IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 18, 2019, accepted July 7, 2019, date of publication July 18, 2019, date of current version August 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929754

Cross-Domain Visual Exploration of Academic Corpora via the Latent Meaning of User-Authored Keywords

ALEJANDRO BENITO-SANTOS, (Member, IEEE), AND ROBERTO THERÓN SÁNCHEZ
Visual Analytics and Information Visualization Group, Department of Computer Science and Automation, University of Salamanca, 37002 Salamanca, Spain
Corresponding author: Alejandro Benito-Santos (abenito@usal.es)

This work was supported by the CHIST-ERA programme under national (MINECO Spain) Grant PCIN-2017-064.

ABSTRACT Nowadays, scholars dedicate a substantial amount of their work to the querying and browsing of increasingly large collections of research papers on the Internet. In parallel, the recent surge of novel interdisciplinary approaches in science requires scholars to acquire competencies in new fields for which they may lack the necessary vocabulary to formulate adequate queries. This problem, together with the issue of information overload, poses new challenges in the fields of natural language processing (NLP) and visualization design that call for a rapid response from the scientific community. In this respect, we report on a novel visualization scheme that enables the exploration of research paper collections via the analysis of semantic proximity relationships found in author-assigned keywords. Our proposal replaces traditional string queries with a bag-of-words (BoW) extracted from a user-generated auxiliary corpus that captures the intentionality of the research. Continuing along the lines established by other authors in the fields of literature-based discovery (LBD), NLP, and visual analytics (VA), we combine novel advances in the fields of NLP with visual network analysis techniques to offer scholars a perspective of the target corpus that better fits their research interests. To highlight the advantages of our proposal, we conduct two experiments employing a collection of visualization research papers and an auxiliary cross-domain BoW. Here, we showcase how our visualization can be used to maximize the effectiveness of a browsing session by enhancing the language acquisition task, which allows for effectively extracting knowledge that is in line with the users' previous expectations.

INDEX TERMS Academic corpora, digital humanities, document exploration, human-computer interaction, knowledge elicitation, latent semantic analysis, literature-based discovery, visualization.

I. INTRODUCTION

A. THE PROBLEM OF INFORMATION OVERLOAD

Recently, the adequate planning and scoping of research efforts has become a key task in academia. For this reason, scholars from all disciplines are spending more time seeking an adequate strategic position within a research body that allows them to develop their work according to practical societal needs and expectations. In this context, the use of electronic scientific databases has become a widespread practice among scholars worldwide. However, this task is becoming increasingly more difficult as databases increase in size. For this reason, efforts are currently being made within the scientific community to systematize and automate the production of literature reviews on practically the totality of scientific topics. The purpose of these kinds of publications is to collect and critically analyze multiple existing studies related to a given set of research questions to offer an exhaustive summary of the literature to the interested reader [1], [2]. The main reason for their popularity lies in their ability to provide scholars with the necessary foundations to start a new research endeavor, removing the need to perform a reading in full of the existing literature to gain insights into a given discipline. An essential step of literature reviews is the selection of sources that are obtained utilizing textual queries launched against an online database. An accepted common approach is to categorize and retain results that match specific inclusion criteria defined by the researcher. However, this procedure contains certain flaws that we identify at the beginning of our study and we aim to resolve. Firstly, while online search tools

The associate editor coordinating the review of this manuscript and approving it for publication was Chang Choi.

have been greatly enhanced in recent years and they generally succeed in the task of retrieving scientific publications from online sources, the usability of these tools in certain research contexts is still at stake due to the vast complexity and size of available collections, which may overwhelm the user. This problematic, known as *information overload*, is a long-standing issue in science that we describe here by quoting David M. Blei, one of the creators of the popular topic model latent Dirichlet allocation (LDA) [3]: "*As more information becomes available, it becomes more difficult to find and discover what we need.*" In relation to this matter, the task of fitting results retrieved from online search engines into a coherent picture is hard to achieve [4]. In our opinion, this unwanted behavior may be partially due to the extreme difficulty of expressing the nuances of the research aim in a textual query, a fact that limits the browsing experience to receiving a series of keyhole views of the subject under study that scholars are left to interpret.

B. LANGUAGE AND INTERDISCIPLINARITY

As a result of the increasing specialization in the sciences, many researchers have turned their attention to other disciplines, seeking help in solving research questions in a great variety of subjects. For these reasons, it has become more common to find multidisciplinary teams collaborating towards achieving the same research aim. Therefore, this particular configuration poses specific challenges that need to be addressed at all levels of collaboration. Within this collaboration, the use of language and the acquisition of communication skills has been identified as key in the development of interdisciplinary research. [5]. Therefore, this fact calls for the application of state-of-the-art linguistic models to: 1. enable meaningful interpretations of vast amounts of scientific literature at once, and 2. rapidly acquire domain-specific language that facilitates cross-domain communication between stakeholders. This problematic provides a conceptual framework for our work. Our method enables the extraction of relevant, non-obvious knowledge from a large document corpus through a high-level query expression (a bag-of-words [BoW]) that is supplied by an auxiliary or *query* corpus. In the context of interdisciplinary research, it aims at providing the user with a purposeful perspective of the target corpus that could be employed as a starting point in a hypothetical new research effort.

C. ANALYZING THE MEANING OF KEYWORDS

In order to provide a successful automatic implementation of the ABC model in the domain of computer science (CS), we rely on a semantic analysis of the author-assigned keywords in the collection. While probabilistic and predictive models, such as LDA or word2vec, have been successfully applied in the past to measure semantic document similarities through co-citation or co-authorship analyses [6], approaches that model the semantic space of author-assigned keywords are scarce in the current literature. Subsequently, centering the analysis task on author-assigned keywords presents its own challenges that differ from those related to other sorts of co-occurrence analyses that we aim to address in this research. For example, keywords are a very sparse feature of research papers, which implies that only a small portion of the phenomena is present in each observation. This particularity renders predictive semantic models inadequate in the context of *narrow-domain* research, in which the reduced size of available literature and the absence of a gold standard dataset may be limiting factors for the analysis. Particularly, highly sparse and small-sized corpora may produce overfitting issues that cannot be easily resolved by manual or automatic means [7]. Moreover, augmenting the size of the corpora could broaden the scope of the research topic too much in those contexts, risking the generation of relevant results.

While the sparsity could be partially addressed by performing an automatic keyword extraction based on the papers abstracts or full texts, in this study we employ author-assigned keywords as the main input for our analysis method because 1. we assume that they provide the best and most concise possible description of the contents of a paper that can be easily retrieved by automatic means from a majority of scientific publications and databases; 2. they effectively retain the original authors' intentionality because they are not constrained by any taxonomy imposed by publishers or other third-parties, which has an immediate positive impact in the acquisition of fine-grained, domain-specific language uses; and 3. author-assigned keywords do not introduce added complexity (i.e., preprocessing, cleaning, extraction, model validation) on the analysis task, which we felt could fall out of scope for a first approach to the problem. Regarding this matter, we refer the reader to Section VI, in which we discuss some future lines of work that aim to incorporate automatically generated keywords into our visualization scheme.

The main contributions of this paper are outlined hereafter: first, we propose a semantic analysis of author-assigned keywords found in the primary and auxiliary corpora to form a set of keyword vector representations from which we derive proximity data. Second, we provide a method to organize and visualize proximity data in such a manner that it enables a meaningful exploration of local structures found in the proximity data. Finally, we represent the original documents in the semantic space defined by the keywords, which has the positive effect of providing a close-loop view of the target collection to the user. This procedure is explained in this paper as follows: in Section II, we introduce relevant contributions that have inspired our work. Here, we also introduce latent semantic analysis (LSA), the distributional semantic model that we employed to generate a vector space model (VSM) of author-assigned keywords. Section III describes the auxiliary and main corpora that were used during our experiments. In IV, we describe the transformations and algorithms that were applied to the data in order to obtain a joint visualization of the keywords and document spaces, which is exemplified in Section V with two use-cases in the context of the interdisciplinary field of visualization in the digital humanities (DH).

Our contribution is completed by outlining known limitations of our method and future lines of work (Section VI) and, finally, by providing some conclusions in Section VII.

II. RELATED WORK

Our work is inspired by previous research in the areas of information science, NLP, interactive exploration of research paper collections and visualization of proximity data derived from LSA models. Below, we introduce a selection of past contributions in these areas that have greatly influenced the work presented in this paper.

A. LITERATURE-BASED DISCOVERY

At the beginning of our study, we identified literature-based discovery (LBD) as a potential solution to the problems of information overload and interdisciplinary vocabulary acquisition previously presented. LBD is a widespread knowledge extraction technique that was introduced in the 1980s by Don R. Swanson, an American information scientist who made important contributions in the biomedical domain. The main idea behind this form of discovery, namely the *ABC Model*, is not to generate new knowledge through laboratory experiments, but to seek to unveil existing connections in a body of literature that were previously unknown to the scientific community. The procedure employs a syllogism to identify potential knowledge associations in two disjoint bodies of scientific literature. Given two concepts A and C pertaining to the two bodies, respectively, the model finds that A and C are related if they both relate to another intermediate concept B. Swanson employed this simple technique to make several relevant medical discoveries, such as the effectiveness of fish oil as a treatment for Raynaud's disease (a circulatory disorder) [8], among others [9]. The ABC model supports two variants for *open* and *closed* discovery (Figure 1). In the open discovery mode, the process is started with an initial user-provided term to detect interesting term associations B and C and it is often employed to *generate hypotheses*. Conversely, in the closed variant the user initially defines two concepts, A and C, and the model reveals hidden associations (B-concepts). This second approach is generally used for *hypothesis testing and validation* [10], [11]. Our proposal aims to enhance the first variant of the ABC model and tries to go beyond typical co-word analysis by incorporating semantic analysis techniques. Throughout the rest of this paper, we will refer to A, B and C terms of query, link, and target, respectively.

While LBD was initially performed by manual means, different computational and semantic analysis techniques have been applied in the past to automate the process. Among these contributions, we highlight two that are specially relevant to this study: the works by Gordon and Dumais [12] and Cameron *et al.* [13]. In the first case, the authors employ LSA to drive the LBD process in a collection of Medline documents. In the second case, the authors make use of graph-based approaches to generate bridging or link terms under the close variant of LBD. In this contribution, we draw from

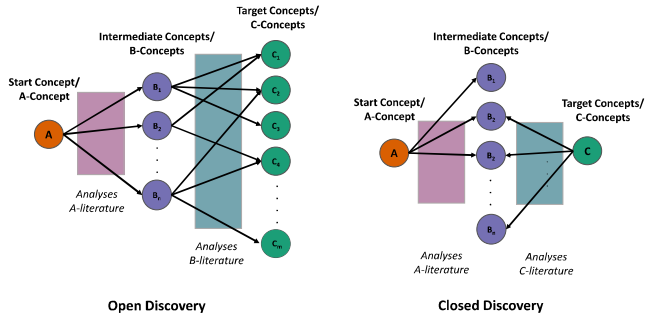


FIGURE 1. Open and closed discovery models in Swanson's ABC Model [8]. Our proposed visualization scheme enables automatic open LBD in narrow-domain research contexts. (Figure adapted from [11]).

similar graph filtering and representation techniques of proximity data (Section II-C) to propose a visually-enabled LBD in the CS realm, as opposed to a majority of past contributions that were limited to the biomedical domain. Furthermore, and in contrast to the works presented in this section, our work seeks to enhance the LBD process by proposing visualizations that assist the user in the task of jointly learning an embedding (Section IV).

B. VISUAL ANALYTICS OF SCIENTIFIC LITERATURE

Visual exploration of scientific literature collections is a topic that has been addressed extensively in the past by several different means, as the analysis of multivariate data is one of the most popular approaches taken by scholars in this field [1], [2]. Many of these contributions propose interaction techniques to filter, aggregate and browse a corpus of research papers employing derived metadata such as publication year, affiliation, authors and keywords, to name a few. In [14], [15] the authors propose VA systems to support and disseminate literature reviews. Beyond the display and filtering of metadata, the current literature has an abundance of examples of document exploration supported by network analysis techniques, which mainly rely on the construction of co-occurrence matrices from authorship [16], citation [17] and keyword data in the corpus. In the simplest cases, the exploration of the co-occurrence matrix is enabled by covariance studies [18] of the events in consideration with the goal of unveiling the underlying patterns of interest in the data. Whereas these kinds of statistical analyses may be useful enough to produce quantitative mappings and visual displays of scientific corpora, scholars must rely on ad-hoc interpretations of the results obtained, which may be prone to bias and error. This issue is usually addressed by more complex NLP techniques that facilitate the understanding of the underlying *semantics* of the collection. In this regard, CiteRivers [19] demonstrates the advantages of entropy analysis in the discovery of citation patterns. Similarly, [20] combines network analysis techniques with a textual importance index to produce dendrograms and graph visualizations of citation patterns. Metro Maps [4] measures the coherence and coverage of documents to produce visual summaries of query results in an online scientific database. One major drawback we detected in these proposals is that they rely on the usage

of a single text query to obtain their initial results. In our approach, this simple query string is replaced by an entire auxiliary corpus that is used as a complex query expression through which the target collection can be *seen*.

Continuing with the analysis of scientific literature via linguistic models, the surge of novel linguistic models such as LDA [3] or skip-gram negative sampling (SGNS) [21] has also had a profound impact on the design of visual document exploration tools. ParallelTopics [22] utilizes LDA to enable users to interactively explore a collection of research papers. Termite [23] allows the interactive refinement of topic models in a dataset comprising more than 14,000 publications. UTOPIAN [24] achieves similar results through non-negative matrix factorization (NMF) of keywords, documents, and topics, producing embeddings that are ultimately projected in a 2D space node-link diagrams. Notably, cite2vec [25] achieves a joint projection of keywords and documents by capturing citation contexts in word vector embeddings. Among these works, it is a common practice to employ dimensionality reduction techniques such as T-SNE to project the semantic high-dimensional space into the 2D plane, producing general perspectives of the dataset. Although T-SNE is able to preserve many interesting qualities of the semantic space, projecting the entire keyword space into the same display makes the appreciation of details in proximity data a harder task to achieve, even if the appropriate interaction techniques are correctly applied. Rather, our approach focuses on producing visualizations in which overlapping or redundant terms are removed while preserving interesting qualities of the topology of the semantic space that the user is interested in exploring. In this way, we focus on the display of local structures found in proximity data derived from the semantic space, which has a positive effect on the understanding of subtopics and other fine-grained information.

C. VISUALIZATION OF PROXIMITY DATA

The visualization of proximity data has also been addressed extensively in the literature. Worth noting is the graph-based psychometric scaling technique known as *pathfinder network scaling* [26]. *Pathfinder network scaling* aims to reveal structural patterns in proximity data by means of a graphical network representation known as pathfinder network (PFNET). PFNETs have been successfully employed in a great variety of contexts such as geoscience [27], biomedicine [28] or software engineering [29], to name a few. Other authors have found the adequacy of PFNETs to represent different cognitive structures and mental models to explain and enhance the learning process at undergraduate and expert levels [30]–[32]. The use of PFNETs to create visual science maps is also well documented in the literature. The majority of these studies rely on the construction of co-citation networks by different means that are ultimately visualized in a PFNET. The authors in [33], [34] combine co-citation and PFNETs to support the process of literature review with the aim of identifying new research opportunities. In a similar approach to ours, the authors in [35], [36] employ LSA and PFNETs to construct visualizations of academic corpora. PFNETs, however, focus on providing a general picture of the similarity matrix, producing large visualizations that may not be adequate to jointly explore keywords and documents as we propose in this research. Although we draw some concepts from PFNETs, such as the use of force-directed layout algorithms to visualize proximity data, our solution is specifically designed to resolve the challenges of interdisciplinary research by producing a coherent joint projection of keywords and documents found in local structures, rather than providing general overviews.

D. LATENT SEMANTIC ANALYSIS

In previous sections, we discussed some of the properties of author-assigned keywords and the reasons why we chose them as the basis for our study. Given the inadequacy of generative and predictive models, we selected LSA, a DSM, to define a semantic space of keywords. LSA is a theory of language and DSM that extracts and represents the contextual-usage meaning of words by applying statistical calculations to a corpus of text [37]. LSA (or Latent Semantic Indexing [LSI], as it is known in the information retrieval community) assumes that the occurring patterns of words in a variety of contexts are able to determine the degree of similarity among such words [38]. LSA is a fully unsupervised method that, unlike the case of predictive semantic models, does not employ any knowledge base or human-generated dictionary. Rather, it relies solely on the analysis of raw text. Because LSA originated in the psychology community, since its implementation it has been thoroughly evaluated to measure its accuracy in replicating human judgments of meaning similarity [39]. The similarity estimates derived by LSA are not based on simple contiguity frequencies or co-occurrence. Rather, they depend on a deeper statistical analysis that extracts the underlying semantics from a corpus. This kind of analysis has the positive effect of producing results that are conceptually similar in meaning to a given query term, even if these results do not share specific words with the search criteria. Beyond that, some authors have stressed the role of LSA as a fundamental computational theory of the acquisition and representation of knowledge that is closely related to the inductive property of learning, for which people seem to acquire much more knowledge than appears to be available from experience [40]. Although previous visualization schemes have been proposed to better understand LSA models [41], to the best of our knowledge, ours is the first to apply these techniques in combination with Swanson's ABC model introduced in previous sections.

1) SINGULAR VALUE DECOMPOSITION

To produce a semantic analysis of the words in a corpus, LSA makes use of a well-known linear algebra matrix decomposition method called singular value decomposition (SVD), which we briefly summarize for the reader hereafter: SVD is used to decompose a given matrix M into the product of three

matrices $U\Sigma V^T$, where U and V are orthonormal ($U^TU = V^TV = I$) and Σ is a diagonal matrix of sorted singular values of the same rank r as the input matrix. Let Σ_k, where $k < r$, be the diagonal formed by the k first singular values of Σ and let U_k and V_k be the matrices that result from keeping only the first k columns in U and V. The matrix $\hat{M} = U_k\Sigma_k V_k^T$ is the rank k matrix that minimizes the Frobenius norm between the input matrix M and any other rank-k matrix, that is $\hat{M} \in arg\,min\|M - \hat{M}\|_F$. Thus, the resulting matrix is the best k-dimensional approximation to the original in the least-squares sense (minimizing covariance). Lately, SVD has again gained interest in the NLP community due to recent studies [42] that prove that dense word vectors resulting from this factorization have similar properties to those obtained from the word embedding optimization of predictive models [21]. Furthermore, these vectors have proven to excel in word-similarity tasks while minimizing hyper-parameter tuning [7], [43], which is another controversial feature of predictive models [42].

III. DATASETS

Before we continue to explain our proposed visualization scheme, in this section we comment on two document collections that were employed during our experiments. In the first sections, we discussed some of the problems related to the selection of an appropriate query string during the extraction phase of mapping studies and literature reviews, which we aimed to leverage in this work. To this end, we replace this query string with a BoW obtained from author-assigned keywords in the auxiliary corpus. This first BoW represents the intentionality of the research; that is, it provides a high-level semantic expression that is representative of the kind of knowledge the researcher is interested in extracting from the target corpus. We construct this hypothetical situation in the context of two inherently interdisciplinary bodies of knowledge, the DH and visualization, which we introduce below.

A. QUERY CORPUS: DIGITAL HUMANITIES VISUALIZATION PAPERS

The DH are an interdisciplinary area of scholarship in which computational methods are applied in the resolution of research questions related to traditional humanities disciplines, such as history, philosophy, linguistics, literature, art, archaeology, music, cultural studies and social sciences. This process usually involves the "application of developed computational methods" [44] in a variety of fields of computer science, such as topic modeling, digital mapping, text mining, information retrieval, digital publishing or visualization, in "novel and unexpected ways" [44]. Particularly, in recent years visualization has become a hot topic in the DH as evidenced by the increasing number of visualization-related submissions to the annual DH conference. This surge has also had an impact on the visualization community, who have turned their attention to the DH as a vibrant new area of application for novel visualization techniques. An excellent example of this recent interest is the Workshop on Visualization for the DH (VIS4DH),[1] which has taken place as a parallel session to the IEEE Vis Conference since its first edition in 2016. One of the recurrent discussions of this workshop has orbited around the idea of how to produce significant visualization advances in the context of the DH. Whereas visualization techniques have been showcased in a large number of computing problems related to the humanities, some authors have warned of an increasing tendency in the DH visualization community to apply standard visualization techniques (such as force-directed graph layouts or word clouds) to the resolution of intrinsically distinct research questions. This tendency, as these authors note, might be impeding the production of valuable visualization research in the humanities [45], [46], therefore they stress the need to incorporate appropriate methodologies and evaluation techniques into the design process of the humanities.

According to the context presented in the previous paragraphs, the first dataset was constructed from metadata describing papers published in the DH conferences between years the years of 2015 and 2018 [47]–[49]. Given the broad range of themes present in this conference, we limited our search to papers that fell in the domain of visualization; that is, papers that contained the word "visualization" either on their title, subject or any of their keywords. We also completed this data with author keywords associations extracted from papers presented in the three editions of the Workshops on Visualization for the DH between years 2016 and 2018. This composition ensures that we have a varied and rich BoW to query a larger, general-purpose target corpus. The humanities-visualization dataset accounts for 257 documents, containing 728 unique keywords that appear a total of 1,131 times, which gives an average of 4.40 keywords per paper. In Figure 2, a histogram showing the frequency of the 20 most used keywords is presented.

B. TARGET CORPUS: DATA VISUALIZATION RESEARCH PAPERS

The second document collection is related to the general topic of visualization. Visualization is a major research theme in computer science that relates to the generation of graphics, diagrams, images and animations that help to enhance the comprehensibility of the underlying data and computational algorithms at play in a broad range of computer-related domains. For these reasons, visualization research papers provide a rich and varied set of keyword associations to explore and to connect to other different knowledge domains (e.g., the humanities). The dataset comprises meta-data from more than 3,000 research papers presented at the IEEE Visualization set of conferences: InfoVis, SciVis, VAST and Vis from 1990-2018 and it was recently compiled by a group of experts in visualization [50]. The dataset is publicly accessible[2] and actively maintained and updated by its authors.

[1] http://vis4dh.dbvis.de/
[2] https://vispubdata.org

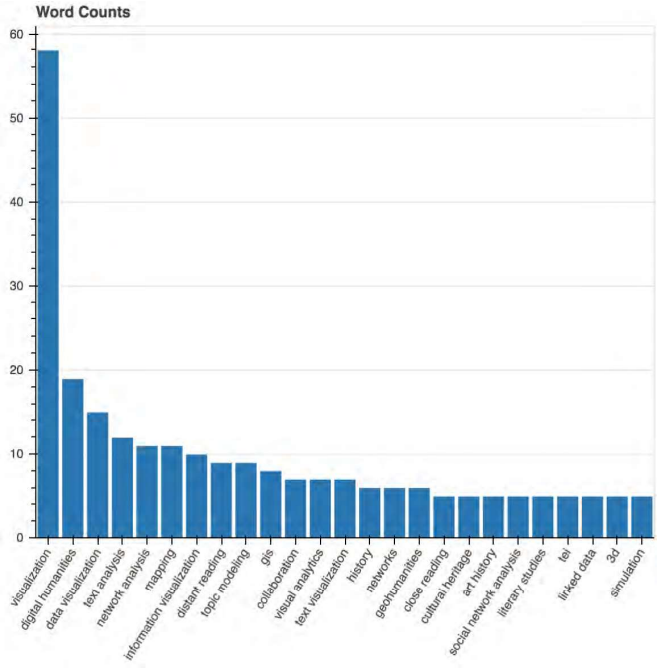


FIGURE 2. 20 most used author keywords in the query humanities-visualization dataset. Rank-based stop word detection is not trivial in this case given that some informative keywords (#4 "text analysis", #5 "network analysis") have higher ranks than some stop word candidates (#7 "information visualization" or #12 "visual analytics").

Data visualization research papers represent a rich corpus with multiple connections to other fields of modern science such as astronomy, sports, humanities, biology and machine learning, among others. To date, the dataset contains 3,102 research papers, of which 2,123 contain author keywords. The number of unique keywords in this dataset is 5,108, appearing a total of 9,877 times, which results in an average of 4.64 keywords per paper.

IV. METHOD

Our document exploration method comprises two main phases. The first involves all the necessary steps to generate a keyword-to-keyword similarity matrix from an LSA of the corpus. The second phase focuses on the querying, filtering and visualization of this similarity matrix. As we introduced in previous sections, our method aims to remove the need to provide a textual query to extract knowledge from a given target corpus C_t by relying instead on an auxiliary user-generated query corpus C_q. This distinction allows us to form two BoWs from keyword associations found in the query and target corpora, which are used as the two main inputs of our scheme. As we explain in Section V, the query corpus can be freely composed from the user's reference manager or from any other source she or he considers relevant to the study. Under this assumption, we expect the user to be familiar with the language of the query dataset whereas the target corpus is to be explored. At the end of the process, our method allows the user to query the target corpus by using keywords exclusive to the query corpus, effectively skipping the need for a language acquisition stage which may be highly time-consuming.

A. SIMILARITY MATRIX GENERATION

In this section, we provide the details on how our proposed method generates a distance matrix $\mathbf{D}$ from the two BoWs provided as inputs. The generation of this matrix relies on the LSA method, with some modifications that we introduce as follows: formally, we want to connect a query corpus $C_q = \{d_{q_1}, d_{q_2}, \ldots, d_{q_m}\}$ to a larger target corpus $C_t = \{d_{t_1}, d_{t_2}, \ldots, d_{t_n}\}$ with $n \gg m$. In our scheme, any given document is assumed to have a variable number j of author-assigned keywords $d_a = \{k_1, \ldots, k_j\}$

1) TOKENIZATION AND STEMMING

Prior to the application of the semantic model to our data, we perform tokenization and stemming on the author-assigned keywords. In the tokenization process, we split each multi-term keyword into its constituent parts, which are then stemmed and ultimately added to the BoW. Note that tokens appearing two or more times in the same document were counted as one. We noticed that, in our case, the inclusion of these two word pre-processing techniques was highly beneficial for the following reasons: the first and most obvious is that it provides an automated manner to match a high number of different linguistic keyword variations of the same concept (e.g., singular and plural), a circumstance that, unlike its occurrence in keyword taxonomies, can be observed in uncontrolled keywords due to their closer proximity to natural language. Second, it allows for the detection and subsequent removal of embedded stop words: i.e., words that do not carry any real meaning in the context of the collection and that might not appear on their own in the corpus. Take, for example, the multi-term keywords "*visual document analysis*" and "*visual citation analysis*". Although at a high level these two concepts are clearly related (because they represent two specializations of visual analysis), making a more clear distinction between them might not be immediately obvious if they are found in a corpus related to VA. In this case, the particles "visual" and "analysis" can be interpreted as noise because they do not add value to our understanding of the contents of the corpus. However, all three particles could carry important significance in other contexts.

The significance of a word can be generally explained by calculating the probabilities of seeing this word in the whole corpus: the less likely it is for a word to be seen, the more information can be assumed to carry. Therefore, in the multi-term keywords "*visual document analysis*" and "*visual citation analysis*," the discriminant terms are "document" and "citation" since it is less likely that they appear in the corpus. Without the tokenization and stemming of keywords, this fact could go unnoticed by the potential linguistic model to be applied at a later stage. In addition, the tokenization and stemming step effectively modifies the distributional model of all keywords over C. In our context, this had the following two positive impacts: first, it helped to reduce the sparsity

of keywords; second, the new distributional model of the keywords was better captured by LSA, which assumes a Gaussian distribution [51]. Although previous studies [18], [52] employ a power-law distribution to explain the phenomena of author-assigned keywords, recent studies also show this kind of distribution may be much rarer than initially thought [53]. For this reason, we identified that it is key to understand the particularities of the distributional model in order to propose a consistent analysis solution. In Figure 3, the *pre-* (top) and *post-* processing (bottom) distributional models are shown. We used the Python package "powerlaw" [54] to plot the complementary cumulative distribution function (CCDF) of the empirical keyword frequency data (black, solid), along with other fitted candidate distributions (dashed). In our example corpora, we could not find evidence that author-assigned visualization keywords follow a power-law distribution. Rather, we observed they could be better fitted to a Gaussian or an exponential distribution. According to these results, we decided not to base our method on the analysis of the first k-ranked keywords but employ other statistical artifacts such as LSA.

At the end of the processing step, the resulting tokens define a vocabulary V_g of size n_g that we split into three disjoint sets: V_q (query), V_t (target) and V_l (link), according to their provenance; that is, tokens in V_q, V_t and V_l can exclusively be found in C_q, C_t, or both, respectively, so that $V_g \doteq V_q \sqcup V_t \sqcup V_l$.

In our experiments, we performed a manual cleaning in which we removed obvious typographic errors and standardization of keywords; that is, the most common form of a keyword was preferred (e.g. "hci/human-computer-interaction" or "xai/explainable artificial intelligence"). Stemming was performed on the keywords using the Porter stemming algorithm [55]. Then, stems matching the expressions "visual," "digit," "human," "humanit," and "humanist" were discarded as they represent the global purpose of the study ("visualization" and "digital humanities"). After tokenization and stemming of keywords, we obtained 2,720 unique keywords that were distributed among the three considered vocabularies: query, link and target ($|V_q| = 257, |V_t| = 2143, |V_l| = 320$).

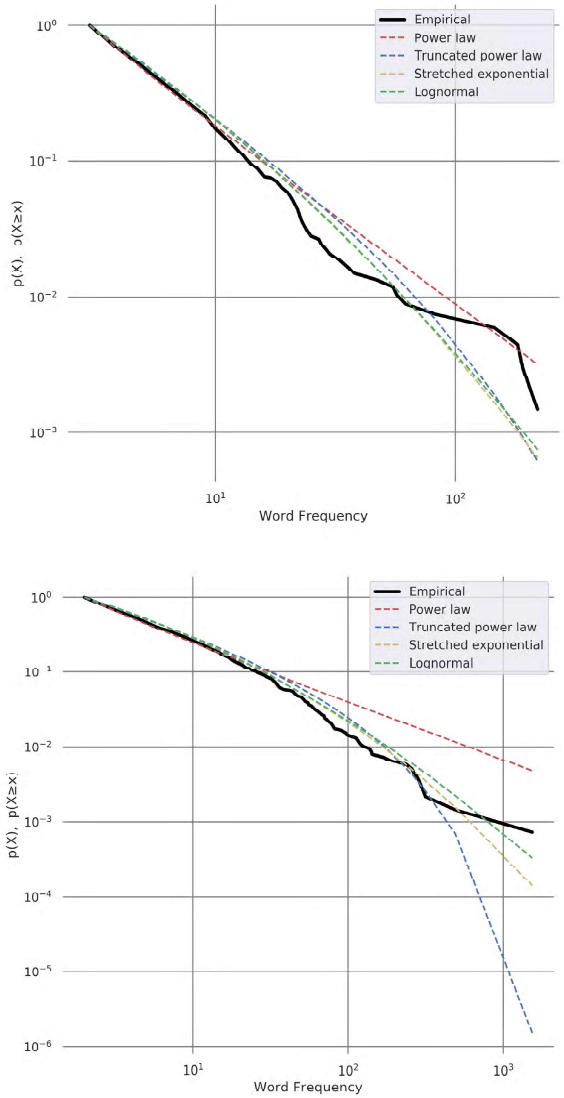


FIGURE 3. Pre (top) and post (bottom) stemming empirical (black) and theoretical (red: power law, blue: truncated power law, yellow: stretched exponential and green: lognormal) keyword frequency data CCDFs. Using the KS-test we could not find statistically significant evidence in any of the two cases that supported that keywords followed a power law, neither before ($p_\sigma = 0.054, gof = 0.0311$) nor after ($p = 0.0, gof : 0.0431$) tokenization/stemming. Moreover, we found evidence that these results could be best described with a stretched exponential, a lognormal distribution, or to a lesser extent, a truncated power law distribution.

2) POINTWISE INFORMATION MATRIX

In previous sections, we explained that LSA extracts latent semantics by factorizing a co-occurrence statistics matrix $\mathbf{M}$. This matrix can be built via different methods, such as term-frequency (TF) or term frequency-inverse document frequency (TF-IDF). In our case, we detected that narrow-domain corpora produce a great overlapping of insignificant words (noise) that we wanted to eliminate. To this end, we relied on a well-known metric of information science, pointwise mutual information (PMI) [56] because: 1. it provides an efficient manner to remove repetitive terms from the analysis and 2. when used in conjunction with LSA/SVD, it is capable of generating linguistic models that excel in distributional similarity tasks [43]. The usage of the smoothed PPMI matrix in LSA favors the detection of infrequent and informative relationships occurring in the high-dimensional semantic space over uninformative terms. This feature helps to provide a view of the target corpus that is based on the specifics of the user-generated query corpus and to identify keyword pairs that share a common latent meaning. PMI encodes the probability for a pair of tokens to be seen together in a document with respect to the probability of seeing those two same tokens in the whole corpus. This probability is defined as the log ratio between w and c's joint probability and the product of their marginal probabilities. These probabilities can be extracted empirically from the corpus by counting the number of times w and c

appear in the same document divided by the times they can be seen in other documents. In this paper, we do not consider the order in which the terms appear within a document and, therefore, the word-context matrix is built solely on co-occurrence. Similarly, the term-document matrix is a sparse binary matrix whose entries are defined as $B(t, d) = \{1 \text{ if } t \text{ occurs in } d \text{ or } 0 \text{ otherwise}\}$.

$$PMI(w, c) = \log \frac{\hat{P}(w, c)}{\hat{P}(w)\hat{P}(c)} = \log \frac{\#(w, c) \cdot |C_T|}{\#(w) \cdot \#(c)} \tag{1}$$

Following recommendations in the recent NLP literature [43], we employ a smoothed version of the PMI matrix. During our experiments, we found that setting the smoothing factor α to 0.95 yielded the best results in the similarity task, which is in line with observations from other studies [7].

$$SPMI(w, c) = \log \frac{\hat{P}(w, c)}{\hat{P}(w)\hat{P}_\alpha(c)} \tag{2}$$

where the smoothed unigram distribution of the context is:

$$\hat{P}_\alpha(c) = \frac{\#(c)^\alpha}{\sum_c \#(c)^\alpha} \tag{3}$$

The pairwise results are stored in a smoothed PMI matrix M^{SPMI} that matches the original dimensions of F, $|V_T| \times |V_T|$. A common problem with M^{SPMI} is that it contains entries of the form $\text{PMI}(w, c) = \log 0 = -\infty$ for word-context pairs that were never observed. This issue is solved in the NLP literature by using *positive* PMI (PPMI), in which the negative entries are replaced by 0:

$$M = SPPMI(w, c) = \begin{cases} SPMI(w, c) & \text{if } SPMI(w, c) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Once the keywords have been tokenized and stemmed, the next step of our method relies on counting the number of times each unique token appears in the query and target BoWs. Similarly, we calculate skipgram counts in order to measure the number of times two tokens can be seen together. The skipgrams count is employed to construct a $N \times N$ sparse matrix in which each cell represents the absolute count of observed associations between any two given tokens. At this stage, a binary term-document sparse matrix T is also created. This binary matrix is employed in the last step of the method to project the results onto a document space and produce a set of paper recommendations.

With vocabulary V_t, we build a square term-context frequency matrix $F \in \mathbb{R}^{|V_g| \times |V_g|}$ and a binary term-document matrix $B \in \mathbb{B}^{|V_g| \times |C_g|}$. The word-context frequency matrix captures how many times two terms appear together in the corpus. Following [42], this translates into $\#(w, c) \cdot |C_g|$. For example, if a document contains the following set of keywords: {*social*, *network*, *analysis*, *graphs*}, the context of *"social"* in this document is {*network*, *analysis*, *graphs*}. Finally, we retain the provenance of each token by indexing the square matrix $\mathbf{M}$ in the following manner:

$$M_i = \begin{cases} 0 \le i < |V_q| & \iff M_i \in V_q \\ |V_q| \le i < |V_q| + |V_l| & \iff M_i \in V_l \\ |V_q| + |V_l| \le i < |V_g| & \iff M_i \in V_t \end{cases} \tag{5}$$

3) LATENT SEMANTIC ANALYSIS

The next step we apply makes use of SVD to factorize the sparse matrix M^{PPMI}. This factorization produces dense vector representations of the keywords in our dataset and captures their latent meaning according to the principles explained in previous sections. Notice that in our case, the input matrix M is the symmetric matrix M^{SPPMI}, because $PMI(w_1, w_2) \equiv PMI(w_2, w_1)$ for any pair of tokens w_1 and w_2, which results in $M^{\text{PPMI}} \approx \hat{M}^{\text{PPMI}} = U_k \Sigma_k U_k^T$. Now, the rows of the resulting matrix U_k are the dense vector representations of all the keywords in vocabulary V_T.

Recent studies [51], [57] support that the selection of the number of singular values k in SVD has an important impact on the interpretability of the results: selecting too few dimensions hinders the extraction of meaningful patterns, while picking too many could reveal irrelevant connections, adding noise to the analysis process. During our experiments, we empirically determined that setting k to the minimum recommended (50) [51] rendered the best results, although we are aware that this parameter may vary in other datasets. In [51], the authors comment that "it has been conjectured that in many cases, such as language simulation, that the optimal dimensionality is intrinsic to the domain being simulated and thus must be empirically determined." Finally, we performed L2 normalization on the resulting word vectors for ease of use and performance optimization of the subsequent steps of our algorithm.

4) DISTANCE MATRIX FROM DENSE WORD VECTORS

One of the most popular (dis)similarity measures employed in NLP is the cosine of the angle formed by two word vectors [57]. This measure discards the length of the vectors and quantifies the difference in their direction in the multidimensional space. We selected this similarity measure because, as reported by other studies, it is adequate to represent cognitive similarity beyond simple linguistic similarity [57]. The formula of the cosine is well known and can be applied easily to the LSA vectors to build a distance matrix D:

$$D(x, y) = cos(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}} \tag{6}$$

Analogously, the similarity between two vectors can be expressed as:

$$S(x, y) = 1 - D(x, y) \tag{7}$$

As a final step, we employed the similarity matrix S to detect and merge synonyms (i.e., token pairs with $S(x, y) \approx 0$), which resulted in a reduction in vocabularies sizes ($|V_q| = 176$, $|V_t| = 1745$, $|V_l| = 320$).

B. ANALYZING INTER-GROUP SIMILARITIES

The second stage of our method focuses on exploring the similarity matrix S that was obtained in the last step. To overcome the conceptual distance between the query and target corpora, we look for structural patterns in the similarity relationships between keywords in the query vocabulary and those found exclusively in the target vocabulary. For this task, we rely on the construction of a complete graph G using the distance matrix D, which enables us to analyze the similarity between nodes (tokens) using different scaling techniques to reduce the complexity of the resulting graph. In order to map all tokens in V_t to their counterpart in V_q, we identify the shortest path that connects a token in V_t to any other token V_q. Formally, we can define the set of shortest paths P'_j from the token j in V_t to all tokens in V_q as the sequence of node pairs $(t^t_j, t^r_{k1}), (t^r_{k1}, t^r_{k2}), \ldots, (t^r_{kl}, t^q_i)$ with $r \in \{q, l, t\}$. Given that all pairs are edges representing distances, the sum of all distance pairs in a path in P' gives the total distance between the token t^t_j and every other token in V_q. Therefore, a shortest path P exists in P', connecting the node t^t_j to another node t^q_i that, by (7), yields a maximum similarity over all other alternative paths to tokens in V_q. Note that when $|P| = 1$, the similarity score *sim* is equal to the value of the similarity matrix S at $S(t_j, t_i)$.

$$sim(t^t_j, t^q_i) = 1 - \min_{P \in P'_j} \{ \sum_{k=1}^{l} dist(t_k, t_{k+1}) \mid (t_k, t_{k+1}) \in P \} \tag{8}$$

By (8), the path $P_{t^t_j}$ that maximizes the similarity score *sim* is a *significant* path of the target token t^t_j in G because it connects it to its most similar counterpart in V_q. These paths can be easily computed by a multi-source version of the Dijkstra algorithm.

After all shortest paths have been calculated, we can group similar nodes by the number of shared links in their respective paths from V_t to V_q. In this way, the sets of target nodes that present structural similarities in their relationship with the query dataset can be grouped together. This builds upon the idea that nodes related to the same topics are likely to share more links in the shortest paths that relate them to tokens in V_q, while the shortest paths of dissimilar nodes have few or no links in common. [58]. Particularly, the subgraph resulting from merging two or more shortest paths with common elements $P_1, P_2 \ldots P_n$ is a spanning tree of its nodes in G. This procedure is illustrated in Figure 4. On the left, two shortest paths for tokens t^t_1 and t^t_2 are shown. As $|V_t| \gg |V_q|$, some paths will share at least a common destination token in V_q, t^q_1 in the example. Input paths are ultimately merged into the same tree $T_{t^q_1}$.

After merging paths with common elements, we obtain a set of trees $T = \{T_1, T_2 \ldots T_i\}$ for each token $t^q_i \in V_q$ present in any path in P. Note that at this point not all tokens in V_q can be found in T, whereas all tokens in V_t are found. The solution to this issue is trivial and can be solved by adding a token t^j_q not present in T to the MST of its nearest neighbor, given that there are not any other shorter paths connecting t^j_q to any other token in V_t. At the end of this process, any tree, or a combination of trees in T, along with related documents, can be represented in a visualization according to the procedure outlined in the next sections. In Figure 5 we provide some of the paths obtained by this method in our experiments.

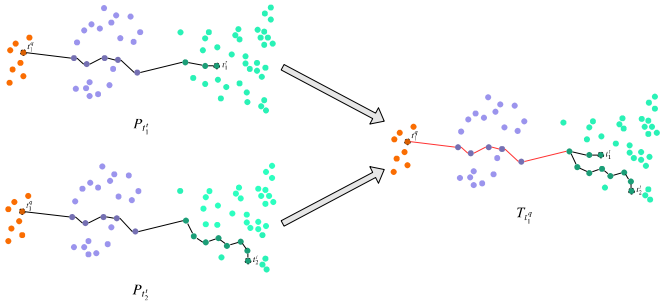


FIGURE 4. Shortest paths $P(t^1_t)$ (left, top) and $P(t^2_t)$ (left, bottom) connecting tokens t^1_t and t^2_t to their closest neighbor in V_q. The proposed method detects coincident tokens in the resulting paths and constructs the spanning tree that contains them. This results in a partition of the dataset in which tokens in V_t are grouped together if they relate to V_q in a similar manner.

During our experiments, we were able to generate paths for 138 distinct query tokens. On these paths, a total of 1,745 target tokens were represented, along with 85 other link tokens.

COMBINING SIGNIFICANT PATHS

Apart from the visualization of a single tree, our visualization scheme also supports the combination of two or more query terms to represent related keywords and documents. Given that by definition all trees in T are disjoint subgraphs of G, we can find an MST in G that contains all vertices in $T_1, T_2, \ldots T_n$ and which presents the minimum edit distance of all possible MSTs to the sum of all subgraphs. This reasoning is depicted in Figure 6, where we show the process of combining the tree of Figure 4, $T_{t^q_1}$. with another tree $T_{t^q_2}$. The tree resulting from the combination of the two paths has similar properties to any other tree in T and, thus, can be displayed in the same manner as we describe in the next section.

C. DOCUMENT EXPLORATION VIA KEYWORD PROXIMITY

In the last stage, the user is expected to provide a set of keywords to explore the collection. Following the reasoning explained in previous sections, the user employs keywords specific to the query vocabulary to obtain affine keywords and documents from the target corpus. These elements are presented to the user in a visualization that shows exploration paths related to the input query expression. The user is then able to progressively form a mental image of the target corpus by following these paths and optionally perform further research on the list of document suggestions that are displayed in the same visualization space. In this section, we comment on the necessary steps that were taken to produce this expected output.

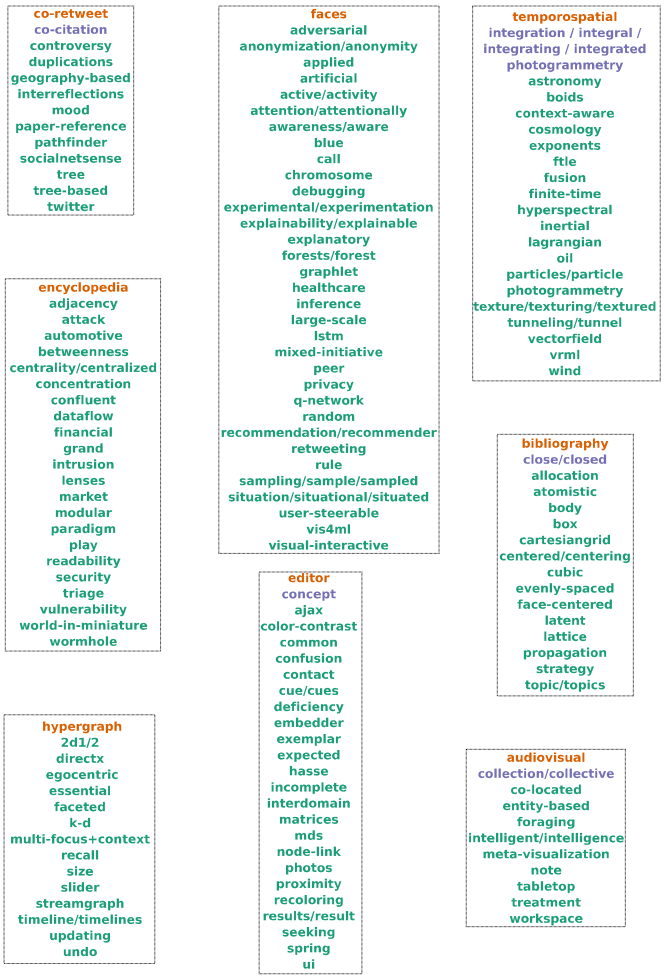


FIGURE 5. Keyword components (query, link, target) of some of the trees obtained by our method. Tokens were translated into their original keyword forms for clarity's sake. Each tree can be interpreted as a topic formed by a group of keywords that are highly related to the same element in V_q.

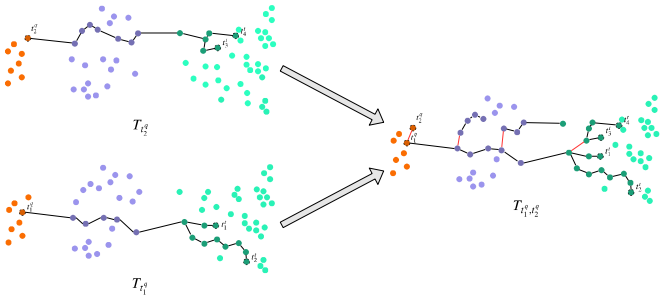


FIGURE 6. Paths for the user-provided terms $T_{t_1^q}$ and $T_{t_2^q}$ are combined into a new path that results from calculating the MST of nodes in the two paths. This procedure ensures that the two paths are presented in the most coherent possible way in the visualization.

The visualization employs a single tree as input, which can be one of the trees in T if only a single keyword is provided, or a tree resulting from combining two or more trees in T. The tree is drawn in the plane using the Kamada-Kawai layout algorithm [59], where tokens are depicted as vertices

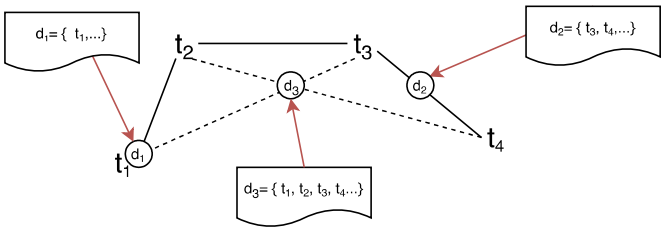


FIGURE 7. Documents are projected into the 2D representation of the semantic subspace defined by T. d_1 is projected to its only component in the subspace, t_1. Similarly, d_2 contains terms t_3 and t_4 of and therefore it is projected at the mid-distance of the link between the two terms. Finally, documents such as d_3 that contain three or more terms are projected at the centroid of the convex hull formed by the positions of such terms in the plane.

(text) and cosine distances as edges (solid lines) in the network. Query, link and target keywords are shown in orange, blue, and green, respectively. Tokens are translated into their original forms to ensure the readability of the results. In a subsequent step, the visualization is completed by representing documents into the semantic subspace defined by T. Firstly, the TD matrix is filtered to obtain documents that contain any of the terms in T. Note that each of the resulting documents may contain one or more terms (components) of the semantic subspace T. Then, the documents are projected according to their components' positions in the plane, as assigned by the Kamada-Kawai layout (see Figure 7).

Documents are represented as dots in the visualization and follow the same color scheme as keywords: documents in the query corpus are shown in orange, whereas those appearing in the target dataset are shown in green. Whenever two or more documents share the same position in the plane, they are aggregated in a visual encoding (the size of the circle). We represent the links between a document and their related components in the plane with a dashed line, which facilitates the task of identifying relationships between terms and documents.

V. EXPERIMENTS

In this section, we demonstrate the advantages of our method with two use-cases framed in the context of visualization in the DH. These experiments can be reproduced at the following location: https://doi.org/10.24433/CO.7350089.v1, whereas the code is publicly accessible at: https://github.com/ale0xb/keywords-vis.

A. DISTANT READING OF SHAKESPEARE'S PLAYS

In the first use case, we show how our visualization scheme can be used to relate theoretically distant subjects specific to the humanities to the subject of visualization. Concretely, we demonstrate how a scholar could extract knowledge from the target document collection using the query term "Shakespeare." We retrieve all the shortest paths ending in "Shakespeare " and plot them in the plane following the procedure explained in Section IV. The joint documents-terms visualization is shown in Figure 8.

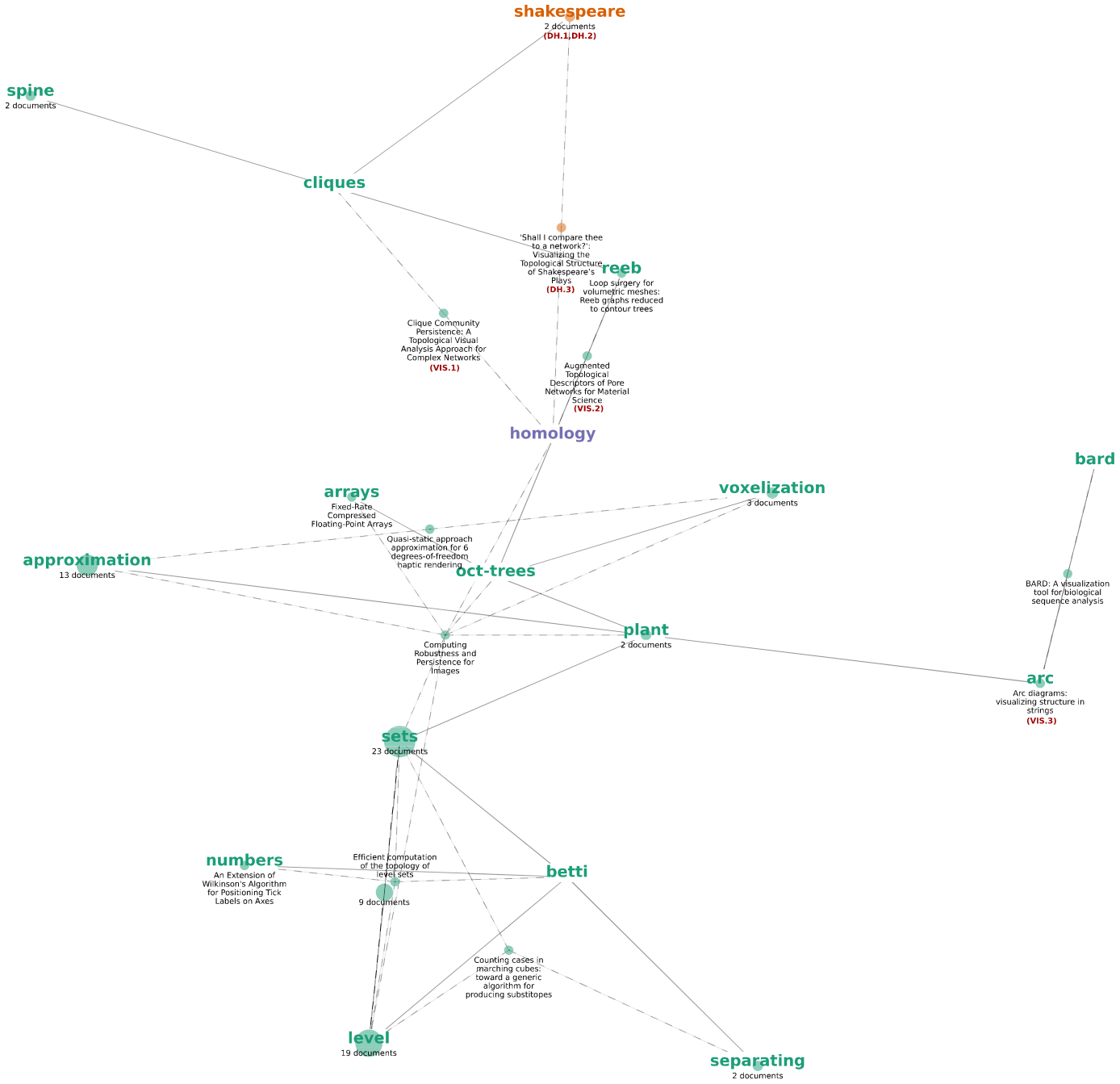


FIGURE 8. Visualization of the tree related to the query term "shakespeare" introducing at the top the concepts "persistent homology" and "topological data analysis".

The visualization is able to preserve similarities in the high-dimensional semantic space by placing nodes with high cosine similarity closer in the plane. The term "shakespeare" is placed at the top of the image. From a first impression, it can be observed that there are three documents (see Table 1) containing the term "shakespeare" in the DH corpus (shown in orange): two documents appear at the same position as "shakespeare," whereas the third one is shown closer to the link word "(persistent) homology" (in blue). Other vis-specific keywords (in green), such as "spine," "cliques," or "reeb," are drawn next to "shakespeare." These particles introduce the topic of *topological data analysis*, because document DH.3 includes the unexpected term "topology" among its keywords. On the contrary, the other two documents (DH.1, DH.2), which include the keyword "shakespeare", display general terms such as "networks," "exploratory" or "social" that do not generate high similarities in the semantic space and, therefore, these are not shown in the graph. Following the path formed by the terms "reeb" and "homology," the topic of "topological data analysis" specializes into "persistent homology," an algebraic method of discerning the topological

TABLE 1. Research papers commented in the description of the two use-cases presented in Section V (Experiments).

Use case	Collection	ID	Title	Keywords
1	DH	DH.1	Personae: A Character Visualisation Tool for Dramatic Texts	visualization, networks, drama, exploratory, shakespeare.
		DH.2	Analyzing Social Networks Of XML Plays: Exploring Shakespeare's Genres.	social networks,shakespeare, genre, drama, xml.
		DH.3	'Shall I compare thee to a network?': Visualizing the Topological Structure of Shakespeare's Plays.	visualization, shakespeare, social network analysis, topology, persistent homology.
	VIS	VIS.1	Clique Community Persistence: A Topological Visual Analysis Approach for Complex Networks.	persistent homology, topological persistence, cliques, complex networks, visual analysis.
		VIS.2	Augmented Topological Descriptors of Pore Networks for Material Science.	reeb graph, persistent homology, topological data analysis, geometric algorithms, segmentation, microscopy
		VIS.3	Arc Diagrams: Visualizing Structure in Strings.	string, sequence, visualization, arc diagram, music, text, code
2	DH	DH.4	Mapping Imagined and Experienced Places: An Exploration of the Geography of Willa Cather's Writing.	willa cather, mapping, gis, spatial turn
		DH.5	Monroe Work Today: Unearthing The Geography Of US Lynching Violence.	racial violence, lynching, gis
	VIS	VIS.4	Hotmap: Looking at Geographic Attention.	geographical visualization, gis, heatmap, server log analysis, online mapping systems, social navigation
		VIS.5	Semotus Visum: A Flexible Remove Visualization Framework	remote visualization, client server
		VIS.6	Dynamic Map Labeling	map labeling, dynamic maps, human-computer interface, label placement, label selection, label filtering, label consistency, computational cartography, gis, hci, realtime, preprocessing
		VIS.7	Spatial Text Visualization Using Automatic Typographic Maps	geovisualization, spatial data, text visualization, label placement
		VIS.8	Dynamic Visualization of Graphs with Extended Labels	graph label placement, dynamic animation, graph visualization, information visualization
		VIS.9	Particle-based labeling: Fast point-feature labeling without obscuring other visual features	interactive labeling, dynamic labeling, automatic label placement, occlusion-free, information visualization
		VIS.10	An Extension of WIlkinson's Algorithm for Positioning Tick Labels on Axes.	axis labeling, nice numbers

features of data, which is another interesting term as found by our model. Documents "Clique Community Persistence" (VIS.1) and "Augmented Topological Descriptors of Pore Networks" (VIS.2) treat this matter in the context of *graph cliques* and *reeb graphs*, respectively. Interestingly, it can be observed that document VIS.1 shares two common authors with document DH.3 (see the full dataset in supplementary materials).

In this case, LSA was able to detect the similarity in latent meaning between the terms "cliques" and "shakespeare" ($dist(shakespeare, cliques) = 0.1773$) by employing the unusual terms "homology" and "topology/topological." This first example shows the advantages of our proposal: The algorithm is able to detect the context of "shakespeare" (social network analysis) and extract relevant terms and documents that are presented in the visualization. In this way, the user can learn about community cliques and persistent homologies, which are statistically significant to the topic at hand. Although there are other documents with the keywords "social network" (7 hits) or "social network analysis" (2 hits) on the VIS collection, those are mostly related to different applications, such as the mapping of intellectual structures or visualization of online communities. Furthermore, none of these manual searches would have returned document VIS.1, although a close reading of this publication reveals that its background is "social network analysis," despite the authors not stating it in their selection of keywords for this document.

Continuing with other elements placed below "homology," we can identify documents and keywords related to "persistent homology" and the visualization of topologies in a variety of contexts. The informative term "oct-tree" (a hierarchical algorithm) is placed at the centre of the polygon formed by the terms "approximation, " "plants," "sets," "voxelization" and "arrays," For example, the paper "Computing Robustness and Persistence for Images" (VIS.2) informs on a visualization technique to depict the robustness of homology classes in 3D images of plant roots. Other documents, containing only one of the keywords in this polygon could be regarded as *complementary* readings to understand the central idea of the subtopic.

On the right side of the graph, it is worth noting the link connecting the terms "plant" and "arc" that introduce text visualization techniques that are also relevant to the topic of the analysis of dramatic texts. Despite relatively high distance of these two keywords to "shakespeare" ($dist(shakespeare, arc) = 0.6574$, $dist(shakespeare, bard) = 0.6773$), the design favors the inclusion of terms that produce documents relevant to the topic. In this case, the term "plant" provides a context to present *arc diagrams* (VIS.3), a popular network visualization technique to represent repetition patterns found in text

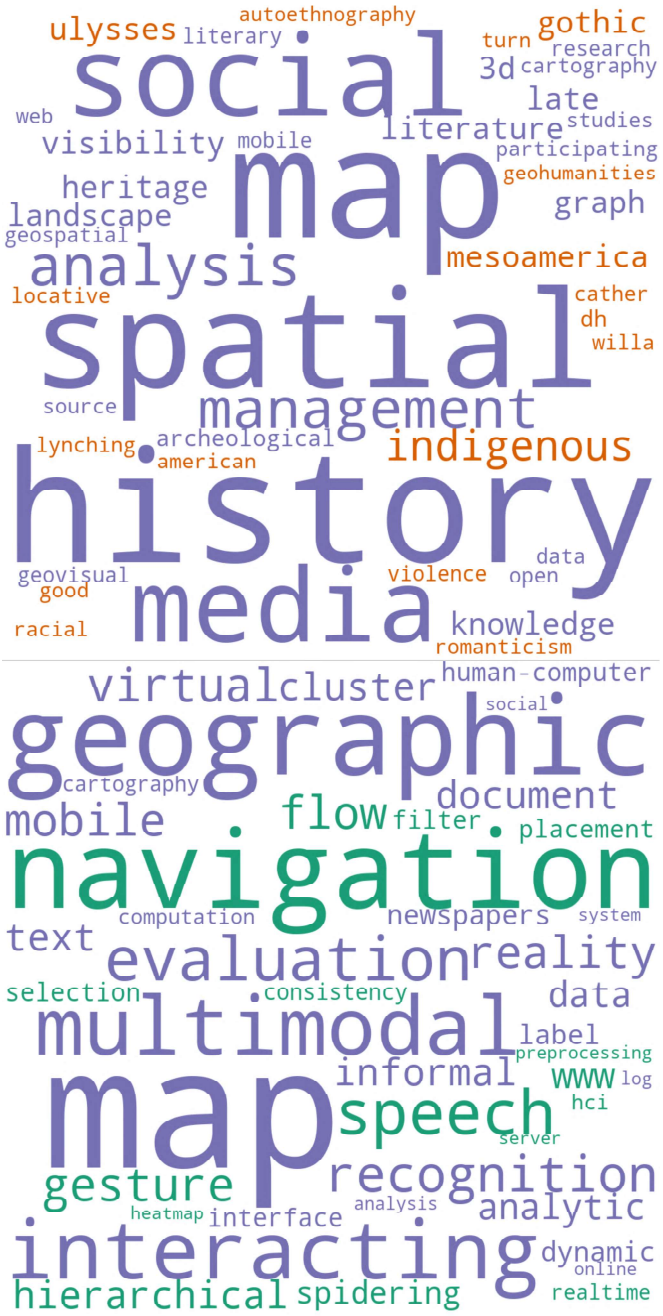


FIGURE 9. Word clouds showing the context of the link keyword "gis" in the query (top) and target (bottom) datasets. The SPPMI statistics matrix, in combination with LSA, is able to identify recurrent context terms such as "map " or "spatial," favoring the establishment of fine-grained affinities that are not built exclusively on first-order co-occurrence.

strings. As presented by the author in the original publication, a natural approach is to apply this technique to analyze DNA sequences (which explains its proximity to the term "plants"). However, arc diagrams are also highly related to the topic of text analysis in the DH: in his paper, the author demonstrates the capabilities of his proposal by visualizing musical compositions in a second use case. This finding reveals a technique that is related to the latent topics of text analysis and graph visualization. Therefore, it may be worth considering when designing a novel visualization in the context of the provided query term.

B. COMBINING SEARCH TERMS

In the second example, we demonstrate how different search terms can be combined in the same visualization to obtain a broader perspective of a given topic, in this case, GIScience in the humanities. To obtain the desired effects, we purposely choose two terms "willa" and "racial" to explore the VIS corpus. Both keywords appear once in two different publications related to the work of the American writer and Pulitzer winner Willa Cather (1873-1947) and of Monroe Work (1866-1945), an American sociologist famous for documenting lynching activity in the United States during the 19th and 20th centuries. The two contributions rely on the use of interactive maps and other GIS techniques to map the intellectual activity of the two individuals, a fact that the authors state in their keyword selection by including the keyword "gis" (see bottom of Table 1). This keyword appears in 10 and 5 publications in the DH and VIS corpora, respectively. In Figure 9, we depict the word cloud of the contexts of "gis" in both datasets.

The MST of members in the two paths of "willa" and "racial" is plotted in Figure 10. The resulting representation places the query terms close together at the center of the image. We can identify three main links departing from the nodes marked in orange, which lead to different subtopics that we discuss below: the shortest path of all displayed contains only one link (racial, server), and highlights two papers, VIS.4 and VIS.5, as VIS.4 is directly related to the general topic represented by the network while the other fits better as additional reading. In this case, the algorithm has detected a component related to web technologies in the latent meaning of "gis." This effect can also be observed in the word clouds of Figure 9, where we can find terms such as "web," "www," "log," "server" or " online." Among all these associations, "server" presents the closest cosine distance to "racial" ($dist(racial, server) = 0.2749$); thus, it is shown in the visualization. If we look at the upper part of the graph in Figure 10, it is worth noting the inclusion of the link keyword "labeling" (in blue), which generates interesting associations with other nodes in the graph. Next to the query node "racial" we find a document containing many of its nearest neighbors, "Dynamic Map Labeling" (VIS.6). This document is especially important since its verbose keyword description introduces specific subjects related to map labeling. Following other dashed links starting at the "labeling" node, we can observe this effect: the link (labeling, placement) produces two documents (VIS.7 and VIS.8) plus a third one (VIS.9), surging from the inclusion of the keyword "occlusion-free." In the same manner the pair "labeling, nice" generates a document (VIS.10) that, although it is not directly related to the topic of GIS, is deemed relevant because its contribution relates to the positioning of labels. Going up, the rest of the path introduces other aspects related to cartography, such as Mercator projections, digital and thematic maps and other specific techniques of interest as found by our method. In the lower side of the graph, the subtheme is related to the depiction of statistical significance

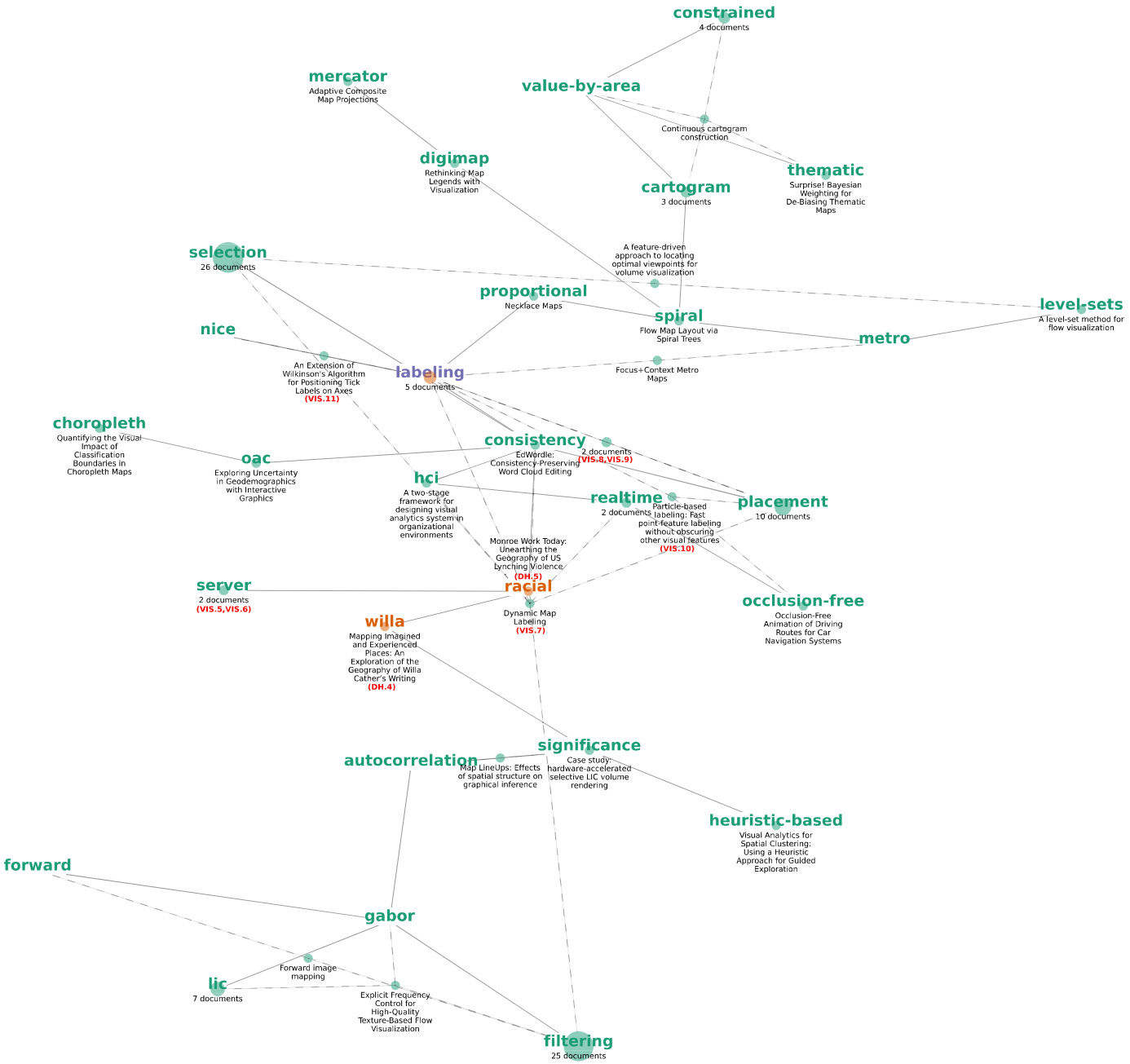


FIGURE 10. Subgraph formed by nodes in the shortest paths of "willa" and "racial." The resulting network informs on techniques related to the topic of GIS.

and autocorrelation in maps, which is ultimately connected to image mapping and display techniques, such as line integral convolution (LIC).

VI. LIMITATIONS AND FUTURE WORK

During our research, we detected certain limitations in our method that we outline and link to future lines of work below: One first obvious yet important limitation of our proposal is that it depends on an appropriate selection of keywords by the original authors of the academic papers. Selecting keywords for a publication is not a trivial task that, in our humblest opinion, is not given enough attention. The task of assigning keywords to a publication presents scholars with the following dilemma: on the one hand, keywords must be easily recognizable within the relevant area of knowledge in order to make the publication *discoverable* to other scientific peers. On the other hand, the selected keywords need to be sufficiently granular to make a given work *distinguishable* from others of a similar nature. The right combination of keywords is a balanced choice that accomplishes these two objectives simultaneously. However, as we observed during our investigation, this is not always the case. We often found relevant papers whose selection of keywords was ill-defined, a fact that negatively impacted the discoverability of such

publications. A potential solution to this issue to be explored in further developments was pointed in Section I-C when we referred to other works [60] that rely on an automatic extraction of keywords through the analysis of the papers' full texts or abstracts. Although the inclusion of these techniques could partially address the reliability issue in the primary sources, their impact on the vocabulary acquisition task needs thoroughly evaluated in future experiments dealing with different research subjects from the one employed in this study.

Another important limitation of our method is that LSA cannot handle polysemy (words with multiple meanings) effectively. It assumes that the same word means the same concept in the whole corpus, which represents a problem for words that acquire different meanings depending on the context in which they appear. Polysemy is an inherent problem to interdisciplinary research, which unfortunately cannot be resolved by LSA alone. Whereas the impact of this unwanted behavior is negligible in small vocabularies, such as the one we employed, we are aware that the stemming procedure that is applied to keywords might be problematic in bigger datasets. During our experiments, this behavior could be observed in the mismatching of different keywords that shared a common root but have different meanings (i.e., "colonoscopy/colonization" or "factory/factorial"). Some solutions have been proposed in the literature to address this kind of issue, such as the inclusion of syntactic dependencies in the construction of the PMI matrix [61]. Syntactic analysis could represent a useful alternative to mark explicit distinctions between occurrences of the same token in different multi-word keywords, in which a token may play different syntactic roles (e.g., noun, adjective). In a different approach, the polysemy problem could also be addressed through interactive term tagging. The user could generate new terms by annotating different meanings of the same token in an opposite approach to synonym detection. Not only would this interactive application be able to resolve this problem, but it could also enable a smarter exploration task in which other parameters could also be live-tuned, such as the stemming algorithm (e.g., Lancaster, Porter, Snowball), the number of singular values, smoothing factor of SPPMI or the selection of stop-words. For these reasons, the construction of an interactive application based on the methods explained in this paper represents a path that we are keen on exploring in the future.

Finally, as we introduced in Section I-C, we will seek to enhance the LBD process by supporting its close variant, which will be key in designing formal evaluations of our visualization scheme. Traditionally, the validation of results obtained in LBD has been achieved by two means: intersection [62] and expert evaluation [63]. Our intention is to combine these methods with well-established interaction and visualization evaluation practices [64] to further asses the validity of the showcased techniques and to identify further requirements for future works along this line.

VII. CONCLUSION

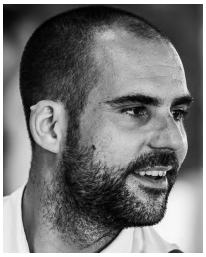
In this paper, we described an automatic method to enhance the open LBD process by visual means. The proposed method allows users to explore author-assigned keywords and related documents in two disjoints bodies of scientific literature which can accelerate the discovery of visualization techniques appropriate for a narrow-domain research interest. Our approach enables scholars to inspect local structures in proximity data derived from the latent meaning of keywords, facilitating both the progressive learning of new concepts and the acquisition of domain-specific vocabulary in a seamless manner. Furthermore, the method eliminates the need for a manual selection of terms to query the collection. Instead, we rely on a set of keyword associations extracted from an auxiliary corpus, which provides a semantic expression that is rich enough to capture specific user needs concerning a predefined multidisciplinary research purpose. Documents from the target and auxiliary corpora are jointly projected into a 2D representation of keyword proximity derived from the high-dimensional semantic space, offering the user multiple learning paths that can be readily incorporated into future research. Moreover, new keywords learned through the use of our visualization could be utilized to perform classical text queries in an online scientific database, bringing new potential data sources into question.

REFERENCES

[1] P. Federico, F. Heimerl, S. Koch, and S. Miksch, "A survey on visual approaches for analyzing scientific literature and patents," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 9, pp. 2179–2198, Sep. 2017.
[2] J. Liu, T. Tang, W. Wang, B. Xu, X. Kong, and F. Xia, "A survey of scholarly data visualization," *IEEE Access*, vol. 6, pp. 19205–19221, 2018.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[4] D. Shahaf, C. Guestrin, and E. Horvitz, "Metro maps of science," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, KDD*, New York, NY, USA, 2012, pp. 1122–1130.
[5] L. J. Bracken and E. A. Oughton, "'What do you mean?' The importance of language in developing interdisciplinary research," *Trans. Inst. Brit. Geographers*, vol. 31, no. 3, pp. 371–382, Sep. 2006.
[6] M. Liu, Y. Chen, B. Lang, L. Zhang, and H. Niu, "Identifying scholarly communities from unstructured texts," in *Web Big Data* (Lecture Notes in Computer Science), Y. Cai, Y. Ishikawa, and J. Xu, Eds. Cham, Switzerland: Springer, 2018, pp. 75–89.
[7] A. Krebs and D. Paperno, "When hyperparameters help: Beneficial parameter combinations in distributional semantic models," in *Proc. 5th Joint Conf. Lexical Comput. Semantics*, Berlin, Germany, Aug. 2016, pp. 97–101.
[8] D. R. Swanson, "Fish oil, raynaud's syndrome, and undiscovered public knowledge," *Perspect. Biol. Med.*, vol. 30, no. 1, pp. 7–18, 1986.
[9] D. R. Swanson, "Migraine and magnesium: Eleven neglected connections," *Perspect. Biol. Med.*, vol. 31, no. 4, pp. 526–557, 1988.
[10] S. Henry and B. T. McInnes, "Literature based discovery: Models, methods, and trends," *J. Biomed. Inform.*, vol. 74, pp. 20–32, Oct. 2017.
[11] M. Thilakaratne, K. Falkner, and T. Atapattu, "Automatic detection of cross-disciplinary knowledge associations," in *Proc. Student Res. Workshop (ACL)*, Jul. 2018, pp. 45–51.
[12] M. D. Gordon and S. Dumais, "Using latent semantic indexing for literature based discovery," *J. Amer. Soc. Inf. Sci.*, vol. 49, no. 8, pp. 674–685, 1998.
[13] D. Cameron, R. Kavuluru, T. C. Rindflesch, A. P. Sheth, K. Thirunarayan, and O. Bodenreider, "Context-driven automatic subgraph creation for literature-based discovery," *J. Biomed. Inform.*, vol. 54, pp. 141–157, Apr. 2015.

[14] J.-K. Chou and C.-K. Yang, "PaperVis: Literature review made easy," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 721–730, Jun. 2011.
[15] F. Beck, S. Koch, and D. Weiskopf, "Visual analysis and dissemination of scientific literature collections with survis," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, pp. 180–189, Jan. 2016.
[16] R. Santamaría and R. Therón, "Overlapping clustered graphs: Co-authorship networks visualization," in *Smart Graphics* (Lecture Notes in Computer Science), A. Butz, B. Fisher, A. Krüger, P. Olivier, and M. Christie, Eds. Berlin, Germany: Springer, 2008, pp. 190–199.
[17] H. D. White and K. W. McCain, "Visualizing a discipline: An author co-citation analysis of information science, 1972–1995," *J. Amer. Soc. Inf. Sci.*, vol. 49, no. 4, pp. 327–355, Jan. 1998.
[18] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, "Toward a deeper understanding of visualization through keyword analysis," INRIA, Paris, France, Tech. Rep. RR-8580, Aug. 2014.
[19] F. Heimerl, Q. Han, S. Koch, and T. Ertl, "CiteRivers: Visual analytics of citation patterns," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 190–199, Jan. 2016.
[20] F. N. Silva, D. R. Amancio, M. Bardosova, L. da F. Costa, and O. N. Oliveira, "Using network science and text analytics to produce surveys in a scientific topic," *J. Informetrics*, vol. 10, no. 2, pp. 487–502, May 2016.
[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Jan. 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781
[22] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "ParallelTopics: A probabilistic approach to exploring document collections," in *Proc. IEEE Conf. Vis. Anal. Sci. Technol. (VAST)*, Oct. 2011, pp. 231–240.
[23] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization techniques for assessing textual topic models," in *Proc. Int. Work. Conf. Adv. Vis. Interfaces*, New York, NY, USA, May 2012, pp. 74–77.
[24] C. Jaegul, L. Changhyun, C. K. Reddy, and P. Haesun, "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 1992–2001, Dec. 2013.
[25] M. Berger, K. McDonough, and L. M. Seversky, "Cite2vec: Citation-driven document exploration via word embeddings," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 691–700, Jan. 2017.
[26] R. W. Schvaneveldt, Ed., *Pathfinder Associative Networks: Studies in Knowledge Organization* . Westport, CT, USA: Ablex, 1990.
[27] A. S. Barb, R. B. Clariana, and C.-R. Shyu, "Applications of pathfinder network scaling for improving the ranking of satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 3, pp. 1092–1099, Jun. 2013.
[28] T. Cohen, G. K. Whitfield, R. W. Schvaneveldt, K. Mukund, and T. Rindflesch, "EpiphaNet: An interactive tool to support biomedical discoveries," *J. Biomed. Discovery Collaboration*, vol. 5, pp. 21–49, Sep. 2010.
[29] U. K. Kudikyala and R. B. Vaughn, "Software requirement understanding using Pathfinder networks: Discovering and evaluating mental models," *J. Syst. Softw.*, vol. 74, no. 1, pp. 101–108, Jan. 2005.
[30] D. L. Trumpower and T. E. Goldsmith, "Structural enhancement of learning," *Contemp. Educ. Psychol.*, vol. 29, no. 4, pp. 426–446, Oct. 2004.
[31] S. Veríssimo, V. G. Lopes, L. M. C. Garcia, and R. L. González, "Evaluation of changes in cognitive structures after the learning process in mathematics," *Int. J. Innov. Sci. Math. Edu.*, vol. 25, no. 2, pp. 7–33, Jun. 2017.
[32] W. Chen, C. Allen, and D. Jonassen, "Deeper learning in collaborative concept mapping: A mixed methods study of conflict resolution," *Comput. Hum. Behav.*, vol. 87, pp. 424–435, Oct. 2018.
[33] T. T. Chen, "The development and empirical study of a literature review aiding system," *Scientometrics*, vol. 92, pp. 105–116, Apr. 2012.
[34] A. Godwin, "Visualizing systematic literature reviews to identify new areas of research," in *Proc. IEEE Frontiers Educ. Conf. (FIE)*, Oct. 2016, pp. 1–8.
[35] C. Chen, "Visualising semantic spaces and author co-citation networks in digital libraries," *Inf. Process. Manage.*, vol. 35, no. 3, pp. 401–420, 1999.
[36] C. Chen, J. Kuljis, and R. J. Paul, "Visualizing latent domain knowledge," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 31, no. 4, pp. 518–529, Nov. 2001.
[37] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[38] C. Chen, "Tracking latent domain structures: An integration of pathfinder and latent semantic analysis," *AI Soc.*, vol. 11, nos. 1–2, pp. 48–62, Mar. 1997.
[39] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Proc. 22nd Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2009, pp. 288–296.
[40] T. K. Landauer and S. T. Dutnais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychol. Rev.*, vol. 104, no. 2, pp. 211–240, 1997.
[41] P. J. Crossno, D. M. Dunlavy, and T. M. Shead, "LSAView: A tool for visual exploration of latent semantic modeling," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, Oct. 2009, pp. 83–90.
[42] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Cambridge, MA, USA: MIT Press, vol. 2, 2014, pp. 2177–2185.
[43] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Trans. Assoc. Comput. Linguistics*, vol. 3, pp. 211–225, May 2015.
[44] E. Meeks, "Digital literacy and digital citizenshipldigital humanities specialist," Tech. Rep., 2013. [Online]. Available: https://dhs.stanford.edu/algorithmic-literacy/digital-literacy-and-digital-citizenship/
[45] S. Jänicke, "Valuable research for visualization and digital humanities: A balancing act," in *Proc. 1st Workshop Vis. Digit. Humanities (VIS4DH)*, Baltimore, MD, USA, Oct. 2016, pp. 1–5.
[46] K. Coles, "Think like a machine (or don't)," in *Proc. 2nd Workshop Vis. Digit. Humanities (VIS4DH)*, Phoenix, AZ, USA, Oct. 2017, pp. 1–5.
[47] C. Schöch, *Abstracts and Metadata from the Digital Humanities Conference 2015 in Sydney (DH2015)*. Zenodo, 2018. Accessed: Jul. 23, 2019. doi: 10.5281/zenodo.1321296.
[48] C. Schöch, *Abstracts from the Digital Humanities Conference in Kraków in 2016 (DH2016)*. Zenodo, 2018. Accessed: Jul. 23, 2019. doi: 10.5281/zenodo.1314770.
[49] C. Schöch, *Abstracts from the Digital Humanities Conference in Mexico City 2018 (DH2018)*. Zenodo, 2018. Accessed: Jul. 23, 2019. doi: 10.5281/zenodo.1344341.
[50] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko, "Vispubdata.org: A metadata collection about IEEE visualization (VIS) publications," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 9, pp. 2199–2206, Sep. 2017.
[51] T. K. Landauer and S. Dumais, "Latent semantic analysis," *Scholarpedia*, vol. 3, p. 4356, Nov. 2008.
[52] Y. Liu, J. Goncalves, D. Ferreira, B. Xiao, S. Hosio, and V. Kostakos, "CHI 1994–2013: Mapping two decades of intellectual progress through co-word analysis," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, Apr. 2014, pp. 3553–3562.
[53] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.
[54] J. Alstott, E. Bullmore, and D. Plenz, "Powerlaw: A python package for analysis of heavy-tailed distributions," *PLoS ONE*, vol. 9, Jan. 2014, Art. no. e85777.
[55] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, pp. 130–137, Mar. 1980.
[56] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguistics*, vol. 16, no. 1, pp. 22–29, Mar. 1990.
[57] F. Günther, C. Dudschig, and B. Kaup, "Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies," *Quart. J. Exp. Psychol.*, vol. 69, no. 4, pp. 626–653, Apr. 2016.
[58] X. Huang and L. Wei, "Clustering graphs for visualization via node similarities," *J. Vis. Lang. Comput.*, vol. 17, no. 3, pp. 225–253, 2006.
[59] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs," *Inf. Process. Lett.*, vol. 31, no. 1, pp. 7–15, Apr. 1989.
[60] C. Olmeda-Gómez, M.-A. Ovalle-Perandones, and A. Perianes-Rodríguez, "Co-word analysis and thematic landscapes in Spanish information science literature, 1985–2014," *Scientometrics*, vol. 113, no. 1, pp. 195–217, Oct. 2017.
[61] D. Lin, "Automatic retrieval and clustering of similar words," in *Proc. 36th Annu. Meeting Assoc. Comput. Linguistics*, Stroudsburg, PA, USA, 1998, pp. 768–774.
[62] M. Gordon, R. K. Lindsay, and W. Fan, "Literature-based discovery on the world wide Web," *ACM Trans. Internet Technol.*, vol. 2, no. 4, pp. 261–275, Nov. 2002.

[63] J. L. Hurtado, A. Agarwal, and X. Zhu, "Topic discovery and future trend forecasting for texts," *J. Big Data*, vol. 3, p. 7, Dec. 2016.
[64] C. M. Freitas, P. R. Luzzardi, R. A. Cava, M. Winckler, M. S. Pimenta, and L. P. Nedel, "On evaluating information visualization techniques," in *Proc. Work. Conf. Adv. Vis. Interfaces*, New York, NY, USA, 2002, pp. 373–374.



ALEJANDRO BENITO-SANTOS received the B.Sc. degree in computer engineering and the M.Sc. degree in Intelligent Systems from the University of Salamanca, Spain, in 2016. He is currently pursuing the Ph.D. degree with the Visual Analytics Group VisUSAL (within the Recognized Research Group GRIAL) under the supervision of Dr. R. Therón. He is currently a Research Assistant and a Lecturer with the Department of Computer Science and Automation, University of Salamanca, Spain, where he joined, in 2016. He is a member of the Visual Analytics Group VisUSAL. In his thesis, he applies visual analytics in a broad range of interdisciplinary research contexts, such as the digital humanities, sports science, or linguistics. His interests include human–computer interaction, design, statistics, and education. He has taught HCI and Introduction to Python Programming for Statisticians at the Faculty of Sciences of Salamanca.

ROBERTO THERÓN SÁNCHEZ received the Diploma degree in computer science from the University of Salamanca, the B.A. degree from the Universidade da Coruña, the bachelor's degrees in communication studies and humanities from the University of Salamanca, and the Ph.D. degree from the Research Group Robotics, University of Salamanca. His Ph.D. thesis was on parallel calculation configuration space for redundant robots. He is currently the Manager of the VisUSAL Group (within the Recognized Research Group GRIAL), University of Salamanca, which focuses on the combination of approaches from computer science, statistics, graphic design, and information visualization to obtain an adequate understanding of complex data sets. He has authored over 100 articles in international journals and conferences. In recent years, he has been involved in developing advanced visualization tools for multidimensional data, such as genetics or paleo-climate data. In the field of visual analytics, he develops productive collaborations with groups and institutions internationally recognized as the Laboratory of Climate Sciences and the Environment, France, or the Austrian Academy of Sciences, Austria. He received the Extraordinary Doctoral Award for his Ph.D. thesis.