Visual Analytics and Similarity Search - Interest-based Similarity Search in Scientific Data

Midhad Blazevic, Lennart B. Sina, Dirk Burkhardt, Melanie Siegel and Kawa Nazemi
Human-Computer Interaction and Visual Analytics
Department of Media & Department of Computer Science
Darmstadt University of Applied Sciences, Darmstadt, Germany
kawa.nazemi@h-da.de, midhad.blazevic@stud.h-da.de, lennart.b.sina@stud.h-da.de, dirk.burkhardt@h-da.de

2021 25th International Conference Information Visualisation (IV) | 978-1-6654-3827-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IV53921.2021.00041

***Abstract*—Visual Analytics enables solving complex analytical tasks by coupling interactive visualizations and machine learning approaches. Besides the analytical reasoning enabled through Visual Analytics, the exploration of data plays an essential role. The exploration process can be supported through similarity-based approaches that enable finding similar data to those annotated in the context of visual exploration. We propose in this paper a process of annotation in the context of exploration that leads to labeled vectors-of-interest and enables finding similar publications based on interest vectors. The generation and labeling of the interest vectors are performed automatically by the Visual Analytics system and lead to finding similar papers and categorizing the annotated papers. With this approach, we provide a categorized similarity search based on an automatically labeled interest matrix in Visual Analytics.**

***Index Terms*—Visual Analytics, Similarity, Collaborative Systems, Trend Analytics, Visual Business Analytics**

I. Introduction

Visual Analytics combines interactive visualizations with artificial intelligence and machine learning methods to enable solving complex analytical tasks [1], [2]. Beside analytical tasks that may lead to new insights or to detect unknown patterns in large amounts of data, the search and exploration processes can be supported through Visual Analytics in a proper manner [3], [4]. Exploration is closely related to search and leads to gaining knowledge [5]. As Marchionini, White and Munz describe, there is an interplay of lookup tasks and the exploratory search process, where humans use lookup tasks in order to accomplish exploratory searches. The critical difference is that exploratory searches require more advanced methods and technologies and are cognitively demanding as the user or explorer learns during the search and thus optimizes future search iterations with the knowledge gained [5], [33], [34] . While exploration is an important factor in Visual Analytics, the huge amount of data may lead to spend time without exploring the "right" or useful data.

Considering the exploration process in scientific publications, similar publications to those that the user is interested in, would be important. In particular those that are containing terms, which the user is not aware of. Those kinds of terms rarely given, due to many reasons. Scientists are pretty creative by introducing new terms, technologies, innovations and scientific enhancements lead in a natural way to that kind of terms, and in particular young scientists are not always aware of all related terms that should be investigated.

To face this problem, and enable exploration with the assistance of similarity-based measures, a Visual Analytics system should enable a kind of "annotation in context" [6]. This should allow the user to interact with a visual system and just annotate the "items-of-interest". In case of scientific publications, these are publications that are annotated during the exploration process and enable finding publications that are similar to the annotated ones.

We propose in this paper such a process of annotation in context of exploration that leads to labeled vectors-of-interest and enable finding similar publications based on the interest vectors. We first introduce some existing approaches and systems that make use of similarity and scoring. Our literature review investigates exemplary work in the last fifteen years that make use of different algorithms and matrices for similarity search. Thereafter we introduce our general approach that includes the entire steps from data extraction through visualization and visual annotation to the labeled similarity approach. Our main contribution is a labeled interest-matrix in Visual Analytics to find similar papers that are already categorized and labeled automatically.

II. Related Work

The approach presented in this paper uses topic extraction methods from text and measure the similarity of a set of documents in a "Visual Analytics" system to enable users getting similar document results visually represented. Similarity can have different "levels of granularity", e.g. "word-to-word similarity", "sentence-to-sentence similarity" or "document-to-document similarity" that can be combined as "word-to-sentence" or "sentence-to-document similarity" [7], [8]. Commonly, two vectors of words are considered when measuring the similarity. We introduce some works that present models to calculate similarity with various approaches.

Huang et al. [9] calculated text similarity using the key phrase vector model and combining two similarity measurements to create a final book similarity weight [9]. Furthermore, intending to use this for book recommendations in a recommendation system, they also applied a customer similarity

978-1-6654-3827-8/21/$31.00 ©2021 IEEE
DOI 10.1109/IV53921.2021.00041

measurement based on demographic data, and linked both book similarity to customer similarity layers [9].

Mahmood et al. [10] proposed to examine similarity and dissimilarity measurements to find document relatedness. They extracted keywords or terms from the contents of research papers such as the title and abstract, compared these between papers, and found and compared synonyms for each keyword [10]. Both similarity and dissimilarity measurements are then normalized and used to calculate the document-relatedness [10]. A disadvantage of this approach is the technique limitation as it is based on the word level, thus not taking sentences, etc., into consideration.

Mahmood et al. [11], in a later published case study, developed an Onthology named "Content based Onthology for Research Paper Similarity" (COReS) [11], which models different document similarity techniques to locate relationships between them and thus identify similarity between them [11]. This approach is interesting as it attempts to examine not only similarity based on one calculation, which might have a disadvantage in certain scenarios or domains, but instead considers many, and by doing so might overcome some disadvantages of certain calculations.

Lu et al. [12] described the use of "bibliographic coupling" and "co-citations" to calculate document similarity, more precisely, research paper similarity. Still, they stated that the disadvantage of this approach lies within the algorithm that would not be able to judge new papers correctly, as they have yet to be cited [12]. This is significant for scientific publications as the landscape is ever-changing and evolving daily due to new publications published or uploaded daily. To overcome this Lu et al. used the "common citation x inverse document frequency" (CCIDF) [12] proposed by Lawrence et al. [13], which is considered similar to "tf/idf".

Alschareef et al. [14] also examined citation network similarity and article similarity to create similarity matrices between articles, authors, and venues, which they calculated using cosine similarity [14]. Thus taking a citation-based approach and further developing it, so that it can also examine venues that might showcase similar work by creating a "venue-article latent preference matrix".

Heidarien and Dinneen [15] proposed a new hybrid geometric approach to measure the similarity levels among documents and document clustering that they named "TS-SS" [15]. The "TS-SS" algorithm attempts to calculate the similarity between two vectors based on cosine angle and euclidean distance, as well as the difference between their magnitudes [15]. Heirdarien and Dinneen create a "triangle's area similarity" (TS) based on the Euclidean distance of the vectors and then calculating the magnitude difference between the vectors as this component is needed to calculate the "sector's area similarity" (SS), along with the angular difference of the vectors [15]. TS is then multiplied with SS to create the "TS-SS" [15]. As stated by Heidarien and Dinneen cosine is not robust enough to distinguish similarity at high levels, and the euclidean distance could have problems when analyzing larger datasets than used in that paper [15].

Ng [20] examined the problem that users face when starting research. The difficulty that users have when formulating search queries [20]. Using a deep learning model that analyzes a research paper provided by the user, the model attempts to identify the category of the provided paper [20]. The model then presents results that are ranked with content similarity, peer reviews, the expertise of the authors, and the number of references of documents that are closely related to the provided paper with which the search was initiated [20]. Doc2Vec is used to generate vectors for the documents, in this case, the abstract and title [20]. SentiWord scores, h-index, and PageRank values are all taken into consideration with this approach [20]; thus, this model attempts to combine content similarity with other important factors to provide the researcher with optimal results. The disadvantage of this model lies within its similarity as it only examines the title and abstract. Thus important content information might get overlooked. This approach to calculate similarity based on title and abstract can also be found in Ristanti et al. [21] work, which used cosine similarity to calculate similarities between two documents for a classification task. This work described that the advantage in cosine lies in its ability not to be affected by the length of the document, but instead on a documents' terms and low error rates [21].

Niraula et al. [7] measured the accuracy, "F-measure" and "Kappa" for "semantic similarity and found out that for LDA, the "LDA-Greedy" provides the best results. The "cosine similarity" showed promising results, such as in the work of Sitikhu et al. [25], Thada and Jaglan [26] and in the case study of Soyusiawaty and Zakaria [27]. An interesting approach from Sitiku et al. and other studies involved using the "soft cosine measure" [25] which takes for example similar word meanings into consideration [28], [29]. As Rahutomo et al. have explained that this approach yields difficulties as it tends to yield lower similarity results based on semantic relationships on a dimensional level [29], which can be seen in the results of Sitikhu et al. [25] which lead us to prefer the standard "cosine similarity" approach for our current project. Further studies will however be made into "soft cosine similarity" due to the promising potential of a more semantic focused approach towards topic similarity.

In this section we introduced some works that make use of different similarity and scoring approaches for providing similar data entities. We focused in our investigation on those works that make use of more than two vectors for comparison. Our literature review illustrated that combining vectors as a matrix does not provide any labeled vectors of interest. By combining different vectors as a matrix, the labeling of such a matrix is important for gathering the context of search and exploration. We face this limitation in this work.

III. General Approach

Our goal is to enable an exploratory approach in Visual Analytics by providing users the ability to not only search for similar items based on one other item but also to enable the users to "bookmark" more items of interest and get similar

items based on the bookmarked ones. In our Visual Analytics system, we use scientific publications that are visualized in particular for detecting emerging trends [1]. While the primary focus lies on detecting emerging trends and upcoming technologies, the users are able to search and explore for publications. The following sections describe based on the model introduced in Figure 1.

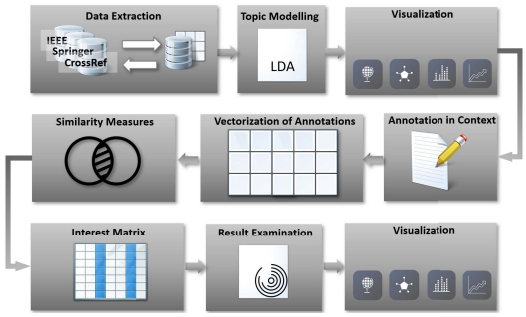

Fig. 1. The general approach with the steps of *Data Extraction*, *Topic Modelling*, *Visualization*, *Annotation in Context*, *Vectorization of Annotations* and *Similarity Measures* based on our previous works [1], [2].

A. Data Extraction

Our model starts with gathering scientific publication data from various resources. We use the application programming interfaces of "IEEE" in particular "IEEE Computer Society (Computer.org)", "Springer" and "Crossref" to gather data. The usage of different resources, in particular "Crossref" together with "Springer" and "IEEE" leads to many duplicates that are stored in different databases. By applying only the "Document Object Identifier", not all duplicates can identified, since many publications do not have a DOI. To identify the entire duplicates, we integrated the following steps that provided in a first examination of the extracted data no duplicates:

- use the DOI, if given, but compare the titles of the publications
- use the title of a publication
- if using the title, compare authors and years of the publication to get sure that these are the same papers

Our investigation of about 20 million publications showed that there are, beside duplicates, an unneglectable number of publications with the same authors (sometimes different order of authors), different titles and very similar content. We assume that these are reprints or extended versions of the papers. In case of reprints with exactly the same content, we categorize the according publications as duplicates. Therefore, the entire text is compared. The DOI is therewith not the best way to identify duplicates. Two exactly identical papers may have different DOIs, since they are published in two books. By including the title, authors and years, these reprints are identified as reprints and as duplicates.

In the first step of our model the entire data is stored in a data-base, including the metadata that provides information about authors, year of publication, publication type and further specific information, e.g. countries of the authors.

B. Topic Modelling

For extracting topics out of the full-texts and abstracts, we have compared the "Latent Semantics Indexing (LSI)" [30] and "Latent Dirichlet Allocation (LDA)" [23] with and without standard methods of "natural language processing (NLP)", e.g. tokenization and lemmatization [2]. Based on this investigation, we found out that LDA without lemmatization leads in two out of three cases to more accurate topics with a K of 20 [2]. We therefore applied the "Latent Dirichlet Allocation" [23]. Because of the high amount of documents, we tested different numbers for topics and got the best results with 500 topics in a dataset of about 20 million documents. We extracted beside unigram topics, also "n-grams" with two or more words that builds again a topic. So we had 500 words and 500 phrases (with two or more words) [1]. Every document was assigned with 20 topics, consisting of 20 words and 20 phrases. Figure 2 illustrates an example for the topic generation. Thereby the words and phrases for the topic "Visual Analytics" are visualized, whereas on the left side the unigrams are visualized with one word and on the right side the phrases. The generated n-grams provide a higher accuracy than the single-word topics, although their distribution percentage is lower.

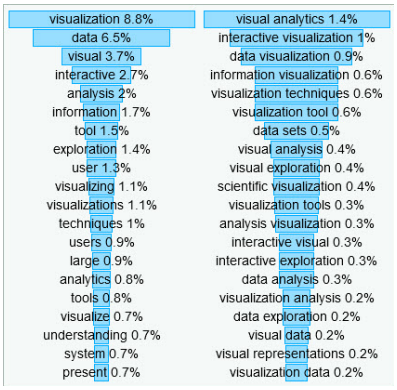

Fig. 2. "Generated single-word topics and n-gram-topics for a search on "visual analytics". [1].

C. Visualization

The visualization of the results is performed with two complementary approaches. We used the "visual information seeking mantra" proposed by Shneiderman [31] with "overview first", "zoom and filter", then "details-on-demand" [31, p. 337]. Beside the"visual information seeking mantra", we visualized the results of a search and used the approach of van Ham and Perer [32] of "search", "show context" and "expand on demand". Figure 4 illustrates the start screen,

Fig. 3. Visualization of the results on the query “visual analytics”. The user has chosen visualizations for the result-set.

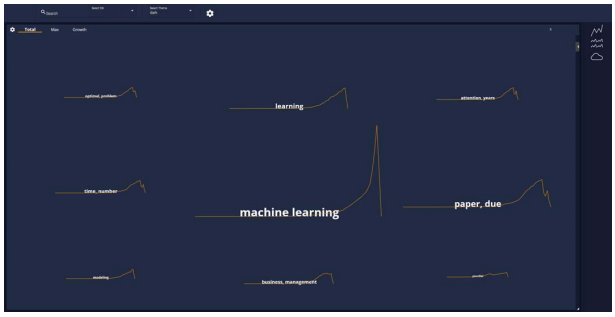

Fig. 4. Overview on emerging topics as starting visualization according to Shneiderman’s mantra [31]. The users are able to click to the emerging topics and get the results. The emerging topics are measured according to our previous work [1].

where the most emerging topics in a certain data base are visualized [1].

The users are able to search for queries or just click on an emerging topic or on a term that was searched very often. These searches are also stored in our data-base. The results can be visualized in different ways. An initial view provides the temporal spread of the result set, so that users can see the amount of publications on a certain query. Beside that users are able to place further visualizations in a web-based application or replace a single visualization. In Figure 3 the user has chosen beside the temporal over (top-left), a temporal spread of the extracted topics (phrases and words) on top-right, a simple graph-visualization for co-authored relations (bottom-left) and a spread topics by year (bottom-right), where the user has chosen the term “machine learning” in context of his search and can see immediately that the term was mentioned in more and more papers in the last three years.

Beside the visualization, three more interaction areas allow the refinement of the results, general settings and the choice of data-bases or allow choosing visualizations. On left a set of facets allow a refinement of search, e.g. years, topics, authors etc. On top general settings on the user interface can be adjusted, e.g. the color-scheme or data-base choice. In Figure 3 no data-base is chosen, so that an aggregated result-set of all data-bases is visualized. Users may choose just a single data-base. On the right bar, the users are able to choose a variety of visualizations, whereas only those are shown that are supported by the underlying data models.

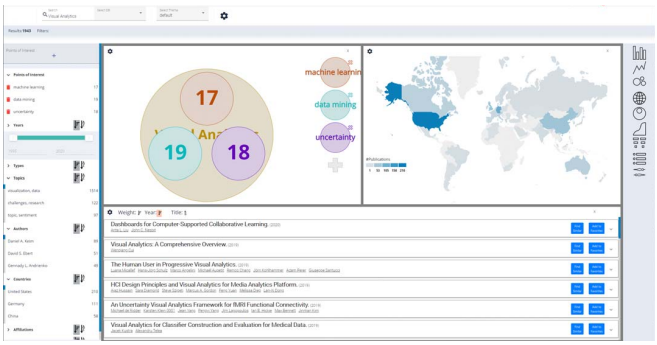

Fig. 5. Visual dashboard with a light-theme: illustrating the graphical search on top-left, a geographical visualization on top-right and a list view on bottom.

In Figure 5 the user has chosen other visualizations for the same result-set. Thereby the default-theme (light-mode) was chosen that makes use of another color-scheme. On top there are two visualizations that enable users to refine the search result. The visualization on top-left is a “graphical search” that allows the user to search within a search result set. In this case, the user chose the terms “machine learning”, “data mining” and “uncertainty”. The circle in the center shows the number of results. The visualization on the top-right illustrates the geographical spread of authors published in a certain field. So it can be seen through the saturation of the color that the United States has the most publications followed by Germany and China. The visualization on the bottom of the UI illustrates the refined search-results as an interactive list with an arrow on the right side that shows more information about a certain publication and two blue buttons, one for showing similar publications and one for adding a publication to “favorites”.

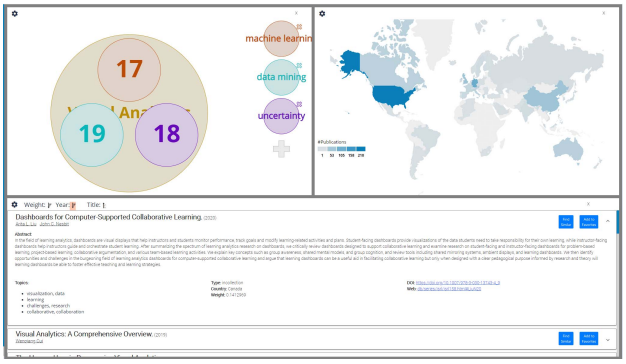

Fig. 6. Open the details of a single document that includes dashboard with a light-theme: illustrating the graphical search on top-left, a geographical visualization on top-right and a list view on bottom.

D. Annotation in Context

Seebacher et al. proposed a simple model for considering influencing factors in similarity search in “Visual Analytics” [6]. They stated that beside data that influence the similarity search, tasks and user should be considered. In particular in “Visual Analytics systems”, the tasks may be quite different [4], [5], [33]. Search task may have an exploratory character or just a lookup. A more precise and dedicated work in this context was provided by Munzner [34], who separated analytical and search tasks.

The aspect of annotating during exploration was mentioned by Seebacher et al. [6], but a real concept was not delivered. How can a “Visual Analytics system” support the annotation process in context of “visual exploration”, search or any other kind of analytical task? We have therefore included two different similarity searches during interacting with a visual system: a simple similarity search based on one publication and an enhanced annotation of publications of interest.

In Figure 6, we are just showing the visualization without the introduced interaction bars. Thereby the user clicks on the arrow to see the publication details that consists in this case

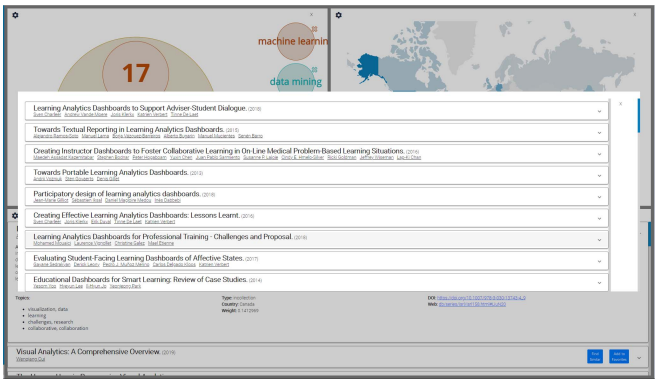

Fig. 7. Similar papers on a new window based on a single choice.

of title, year type of publication country of authors and four topics that was weighted highest amongst the extracted topics.

By clicking on the “find similar” button, similar publications are illustrated in a new window of the application and the user can start to explore the similar publications as illustrated in Figure 7. Thereby all topics of a certain publication that was generated before are used to identify similar papers.

Finding similar papers based on one document’s topic is pretty often performed in such visual applications. While exploring new topics and papers, the users should be enabled to explore and simultaneously annotate publications of interest. This leads to a real annotation in context, thus the user is still in the exploration process (context) that allows him/her to focus on his/her search and exploration. Our system enables users to “annotate” papers of interest during the interaction with the system. Figure 8 illustrates a part of the user interface with the list view. The user has annotated some publication with “add to favorites”. These elements are shown during the entire interaction with the “Visual Analytics system” in a different color and are added to the list of favorites.

Fig. 8. Annotation in context: users can add publication to favorites during interacting with the system.

E. Vectorization of Annotations

The list of annotated papers provides a set of topics. Each paper provides thereby a number of 20 topics that are stored as cookies, since we do not want to collect any kind of user information in our database. Beside the topics, a unique key of each publication is stored, thus we are running the “LDA-Algorithm” [23] every Saturday that may lead to other topics.

Thereby the topics of each paper are compared and identical topics are identified. In case of more than ten identical topics, a single interest vector is generated that contains all topics from both papers. In case of less than ten identical topics, a new vector is generated. So a "Vector of Interest" v_{I_j} may contain N topics as illustrated in Equation 1, whereas topic t_0 is the name of the vector.

$$v_{I_j} = (t_1, t_2, t_3, \cdots t_n) \tag{1}$$

F. Similarity Measures

We calculate the similarity between the vectors of interest v_{I_j} and all topic vectors v_i stored in our database through "cosine similarity" based on the works of Sitikhu et al. [25] described in Section II. We take t_0 of a vector v_{I_j} into consideration when examining the annotated topic vectors and calculating the similarity to all other topics that are stored as vectors v_i each with 20 topics. In this way, we are able to provide discovering topics that might have been yet unknown to the user even though they are of interest. Through generating the "interest matrix" (Section III-G) that helps to identify relevant publications. If the vectors are very different, the similarity will have a value close to -1, and we can conclude that the documents are not similar and can be considered uninteresting to the user, whereas high similarity results into a value close to 1. The calculated θ is illustrated in Equation 2 that illustrates the simple way of similarity measurement.

$$cos(\theta) = \frac{v_{I_j} \cdot v_i}{||v_{I_j}|| \cdot ||v_i||} \tag{2}$$

G. Interest Matrix

Based on the "vectors of interest" described in Section III-E and our similarity measures, an "Interest Matrix" can be generated that illustrates the "areas of interest" with all papers that were annotated and a function that illustrates all similar publications in each area of interest. In Figure 9 the user has annotated some papers during his exploration and interaction with the "Visual Analytics system". According to the describes vectors of interest, the areas are labeled with the first and highest weighted phrase. In this case the eight publications got two vectors, one for "Visual Analytics" and one for "Artificial Intelligence". As it can be seen not all publications have the label in the title but all the publications showed a significant similarity to each other, so that they could be labeled with the highest ranked topic.

IV. Conclusions

We introduced in this paper an approach for annotation in context of exploration that leads to labeled vectors-of-interest and enable finding similar publications based on the interest vectors. Thereby the problem of emerging new terms were investigated that may occur in research and scientific publications more than in other domain. Our approach generates labeled vectors-of-interest that leads to n-dimensional vectors

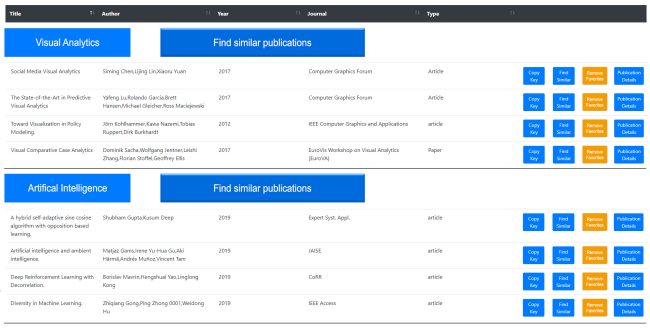

Fig. 9. Open the details of a single document that includes dashboard with a light-theme: illustrating the graphical search on top-left, a geographical visualization on top-right and a list view on bottom.

and finds based on each labeled vector similar publications by using the "cosine-similarity". The approach allows users to interact with a Visual Analytics system without being interrupted in the search and exploration process. The user is able to annotate publications-of-interest in context of visual exploration and find a list with categorized and labeled areas-of-interest. Based on these similarity is calculated that leads to similar paper without even containing exact matching. The vectors are containing topics generated through LDA. To enable a replication of our approach, we introduced each step based on a general model that illustrates the entire process beginning with the "data extraction", through "topic modeling", "visualization", "annotation in context", "vectorization of annotations", "similarity measures" to the "interest matrix", which is our main contribution.

V. Future Work

In the future, we will focus on evaluating the presented interest matrix to identify possible advantages, disadvantages, and use cases. Incorporating this matrix into a recommendation system is also of high interest as the previously mentioned studies attempted interesting approaches with similar goals. We believe that the interest matrix might be able to accomplish some of these goals more optimally. We will also examine the combination of our proposed interest matrix with other similarity measurements to create a detailed evaluation of our model. Furthermore, we will examine combining the interest matrix with other matrix approaches studies have found to yield interesting results, such as venues and citation networks. Although our current work focuses heavily on the interest matrix, we will also consider research paper credibility in future work because not only credible research papers are published but also less credible papers, which plays an important role when recommending papers. Our current approach and system is not evaluated yet with real users. We will evaluate the approach and the partially implemented system with users.

TABLE I
INTEREST MATRIX OF ANNOTED CONTENT

Interest Matrix					
Interest Term	**Topic 1**	**Topic 2**	**Topic 3**	**...**	**Topic N**
Visual Analytics	Interactive Visualization	Data Visualization	Information Visualization	...	Visualization Data
Artificial Intelligence	Neural Network	Machine Learning	Multi Agent	...	Cognitive Psychology

ACKNOWLEDGMENTS

This work was conducted within the research group on Human-Computer Interaction and Visual Analytics (https://vis.h-da.de).

REFERENCES

[1] K. Nazemi and D. Burkhardt, "Visual analytics for analyzing technological trends from text," in *2019 23rd International Conference Information Visualisation (IV)*, Jul. 2019, pp. 191–200, best Paper Award.

[2] K. Nazemi, M. J. Klepsch, D. Burkhardt, and L. Kaupp, "Comparison of full-text articles and abstracts for visual trend analytics through natural language processing," in *2020 24th International Conference Information Visualisation (IV)*. IEEE CPS, Sep. 2020, pp. 360–367.

[3] D. Keim and F. Kohlhammer, Jörn; Ellis Geoffrey; Mansmann, Eds., *Mastering the information age : solving problems with visual analytics*. Goslar : Eurographics Association, 2010.

[4] K. Nazemi, *Adaptive Semantics Visualization*, ser. Studies in Computational Intelligence 646. Springer International Publishing, Studies in Computational Intelligence 646, 2016. [Online]. Available: http://www.springer.com/de/book/9783319308159

[5] G. Marchionini, "Exploratory search: from finding to understanding," *Commun. ACM*, vol. 49, no. 4, pp. 41–46, Apr. 2006. [Online]. Available: http://doi.acm.org/10.1145/1121949.1121979

[6] D. Seebacher, J. Häußler, M. Stein, H. Janetzko, T. Schreck, and D. A. Keim, "Visual analytics and similarity search: Concepts and challenges for effective retrieval considering users, tasks, and data," in *Similarity Search and Applications*, C. Beecks, F. Borutta, P. Kröger, and T. Seidl, Eds. Cham: Springer International Publishing, 2017, pp. 324–332.

[7] N. Niraula, R. Banjade, D. Ştefănescu, and V. Rus, "Experiments with semantic similarity measures based on lda and lsa," in *Statistical Language and Speech Processing*, A.-H. Dediu, C. Martín-Vide, R. Mitkov, and B. Truthe, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 188–199.

[8] I. Dagan, L. Lee, and F. Pereira, "Similarity-based methods for word sense disambiguation," in *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain: Association for Computational Linguistics, Jul. 1997, pp. 56–63. [Online]. Available: https://www.aclweb.org/anthology/P97-1008

[9] Z. Huang, W. Chung, T.-H. Ong, and H.-c. Chen, "A graph-based recommender system for digital library," 01 2002, pp. 65–73.

[10] Q. Mahmood, M. A. Qadir, and M. T. Afzal, "Finding relatedness between research papers using similarity and dissimilarity scores," in *Web-Age Information Management*, F. Li, G. Li, S.-w. Hwang, B. Yao, and Z. Zhang, Eds. Cham: Springer International Publishing, 2014, pp. 707–710.

[11] ——, "Application of cores to compute research papers similarity," *IEEE Access*, vol. 5, pp. 26 124–26 134, 2017.

[12] W. Lu, J. Janssen, E. Milios, N. Japkowicz, and Y. Zhang, "Node similarity in the citation graph," *Knowl. Inf. Syst.*, vol. 11, pp. 105–129, 01 2007.

[13] S. Lawrence, C. Lee Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing," *Computer*, vol. 32, no. 6, pp. 67–71, 1999.

[14] A. Alshareef, M. Alhamid, and A. El Saddik, "Academic venue recommendations based on similarity learning of an extended nearby citation network," *IEEE Access*, vol. PP, pp. 1–1, 03 2019.

[15] A. Heidarian and M. J. Dinneen, "A hybrid geometric approach for measuring similarity level among documents and document clustering," in *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, 2016, pp. 142–151.

[16] D. Sarkar, "Text analytics with python : A practitioner's guide to natural language processing," Berkeley, CA, 2019.

[17] S. Vajjala, B. Majumder, and A. Gupta, *Practical Natural Language Processing*, 1st ed. O'Reilly UK Ltd, 2020.

[18] N. Thakur, D. Mehrotra, A. Bansal, and M. Bala, "Comparative analysis of ranking functions for retrieving information from medical repository," *Malaysian Journal of Computer Science*, vol. 32, no. 1, pp. 18–30, 2019.

[19] A. Singhal, "Modern information retrieval: A brief overview." *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001. [Online]. Available: http://dblp.uni-trier.de/db/journals/debu/debu24.html#Singhal01

[20] Y.-K. Ng, "Research paper recommendation based on content similarity, peer reviews, authority, and popularity," in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020, pp. 47–52.

[21] P. Ristanti, A. Wibawa, and U. Pujianto, "Cosine similarity for title and abstract of economic journal classification," 10 2019, pp. 123–127.

[22] B. Bengfort, "Applied text analysis with python : enabling language-aware data products with machine learning," Sebastopol, CA, 2018. [Online]. Available: https://learning.oreilly.com/library/view/-/9781491963036

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003. [Online]. Available: http://www.jmlr.org/papers/v3/blei03a.html

[24] T. Klove, T. Lin, S. Tsai, and W. Tzeng, "Permutation arrays under the chebyshev distance," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2611–2617, 2010.

[25] P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, "A comparison of semantic similarity methods for maximum human interpretability," in *2019 Artificial Intelligence for Transforming Business and Society (AITB)*. IEEE, nov 2019.

[26] V. Thada and D. Jaglan, "Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm," *International Journal of Innovations in Engineering and Technology*, vol. 2, pp. 202–205, 08 2013.

[27] D. Soyusiawaty and Y. Zakaria, "Book data content similarity detector with cosine similarity (case study on digilib.uad.ac.id)," in *2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 2018, pp. 1–6.

[28] G. Sidorov, A. Gelbukh, H. Gomez Adorno, and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model," *Computación y Sistemas*, vol. 18, 09 2014.

[29] F. Rahutomo, T. Kitasuka, and M. Aritsugi, "Semantic cosine similarity," in *The 7th International Student Conference on Advanced Science and Technology*, 10 2012.

[30] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[31] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *VL*, 1996, pp. 336–343.

[32] F. van Ham and A. Perer, "Search, show context, expand on demand: Supporting large graph exploration with degree-of-interest," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, pp. 953–690, 2009.

[33] R. W. White and R. A. Roth, *Exploratory Search: Beyond the Query-Response Paradigm*, ser. Synthesis Lectures on Information Concepts, Retrieval, and Services. G. Marchionini (Ed). Morgan & Claypool Publishers, 2009, vol. 1. [Online]. Available: http://dx.doi.org/10.2200/s00174ed1v01y200901icr003

[34] T. Munzner, *Visualization Analysis and Design*. Taylor & Francis Inc, Nov. 2014.