomly assigned without duplication.

For each interface, participants completed a set of two tasks. Task 1 was to explore the content in the interface and identify three research areas, with a limit of two minutes. Then, they selected one research area to continue to Task 2, which was to find three authors and three documents relevant to the selected research area, with a limit of three minutes. After performing the tasks, participants filled in an exit-questionnaire and the standard NASA TLX survey to collect their feedback, comments, and experience regarding to the interfaces. The study lasted about one hour for each participant.

5.2 Results and Discussion

Here we report our study results including task performance, questionnaire ratings, and participants' strategies.

Task Completion and Task Time. For the exploration task (Task 1), all the participants using MM completed the task within the 2-minute limit; whereas with the list based interfaces MO and TT, 1 (4.2%) and 3 (12.5%) of 12 participants, respectively, were not able to complete the task. One explanation is that they tended to read the content in the lists from top to bottom, and items in similar areas may be dispersed and take time to view; whereas with MM the items in similar areas are grouped by topic. For the finding items task (Task 2), the same number of participants (3 of 12, 12.5%) did not complete the task within the 3-minute limit for all three interfaces.

The completion times for both tasks are shown in Fig. 3-a. No significant effect is found across the three interfaces. The small

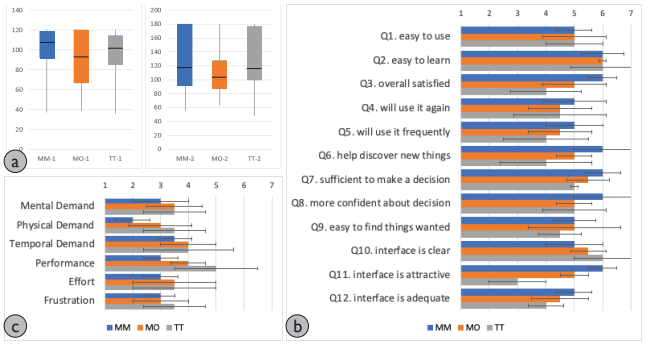

Figure 3: (a) Task completion times in box-and-whisker plots (the whiskers represent min and max values). (b) Medians of questionnaire ratings (higher value is better; error bars indicate IQRs). (c) Medians of NASA TLX results (lower value is better; error bars indicate IQRs).

number of instances where the time limit was hit does not affect the median or quartile values. MM and TT have similar median times, indicating that the new visualization did not slow participants down compared to text-baed interfaces. In general, MO seems a bit faster, which might be because it is similar to classic ranked lists (easier adaption for participants) while having author information (facilitating the exploration). However, there is a large amount of variation in the task times, and thus future studies are needed to provide in-depth understanding. But it worth noting that in the study MO and TT showed the ranked lists of documents roughly within one page, whereas in practice it may spread on multiple pages, which could result in worse performance. As the tasks are mostly exploratory and open-ended, the times are less of an indicator of the interface effectiveness.

Questionnaires Results. The results of the exit-questionnaire are shown in Fig. 3-b. The interfaces have the same median rating for Q1: easy to use and Q2: easy to learn. MM has the highest median rating in all the other questions, except for Q9: easy to find things wanted (MM is tied with MO, and better than TT) and for Q10: interface is clear (MM is the lowest). These results indicate that the proposed visualization, in general, was appreciated by the participants as useful. However, we note that the IQRs are overlapping for most of the questions and the sample size is not very large. For Q10, it is plausible that MM and MO were rated lower, because the interfaces were new to the participants, and they do not have any visualization background. The largest difference in rating is Q11: interface is attractive, where TT was substantially lower than MM and MO. Comments about the visual design of the interfaces noted that TT is *"not as engaging"* as MM, and that TT has too much text which made it *"difficult to focus where I am looking at."*

The result of the NASA TLX questionnaire are shown in Fig. 3-c. MM has the best median scores on all the questions, and MO seems to be in the middle (with some scores tied with MM or TT). This indicates that MM exhibited less cognitive demand for participants. However, the IRQs are overlapping, which means future studies are warranted. The biggest differences exist in physical demand and performance, where the median scores of MM are at least 1.0 lower compared to those of MO and TT.

Task Strategies. Participants reported different approaches to complete the tasks. For MM, 8 of 12 participants reported that they made use of the topic keywords at the beginning of a task to guide their exploration. For MO and TT, looking at the titles first was popular (9 participants), while looking at authors first was not common (2 participants). They tended to go through the list from top down (ordered by search rank), either by looking at the titles (5 participants), or by looking at the abstracts (6 participants). For deciding on which words to formulate their research areas, the reported strategies included highest ranked words (3 participants) and frequent words (1 participant).

6 Conclusion and Future Work

We presented a novel approach for analyzing search results of documents using two-mode networks and topic modeling, to address the know-who and know-what information seeking questions. We demonstrated this approach with publication datasets, and performed an evaluation of how the relevant items are distributed in a 2D layout. We further conducted a user study to investigate how participants used our proposed visualization based on BiDots [34], by comparing it with a simplified variation and with a text-based interface. The results indicate that the combination of icons and display of documents and authors grouped by topic biclusters provide a search interface that is viewed most positively and least taxing on users.

There exist several limitations that we want to address in future work. One direction is to conduct more user studies with realistic tasks over extended periods, as many of the study results need further verification and in-depth investigation. Second, we aim to apply topic models that are more directly based on both documents and authors (e.g., [23]), as well as to employ newer clustering models (e.g., [31]), which may have better information retrieval performance. A third direction is to address the scalability issue of the visualization based on BiDots [34], such as supporting collapsing and expanding of the items in a row or an entire row.

Acknowledgments

This research is supported in part by the NSERC Discvoery Grant and NSF Grant IIS-1850036 and IIS-2002082.

References

[1] M. Ailem, F. Role, and M. Nadif. Co-clustering document-term matrices by direct maximization of graph modularity. In *Proc. of the ACM International on Conference on Information and Knowledge Management*, page 1807–1810, 2015.

[2] F. Beck, S. Koch, and D. Weiskopf. Visual analysis and dissemination of scientific literature collections with survis. *IEEE Trans. on Visualization and Computer Graphics*, 22(1):180–189, 2016.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Proc. of the International Conference on Neural Information Processing Systems*, pages 288–296, 2009.

[5] F. Chen, P. Chiu, and S. Lim. Topic modeling of document metadata for visualizing collaborations over time. In *Proc. of the 21st International Conference on Intelligent User Interfaces*, pages 108–117, 2016.

[6] P. Chiu, F. Chen, and L. Denoue. Picture detection in document page images. In *Proc. of the 10th ACM Symposium on Document Engineering*, pages 211–214, 2010.

[7] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proc. of the International Working Conference on Advanced Visual Interfaces*, pages 74–77, 2012.

[8] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–274, 2001.

[9] M. Dork, N. Henry Riche, G. Ramos, and S. Dumais. Pivotpaths: Strolling through faceted information spaces. *IEEE Trans. on Visualization and Computer Graphics*, 18(12):2709–2718, 2012.

[10] W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In *Proc. of the IEEE Symposium on Visual Analytics Science and Technology*, pages 231–240, 2011.

[11] P. Fiaux, M. Sun, L. Bradel, C. North, N. Ramakrishnan, and A. Endert. Bixplorer: Visual analytics with biclusters. *Computer*, (8):90–94, 2013.

[12] S. Ghani, B. C. Kwon, S. Lee, J. S. Yi, and N. Elmqvist. Visual analytics for multimodal social network analysis: A design study with social scientists. *IEEE Trans. on Visualization and Computer Graphics*, 19(12):2032–2041, 2013.

[13] E. Gomez-Nieto, F. San Roman, P. Pagliosa, W. Casaca, E. Helou, M. de Oliveira, and L. Nonato. Similarity Preserving Snippet-Based Visualization of Web Search Results. *IEEE Trans. on Visualization and Computer Graphics*, 20(3):457–470, 2014.

[14] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth. Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Trans. on Intelligent Systems and Technology*, 3(2), 2012.

[15] G. A. Grothaus, A. Mufti, and T. Murali. Automatic layout and visualization of biclusters. *Algorithms for Molecular Biology*, 1(1):1–15, 2006.

[16] J. Heinrich, R. Seifert, M. Burch, and D. Weiskopf. *BiCluster Viewer: A Visualization Tool for Analyzing Gene Expression Data*, pages 641–652. 2011.

[17] P. E. Hulme. Bridging the knowing–doing gap: know-who, know-what, know-why, know-how and know-when. *Journal of Applied Ecology*, 51(5):1131–1136, 2014.

[18] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson. Understanding research trends in conferences using paperlens. In *Proc. of the CHI Extended Abstracts on Human Factors in Computing Systems*, page 1969–1972, 2005.

[19] Lucene. `https://lucene.apache.org`.

[20] I. V. Mechelen, H.-H. Bock, and P. D. Boeck. Two-mode clustering methods: a structured overview. *Statistical methods in medical research*, 13 5:363–94, 2004.

[21] R. Nakazawa, T. Itoh, and T. Saito. A visualization of research papers based on the topics and citation network. In *International Conference on Information Visualisation*, pages 283–289, 2015.

[22] J. Peltonen, K. Belorustceva, and T. Ruotsalo. Topic-relevance map: Visualization for improving search result comprehension. In *Proc. of the International Conference on Intelligent User Interfaces*, pages 611–622, 2017.

[23] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proc. of the Conference on Uncertainty in Artificial Intelligence*, pages 487–494, 2004.

[24] S. Roweis. Nips conference papers vols0-12 dataset. `https://cs.nyu.edu/~roweis/data.html`, 2002.

[25] J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008.

[26] M. Streit, S. Gratzl, M. Gillhofer, A. Mayr, A. Mitterecker, and S. Hochreiter. Furby: fuzzy force-directed bicluster visualization. *BMC bioinformatics*, 15(S6):S4, 2014.

[27] M. Sun, P. Mi, C. North, and N. Ramakrishnan. Biset: Semantic edge bundling with biclusters for sensemaking. *IEEE Trans. on visualization and computer graphics*, 22(1):310–319, 2015.

[28] M. Sun, J. Zhao, H. Wu, K. Luther, C. North, and N. Ramakrishnan. The effect of edge bundling and seriation on sensemaking of biclusters in bipartite graphs. *IEEE Trans. on Visualization and Computer Graphics*, 25(10):2983–2998, 2019.

[29] T. Uno, T. Asai, Y. Uchida, and H. Arimura. An efficient algorithm for enumerating closed patterns in trans. databases. In *International Conference on Discovery Science*, pages 16–31, 2004.

[30] S. Wu, Y. Zhao, F. Parvinzamir, N. T. Ersotelos, H. Wei, and F. Dong. Literature explorer: effective retrieval of scientific documents through nonparametric thematic topic detection. *The Visual Computer*, pages 1–18, 2019.

[31] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *Proc. of the International Conference on Machine Learning*, pages 478–487, 2016.

[32] M. J. Zaki and C.-J. Hsiao. Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Trans. on Knowledge and Data Engineering*, 17(4):462–478, 2005.

[33] J. Zhao, C. Bhatt, M. Cooper, and D. A. Shamma. Flexible learning with semantic visual exploration and sequence-based recommendation of mooc videos. In *Proc. of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–13, 2018.

[34] J. Zhao, M. Sun, F. Chen, and P. Chiu. Bidots: Visual exploration of weighted biclusters. *IEEE Trans. on Visualization and Computer Graphics*, 24(1):195–204, 2018.