

---

# Supporting Methodology Transfer in Visualization Research with Literature-Based Discovery and Visual Text Analytics

---



## Doctoral Dissertation

Degree of Doctor of Philosophy in Computer Engineering

**Alejandro Benito-Santos**

*Supervised by*

**Roberto Therón Sánchez**

Universidad de Salamanca

Salamanca, Octubre de 2020

This document is prepared for double-side printing.

# Supporting Methodology Transfer in Visualization Research with Literature-Based Discovery and Visual Text Analytics

*A thesis by compendium of publications by*  
**Alejandro Benito-Santos**

Degree of Doctor of Philosophy in Computer Engineering  
*Supervised by*  
**Roberto Therón Sánchez**

**Universidad de Salamanca**

**Salamanca, Octubre de 2020**



This PhD dissertation titled "Supporting Methodology Transfer in Visualization Research with Literature-Based Discovery and Visual Text Analytics", presented by Alejandro Benito Santos to obtain the PhD degree in Computer Engineering, has been carried out within the official PhD Program in Computer Engineering of the Department of Computer Science and Automation of the University of Salamanca under the supervision of Roberto Therón Sanchez, PhD.

The PhD student, Alejandro Benito Santos, and his PhD advisor, Roberto Therón Sanchez, guarantee, by signing this doctoral thesis, that the work has been carried out by the PhD student under the direction of his advisor, and as far as our knowledge is concerned, the rights of authorship have been respected.

Salamanca, October 2020.

The PhD student:

Alejandro Benito Santos

The PhD advisor:

Roberto Therón Sanchez

## **Attribution-NonCommercial-ShareAlike 4.0 International**

*CC BY-NC-SA 4.0*

You are free to:

- **Share** — copy and redistribute the material in any medium or format
- **Adapt** — remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the following terms:

- **Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial** — You may not use the material for commercial purposes.
- **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
- **No additional restrictions** — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.



<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Alejandro Benito-Santos

*A mis padres, por apoyarme y respetarme incondicionalmente a lo largo de los años.*

*Gracias por dejarme fallar. Gracias por educarme en los valores del humanismo, del trabajo, y de la humildad. Y por hacerme creer siempre que podría ser cualquier cosa que quisiera ser. Gracias por enseñarme que hay cosas más importantes que el dinero, y a distinguir lo que está bien de lo que está mal, sois los mejores.*

*A toda mi familia, a mi hermana Raquel, a mis abuelos Juan y Paquita que tanto trabajaron para que yo haya podido escribir esta tesis. A mis primas, tíos y tíos, y muy especialmente a Paqui y Daniel, que han sido lo más parecido que se puede tener a unos segundos padres. Gracias por enseñarme tantas cosas, por dedicar tiempo al niño que un día fui y sentarme en vuestras rodillas delante de un ordenador a aporrear un teclado... ¡Mirad lo que habéis hecho!*

*A Natalia, mi futura mujer, por compartir tu camino conmigo y por darme tu amor día a día, nunca paro de aprender a tu lado.*

*A mi hijo Isaac. Quién me lo iba a decir pero, ahora, sin ti, nada de esto tendría mucho sentido. Espero que algún día puedas leer la tesis de tu padre y te recuerde que el talento sin persistencia y trabajo, no sirve de nada...*



# Agradecimientos

A mi director, Roberto, por confiar en mí, apoyarme, y darme la libertad necesaria para llevar a cabo esta investigación. Muchas gracias por introducirme de tu mano en el maravilloso mundo de la ciencia.

A mis compañeras del grupo de investigación GRIAL, y muy especialmente a su director Fran, a Alicia, Juan, Feliz, Lucía y Andrea por acogerme de la mejor manera posible estos últimos años y saber lidiar conmigo cuando más lo he necesitado.

A mis amigxs lxs kinkis, por seguir a mi lado año tras año a pesar de mis repetidas ausencias debidas a la escritura de esta tesis. Si pudiera vivir mil veces, mil veces os elegiría como compañerxs de viaje. ¡ODIOBEDECER!

También me gustaría agradecer su trabajo a todos los productores de música reggae, roots, dub y DnB que he escuchado durante las largas horas de preparación de esta tesis: King Tubby, Scientist, Mad Professor, Logistics, Martin Campbell... Drop it from the top, till the very last drop! One Love!



# Abstract

The increasing specialization of science has recently motivated a rapid fragmentation of well-established disciplines into communities of interdisciplinary practice. This decomposition can be observed in a type of visualization practice known as problem-driven visualization research. Here, interdisciplinary teams of domain and visualization experts collaborate in a specific area of knowledge such as the digital humanities, bioinformatics, computer security, or sports science. This thesis proposes a series of methods inspired by recent advances in automated text analysis and knowledge representation to promote the adequate communication and transference of knowledge between these communities. The discovered methods were combined into a visual text analytics interface for scientific discovery, GlassViz, that was designed with these aims in mind. The tool was first tested in the digital humanities domain to explore a large corpus of general-purpose visualization papers. GlassViz was adapted in a later study to support different data sources representative of these communities, showing evidence that the proposed approach is also a valid alternative to address the fragmentation problem in visualization research.

**Keywords:** visual text analytics of scientific corpora, scientific and literature-based discovery, keyword analysis, distributional similarity, digital humanities, problem-driven and interdisciplinary visualization research.



# Resumen

La creciente especialización de la ciencia está motivando la rápida fragmentación de disciplinas bien establecidas en comunidades interdisciplinares. Esta descomposición se puede observar en un tipo de investigación en visualización conocida como investigación de visualización dirigida por el problema. En ella, equipos de expertos en visualización y un dominio concreto, colaboran en un área específica de conocimiento como pueden ser las humanidades digitales, la bioinformática, la seguridad informática o las ciencias del deporte. Esta tesis propone una serie de métodos inspirados en avances recientes en el análisis automático de textos y la representación del conocimiento para promover la adecuada comunicación y transferencia de conocimiento entre estas comunidades. Los métodos obtenidos se combinaron en una interfaz de análisis visual de textos orientada al descubrimiento científico, GlassViz, que fue diseñada con estos objetivos en mente. La herramienta se probó por primera vez en el dominio de las humanidades digitales para explorar un corpus masivo de artículos de visualización de propósito general. GlassViz fue adaptada en un estudio posterior para que soportase diferentes fuentes de datos representativas de estas comunidades, mostrando evidencia de que el enfoque propuesto también es una alternativa válida para abordar el problema de la fragmentación en la investigación en visualización.

**Palabras Clave:** analítica visual de textos científicos, descubrimiento científico basado en la literatura, análisis de palabras clave, similitud distribucional, humanidades digitales, visualización interdisciplinar y dirigida por el problema.



# Contents

<b>Agradecimientos</b>	<b>ix</b>
<b>Abstract</b>	<b>xi</b>
<b>Resumen</b>	<b>xiii</b>
<b>Supervisor Authorization</b>	<b>xxi</b>
<b>Contributions List</b>	<b>xxiii</b>
<b>1 PhD Dissertation</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Background and Theoretical Foundations . . . . .	2
1.2.1 Problem-Driven Visualization Research . . . . .	3
1.2.2 Methodology Transfer . . . . .	8
1.2.3 Literature-Based Discovery . . . . .	10
1.2.4 Visual Text Analytics of Scientific Literature . . . . .	12
1.3 Objectives . . . . .	13
1.4 Methodology . . . . .	14
1.5 Summary . . . . .	15
1.5.1 Preliminaries . . . . .	16
1.5.2 Data Collection and Studies on Keywords . . . . .	16
1.5.3 Automatic Knowledge Extraction and Representation . . . . .	17
1.5.4 Applicability to Other Interdisciplinary Domains . . . . .	18
1.6 Conclusions and Future Work . . . . .	19
<b>2 Resumen en Español de las Contribuciones</b>	<b>23</b>
2.1 Contribución #1 . . . . .	23
2.1.1 Resumen . . . . .	23

2.1.2	Objetivos . . . . .	24
2.1.3	Metodología . . . . .	24
2.1.4	Resultados . . . . .	25
2.1.5	Conclusiones . . . . .	26
2.2	Contribución #2 . . . . .	26
2.2.1	Resumen . . . . .	26
2.2.2	Objetivos . . . . .	27
2.2.3	Metodología . . . . .	27
2.2.4	Resultados . . . . .	27
2.2.5	Conclusiones . . . . .	28
2.3	Contribución #3 . . . . .	28
2.3.1	Resumen . . . . .	28
2.3.2	Objetivos . . . . .	29
2.3.3	Metodología . . . . .	29
2.3.4	Resultados . . . . .	30
2.3.5	Conclusiones . . . . .	30
2.4	Contribución #4 . . . . .	31
2.4.1	Resumen . . . . .	31
2.4.2	Objetivos . . . . .	31
2.4.3	Metodología . . . . .	32
2.4.4	Resultados . . . . .	32
2.4.5	Conclusiones . . . . .	33
2.5	Contribución #5 . . . . .	33
2.5.1	Resumen . . . . .	33
2.5.2	Objetivos . . . . .	34
2.5.3	Metodología . . . . .	34
2.5.4	Resultados . . . . .	35
2.5.5	Conclusiones . . . . .	35
<b>A</b>	<b>Copy of the Contributions</b>	<b>37</b>
A.1	Contribution #1 . . . . .	37
A.2	Contribution #2 . . . . .	51
A.3	Contribution #3 . . . . .	57
A.4	Contribution #4 . . . . .	75
A.5	Contribution #5 . . . . .	81





# List of Figures

- |     |   |    |
|-----|---|----|
| 1.1 | Interdisciplinary communication issues: in (A), a poor communication between the two teams leads to channels that build as a cone, limiting the size of the solution space. In (B), the Liaison serves as a translator between the two teams of experts, effectively broadening the scope of the communication channels, and thus enlarging the available solution space. Figure from [SMKS15]. ©2015. The Eurographics Association..   | 6  |
| 1.2 | Methodology Transfer Model by Miller <i>et al.</i> [MSK <sup>+</sup> 19] inspired by the communication model introduced by Simon <i>et al.</i> [SMKS15] and others [KM13, SMM12]. The model maps problems and designs from a source domain (e.g., visualization) to a given target domain (e.g., visual musicology) to find potential solutions in the source domain for existing, unsolved problems in the target domain. Figure from [MSK <sup>+</sup> 19]. ©2019 by Miller <i>et al.</i> . . . . . | 10 |
| 1.3 | The two modes of Swanson’s ABC Model for scientific discovery (extracted from contribution #3). On the left, the open mode the user provides a term which is then used to detect interesting associations in the target literature through existing links to B-concepts. On the right, the closed discovery mode finds intermediate B-concepts to validate experimental findings.. . . . .  | 11 |



# **Supervisor Authorization**

Dr. Roberto Therón Sánchez, con DNI 07976246F, profesor Titular de Universidad del Departamento de Informática de la Universidad de Salamanca

HAGO CONSTAR:

Que como director de la tesis doctoral de Alejandro Benito Santos, con DNI 70889097S, autorizo a presentar la tesis doctoral “SUPPORTING METHODOLOGY TRANSFER IN VISUALIZATION RESEARCH WITH LITERATURE-BASED DISCOVERY AND VISUAL TEXT ANALYTICS” mediante la modalidad de compendio de artículos al disponer de los siguientes artículos publicados:

- [1] A. Benito-Santos and R. Therón Sánchez, ‘A Data-Driven Introduction to Authors, Readings and Techniques in Visualization for the Digital Humanities’, *IEEE Computer Graphics and Applications*, 2020, doi: [10.1109/MCG.2020.2973945](https://doi.org/10.1109/MCG.2020.2973945). **JCR: 1.627 Computer Science, Software Engineering (51/108) Q2.**
- [2] A. Benito-Santos and R. Therón, ‘Pilaster: A Collection of Citation Metadata Extracted From Publications on Visualization for the Digital Humanities’, presented at the 5th Workshop on Visualization for the Digital Humanities, 2020.
- [3] A. Benito-Santos and R. Therón Sánchez, ‘Cross-domain Visual Exploration of Academic Corpora via the Latent Meaning of User-Authored Keywords’, *IEEE Access*, vol. 7, pp. 98144–98160, 2019, doi: [10.1109/ACCESS.2019.2929754](https://doi.org/10.1109/ACCESS.2019.2929754). **JCR: 3.745 Computer Science, Information Systems (35/156) Q1.**
- [4] A. Benito-Santos and R. Therón, ‘GlassViz: Visualizing Automatically-Extracted Entry Points for Exploring Scientific Corpora in Problem-Driven Visualization Research’, presented at the 2020 IEEE Visualization Conference (VIS), Oct. 2020. **A+ GII-GRIN-SCIE (GGS) Conference Rating. Acceptance rate: 36%.**
- [5] A. Benito-Santos and R. Therón, ‘Defragmenting Research Areas with Knowledge Visualization and Visual Text Analytics’, *Applied Sciences*, 2020. **JCR: 2.474 Engineering, multidisciplinary (32/91) Q2.**

Por todo ello firmo esta carta de autorización

  
Digitally signed by  
**THERON SANCHEZ** THERON SANCHEZ  
**ROBERTO -** ROBERTO - 07976246F  
**07976246F** Date: 2020.10.16  
13:31:53 +02'00'

Fdo. Roberto Therón Sánchez  
En Salamanca, a 16 de Octubre de 2020

# Contributions List

## Contribution #1

A. Benito-Santos and R. Therón Sánchez, ‘A Data-Driven Introduction to Authors, Readings and Techniques in Visualization for the Digital Humanities’, IEEE Computer Graphics and Applications, 2020.

- Status: **Published**
- DOI: [10.1109/MCG.2020.2973945](https://doi.org/10.1109/MCG.2020.2973945)
- Impact Factor (JCR 2019): **1.627**
- Subject Category: **Computer Science, Software Engineering**
- Quartile: **(51/108) Q2.**

## Contribution #2

A. Benito-Santos and R. Therón, ‘Pilaster: A Collection of Citation Metadata Extracted From Publications on Visualization for the Digital Humanities’, presented at the 5th Workshop on Visualization for the Digital Humanities, colocated to IEEE VIS 2020, Oct. 2020.

- Status: **Accepted**

## Contribution #3

A. Benito-Santos and R. Therón Sánchez, ‘Cross-domain Visual Exploration of Academic Corpora via the Latent Meaning of User-Authored Keywords’, IEEE Access, vol. 7, pp. 98144–98160, 2019.

- Status: **Published**
- DOI: **10.1109/ACCESS.2019.2929754**
- Impact Factor (JCR 2019): **3.745**
- Subject Category: **Computer Science, Information Systems**
- Quartile: **(35/156) Q1.**

## Contribution #4

A. Benito-Santos and R. Therón, ‘GlassViz: Visualizing Automatically-Extracted Entry Points for Exploring Scientific Corpora in Problem-Driven Visualization Research’, presented at the 2020 IEEE Visualization Conference (VIS), Oct. 2020.

- Status: **Accepted**
- Conference Rating (GII-GRIN-SCIE 2018): **A+**
- Acceptance Rate: **36%**

## Contribution #5

A. Benito-Santos and R. Therón Sánchez, ‘Defragmenting Research Areas with Knowledge Visualization and Visual Text Analytics’, Applied Sciences, vol. 10, no. 20, Art. no. 20, Jan. 2020.

- Status: **Published**
- DOI: **10.3390/app10207248**
- Impact Factor (JCR 2019): **2.474**
- Subject Category: **Engineering, Multidisciplinary**
- Quartile: **(32/91) Q2.**

# Chapter 1

# PhD Dissertation

*It is the time you have wasted for your rose  
that makes your rose so important.*

Antoine de Saint-Exupéry - The Little Prince

## 1.1 Introduction

In the first century AD, the Hispano-Roman writer Lucius Annaeus Seneca wrote: "The abundance of books is a distraction." Since then, many others have noted that the exposure to high volumes of information negatively affects decision-makers in a broad range of themes of science and human knowledge, often leading to a waste of human and computational resources. This problem, known as information overload, has also become a serious issue in modern academia, especially since the emergence of the Internet and global mass communication. Nowadays, scholars devote significant time to the searching, abstracting and sensemaking of online collections of research papers with the aim of extracting useful information to accomplish a given research aim. However, as the amounts of existing available information keep growing at increasingly higher rates every year, and science becomes more specialized, these tasks are becoming harder to complete, often leading to cold-start situations in which the user does not have enough information to perform an initial query. This issue specially affects a type of interdisciplinary visualization practice known as problem-driven visualization research (PDVR) [SMKS15]. In PDVR, domain and visualization experts collaborate to solve an inherently complex domain problem. This kind of collaboration may occur in the context of many different knowledge domains and usually leads to the emergence of workshops, conferences and reference

datasets in a wide variety of focused areas. However, and despite its apparent legitimacy, the emergence of these communities of practice may eventually lead to the formation of poorly-communicated groups of researchers who inevitably will develop redundant solutions for generic, context-independent visualization problems. Ideally, this negative effect could be avoided by allowing an effective transfer of knowledge between these communities [Bur04, MSK<sup>+</sup>19]. Although many available visual text analytics (VTA) solutions allow exploring research paper collections by employing a combination of different supporting techniques —such as natural language processing (NLP) or graph theory—, only a few take into account the particularities and challenges of interdisciplinary visualization research, whose most important related contributions mainly are given in the theoretical plane. Consequently, the work in this thesis aimed at advancing the current state-of-the-art of VTA tools by bringing together recent advances in the fields of NLP and literature-based discovery (LBD), a knowledge extraction technique originating in the biomedical domain that "generates discoveries, or hypotheses, by combining what is already known in the literature." [TFA19] The resulting interactive applications employ the concept of methodology transfer, this is, "the action of utilizing available models that provide solutions to existing and unsolved problems" [Bur04, MSK<sup>+</sup>19] to maximize the effectiveness of a browsing session.

## 1.2 Background and Theoretical Foundations

This thesis draws upon previous works in the areas of visualization, human-computer interaction (HCI), visual analytics, natural language processing, and information science which are discussed in this section. In particular, in subsection 1.2.1 I present the key concept of problem-driven visualization research (PDVR) and its related communication issues, which serve as the application context of the work in this thesis. In addition, I also discuss a model derived by other authors that is built on top of the communication model of PDVR, known as the methodology transfer model (MTM) in Section 1.2.2. This model, as it is explained in later sections, is one of the two main foundations of my own document exploration model that is one of the main contributions of this thesis. The other foundation can be found in Swanson's ABC model of literature-based discovery that I present in subsection 1.2.3. In subsection 1.2.4 I introduce the concept of visual text analytics, which is the discipline in which I frame the document exploration task. In addition, I discuss previous works in the field, with a focus on those related to the extraction of knowledge from

scientific corpora. In relation to the latter, I comment on the sensemaking models that have been incorporated into these works and how they can be used to accelerate knowledge acquisition and the generation of novel research ideas in an academic setting, which also inspired the design of my own model of document exploration derived from the MTM and ABC models.

### 1.2.1 Problem-Driven Visualization Research

Problem-Driven Visualization Research (PVDR), is a type of visualization practice that brings together domain and visualization experts who collaborate in a broad range of different knowledge domains, such as computer security, bioinformatics, or digital humanities, to name a few. Under this setting, domain experts supply driving problems and visualization experts provide expertise in data analysis and visualization techniques to solve non-trivial problems in the given target domain and thus, the combination of these competences is key for the success of the project. This is in contrast to technique-driven visualization research, in which the effort is put on creating "new and better techniques without necessarily establishing a strong connection to a particular documented user need" [SMM12]. Rather, PDVR aims at working "with real users to solve their real-world problems" [SMM12]. However, this collaboration poses several challenges that have been addressed extensively by the HCI and visualization communities in the past, starting by the pioneer work by McCormick *et al.* [McC87]. In their work, the authors suggested that the basis for solving visualization challenges should emerge from domain needs and processes found by collaborative teams formed by scientists, engineers and visualization researchers. Around the same time, Donna Cox advocated for the concept of the "Renaissance team" — a multidisciplinary team of experts who collaborate to solve visualization problems — [Cox87]. Although McCormick and Cox specified with great detail the components of these teams — even their respective skill sets — they did not offer guidance on how these teams could collaborate in a successful manner.

Since then, visualization and HCI researchers have refined the concepts introduced by McCormick and his colleagues with the aim of improving the design studies resulting from these collaborations, making clear distinctions between interaction and collaboration, interdisciplinary and multidisciplinary teams, or independence and interdependence. Kirby and Meyer [KM13] provide a glossary of these terms that are summarized here below.

### Interaction vs Collaboration

1. **Scientific interaction:** It is the most basic form of communication between researchers and usually consists of the passing of data or higher-level expression of information and ideas. In the context of this thesis, this occurs between researchers in different disciplines, but it can be observed more often between researchers in the same discipline. A common scenario for interactions of this kind is when a visualization researcher obtains data from a colleague in another discipline and produces images that are sent back to the sender. As the work with the data progresses and the analysis method is improved, the *trajectory* of the visualization researcher diverges from that of her colleague, resulting in the publication of results in a visualization conference.
2. **Collaboration:** In this type of scientific exchange, the individual research trajectories of the stakeholders are affected in the same manner according to a set of goals. In contrast to the previous example, this exchange is a bidirectional one in which the research trajectory of the other colleague is changed by the images and figures provided by the visualization researcher. During this exchange, both researchers agree on a set of common goals that individually affect them in the same manner and that pairs their trajectories together for the rest of the research endeavor. This special form of interaction is often seen between researchers in different fields.

### Interdisciplinary versus Multidisciplinary

1. Interdisciplinary team: Interdisciplinary teams address problems in an area of science where a discipline gap exists and that need to be resolved by means of hybrid approaches that draw from multiple disciplines. Interdisciplinary team members set research goals in a collaborative manner, and they are all equal partners in terms of workload, responsibility and acclaim for the achievement of such goals. This kind of setting may produce new disciplines: for example, computer science is a field that emerged from the collaboration of researchers working in the gap between applied mathematics and electrical engineering.
2. Multidisciplinary team: A multidisciplinary team tackles problems that are multidimensional according to the number of different disciplines to which their questions and challenges can be mapped, and that require disciplinary confluence to be solved. In this mode, researchers work in parallel in different

tasks according to their respective disciplinary goals. For example, visualization is a case of multidisciplinary research that involves researchers from cognitive science, design, human-computer interaction and computer science, among others.

3. Intradisciplinary team: This third type of teams are formed whenever a discipline becomes too large to be covered by a single researcher. Intradisciplinary teams gather researchers in the same discipline who have different, often complementary skills. Depending on the subject under study, these teams may resemble more an interdisciplinary or a multidisciplinary team. This is the case of uncertainty visualization, in which it is common to see teams formed by visualization experts focusing on cognition and perception, AI explainability, or visual analytics.

According to the presented classification, PDVR mostly occurs in interdisciplinary collaborations that offer the opportunity to conduct design studies in which visualization techniques are applied to solve problems in an existing discipline gap between visualization research and another domain. In 2012, Sedlmair *et al.* [SMM12] conducted an extensive review of the literature of that time describing design studies in the fields of visualization, HCI and social sciences. Resulting from this study, they proposed a design study methodology and a nine-stage framework for conducting design studies in collaboration with domain experts. In addition, they made a clear distinction between their proposed methodology and others with common elements, such as *ethnography*, *grounded theory*, or *action research*.

Beyond this classification, the work by Kirby *et al.* had also a pioneering role in characterizing the role of communication in interdisciplinary design studies, which is one of the hot topics of this thesis. Among their recommendations to sustain viable interdisciplinary collaborations, they explicitly refer to multilingualism as a key skill in this context. Drawing from concepts initially laid out by geographers Bracken and Oughton [BO06], they state that much of the disciplinary training focuses on learning terminology, vocabulary, and nomenclature to structure and communicate ideas. As such, the initial stages of a collaboration typically focus on achieving a common language that is shared by the project’s stakeholders and that allows for an effective flow of ideas between researchers with different backgrounds. In the text, the authors also warn about the difficulty of this language-acquisition stage, which they link to two common collaboration pitfalls: the limited expressiveness and richness of the resulting shared language and the potential negative influence of

dominant personalities among team members, who may bias the language towards a single discipline.

In more recent work, Simon *et al.* [SMKS15] delved deeper into the particularities and issues of interdisciplinary communication in visualization design studies, who exemplified the collaboration as a *metaphor of spaces*. (see Figure 1.1).

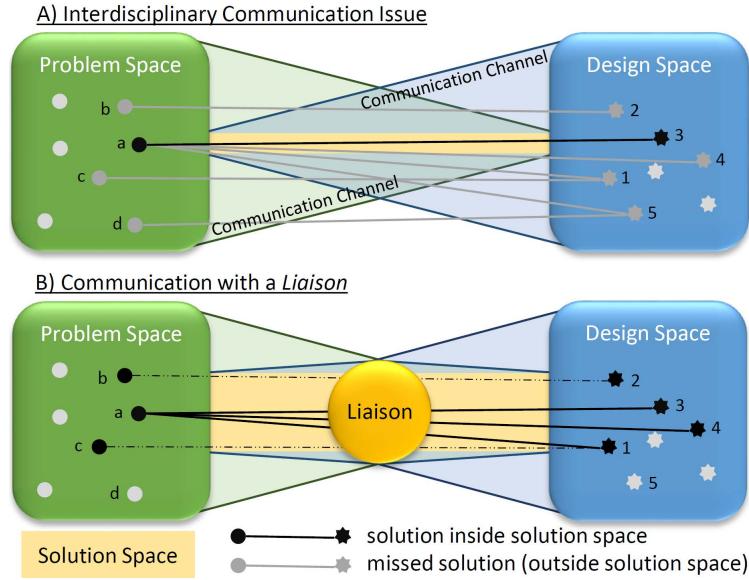


Figure 1.1: Interdisciplinary communication issues: in (A), a poor communication between the two teams leads to channels that build as a cone, limiting the size of the solution space. In (B), the Liaison serves as a translator between the two teams of experts, effectively broadening the scope of the communication channels, and thus enlarging the available solution space. Figure from [SMKS15]. ©2015. The Eurographics Association..

Under the communication model proposed by the authors, domain experts generate the *problem space* with questions about the data, goals, tasks and other constraints. Analogously, visualization researchers span a *design space* with solutions that are adequate for the problems at hand, including data analysis algorithms or visual analysis techniques. The aim of the team in the design study is to find mappings between problems and potential visual solutions that are confined to the solution space (in yellow in the figure). The more grounded knowledge participants have in both domains, the better these solutions will be (the mappings will more correctly relate problem abstractions to adequate designs). The solution space is defined by

the ability of team members in each side of the collaboration to connect their own knowledge to that of their counterparts. At the top of Figure 1.1, this effect is displayed: each team of experts employs its own communication channel to exchange ideas with their counterparts in the other side of the collaboration. A poor communication will only allow for a small portion of the knowledge to reach the other side, resulting in the communication channel adopting a shape of a cone, and effectively limiting the breadth of the solution space. As this communication becomes more effective, the communication channel acquires a trapezoid shape, which in turn enables a broader solution space and, as a consequence, a larger number of mappings between problems and designs (represented as black solid lines in Figure 1.1).

In order to augment the communicative capabilities of the team, and given the common impossibility to find individuals who are experts in both domains, the authors suggest the introduction of a *Liaison* in problem-driven visualization projects. The *Liaison* is a type of project stakeholder who holds knowledge and language in both the visualization and target domains and who mediates between the two teams of experts to foster a better inter-domain communication. Simon *et al.* [SMKS15] describe three types of *Liaison* users, according to their original domain of expertise: the first type is the domain expert who has developed an interest in visualization (**domain *Liaison***) due to previous experience in similar design studies. Despite her low level of visualization literacy, this role of this individual is key to produce a successful abstractions of the problems and validate design alternatives. The second type is the **visualization *Liaison***, a visualization expert that has acquired knowledge in the target domain through experience. The third type (the **interdisciplinary *Liaison***) is the ideal case that was presented at the beginning of this paragraph. This individual has grounded knowledge in both domains but she might often not be at hand (or even exist), much especially when the discipline is novel. As the discipline becomes more established, and as a result of the creation of specialized courses and other specific academic curricula, this kind of *Liaisons* start to be seen more often in visualization design studies. The work in this thesis is aimed at supporting the knowledge acquisition task of the first and second *Liaisons* types, who are arguably the most commonly found types when a new discipline is born and thus, they are the ones who have to face the largest knowledge gap.

### 1.2.2 Methodology Transfer

#### 1.2.2.1 What is Methodology Transfer?

Methodology transfer (MT) refers to the practice of reusing available models to provide solutions for novel, unsolved problems. The practice was first introduced into the visualization domain by Burkhard in 2004 [Bur04] who, inspired by common practices in the domain of architecture, advocated for a transference of knowledge between different stakeholders and communities of practice. To this end, he defined the concept of **knowledge visualization** (as opposed to information visualization) which is "*the use of visual representations to improve the transfer of knowledge between at least two persons or groups of persons.*" Building on this definition, he proposed a knowledge visualization framework that aimed (1) to systemize visualization methods, (2) to identify missing research areas and (3) to mediate between different research areas. The framework is heavily based on the ideas previously introduced by Eppler [Epp04], who made a distinction between information and knowledge visualization. According to Eppler, knowledge visualization needs to be able to transfer insights to answer questions such as "why?" or "how?" that go beyond communicating facts (that provide answers to the questions "what?", "who?", "when?" or "how many"?).

Burkhard's knowledge visualization framework has three well-defined dimensions or perspectives according to the purposes that were previously introduced:

1. **Knowledge Type Perspective:** Aims at identifying the type of knowledge to be transferred, which can be: (1) declarative or know-what, (2) procedural or know-how, (3) experimental or know-why, (4) orientational or know-where and (5) individual or know-who.
2. **Recipient Type Perspective:** Aims at identifying the target group that will receive the knowledge, which can be an individual, a team, an organization or a network of persons. Knowing the context and cognitive background of the recipient(s) is key to find an appropriate visualization method that enables the transfer of knowledge.
3. **Visualization Type Perspective:** Aims at establishing a taxonomy that organizes existing visualization methods in order to mediate between different application areas.

However pioneering and groundbreaking, the contribution by Burkhard was a theoretical one and did not present concrete examples or applications of the model

to areas of visualization practice. Nonetheless, since the knowledge visualization framework was first introduced to the scientific community, many other researchers (including the own Burkhard) have applied it to different areas of interdisciplinary visualization practice such as urban planning [Bur05], decision-making support in the medical domain [ELA16], or education [FvS18]. In general, the current literature displays a trend in which Burkhard’s framework is used to propose novel ways of interdisciplinary collaboration that require a transfer of knowledge between established research areas (in this context, visualization and another one). Moving closer to this thesis’ main area of application, the digital humanities, in recent work Miller *et al.* [MSK<sup>+</sup>19] elaborated on Burkhard’s ideas to frame a novel research field (*visual musicology*) as per the principles of methodology transfer. This contribution is thoroughly discussed below, due to the great influence it exerted on this thesis work.

### 1.2.2.2 Using Methodology Transfer to Frame Novel Research Areas

Musicology is a sub-field of traditional humanities that focuses on producing research-based studies of music. Musicology research typically adopts three main forms, (1) historical musicology, (2) systematic musicology and (3) ethnomusicology, depending on the concrete subject of study within the broad concept of music. Departing from the concept of musicology, Miller *et al.* [MSK<sup>+</sup>19] coined the term *visual musicology* as the study of music supported by visual analytics techniques. As the authors explain, this novel research area poses an under-explored context for collaboration between visualization and musicology researchers who may offer exciting new opportunities for all involved parties [BEC<sup>+</sup>18]. On the downside, this novelty also means that new researchers to the collaboration may encounter important difficulties in finding existing research from which they can obtain knowledge to advance the field. To overcome this important issue, the authors draw on the communication model proposed by Simon *et al.* [SMKS15] that I introduced in subsection 1.2.1. The model is augmented with a further characterization of the three spaces (problem, design and solution) in two different domains in which the concept of methodology transfer is introduced. The augmented model is shown in Figure 1.2. As I introduced previously, one of the main contributions of this thesis is an extension of this model with concepts drawn from NLP and information science with the aim to automate the detection of potential methodology transfers (see subsection 1.5.3 for further details on this).

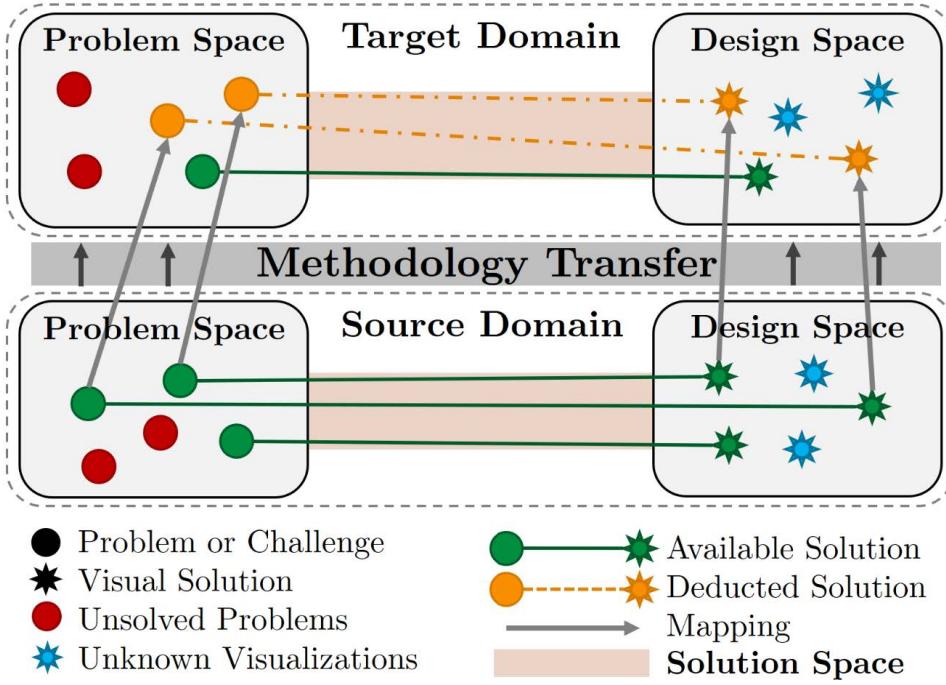


Figure 1.2: Methodology Transfer Model by Miller *et al.* [MSK<sup>+</sup>19] inspired by the communication model introduced by Simon *et al.* [SMKS15] and others [KM13, SMM12]. The model maps problems and designs from a source domain (e.g., visualization) to a given target domain (e.g., visual musicology) to find potential solutions in the source domain for existing, unsolved problems in the target domain. Figure from [MSK<sup>+</sup>19]. ©2019 by Miller *et al.*.

### 1.2.3 Literature-Based Discovery

Literature-Based Discovery (LBD) refers to the kind of knowledge extraction and automated hypothesis generation that generates new insights by logically connecting a-priori independent fragments of information typically found in the scientific literature. Don R. Swanson, an American information scientist, popularized this technique in the 1980s by employing it to make important discoveries in the biomedical domain, such as the treatment for Raynaud's disease (fish oil) [Swa86] or the connection between magnesium intake and migraines [Swa88]. In order to make these discoveries, Swanson followed a simple syllogism, named the *ABC Model*: "if concept A is linked to concept B, and at the same time concept B is linked to concept C, then concept A is associated with concept C, and concept B characterizes the relationship between

concepts A and C". Under this assumption, concept A can be denoted as the starting term/concept, the B concept(s) as the intermediate link(s), and the concept C as the target term/concept. The ABC Model supports two variants for *open* and *closed* discovery (presented in Figure 1.3).

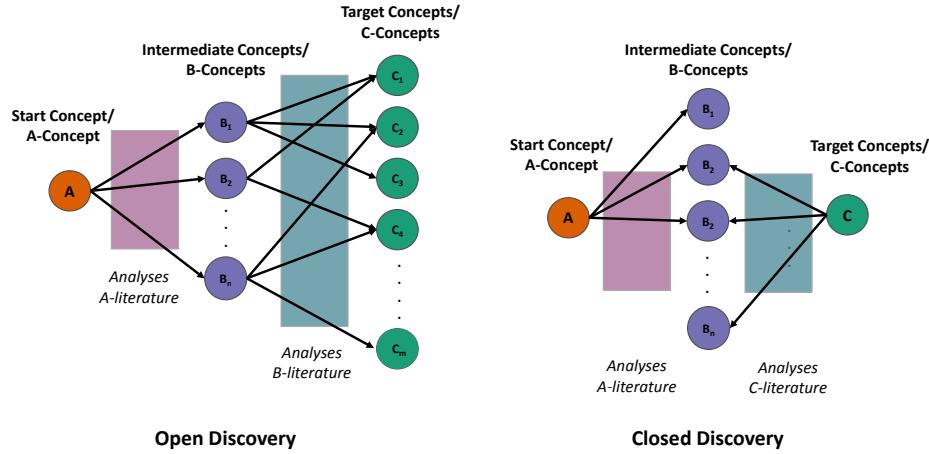


Figure 1.3: The two modes of Swanson's ABC Model for scientific discovery (extracted from contribution #3). On the left, the open mode the user provides a term which is then used to detect interesting associations in the target literature through existing links to B-concepts. On the right, the closed discovery mode finds intermediate B-concepts to validate experimental findings..

In the open discovery mode, the LBD process is started with an initial term provided by the user that is employed to generate term associations of type B and C. The open discovery mode often is used to *generate hypotheses*. Conversely, in the closed discovery mode, the user provides terms of type A and C that are used by the model to reveal intermediate links or B-concepts. This second approach is generally used for *hypothesis testing and validation* [HM17]. Although originally Swanson obtained his findings by manual means, there is a rising interest among information scientists [TFA18] to automate the workflows of LBD, most of which [TFA19] employ similarity scores derived from popular linguistic models known as word embeddings [MSC<sup>+</sup>13, PSM14], which are vector representations of the words in a corpus and are mined from massive online databases of scientific documents. As I discuss in section 1.5, a key contribution of this thesis involved the development of a similarity model for author-assigned keywords, which I demonstrated in the last three contributions of this thesis.

### 1.2.4 Visual Text Analytics of Scientific Literature

The visual solutions implemented in this thesis derive from two larger bodies of visualization research known as visual text analytics and visualization of scientific literature which just recently started to converge into a single discipline: visual text analytics of scientific literature. In this section, I refer to recent contributions in which the trend is to incorporate sensemaking models specific to the exploration of scientific texts (e.g., those adopted by users performing literature reviews).

Visual Text Analytics (VTA) is a novel specialization of a larger research discipline known as visual analytics (VA). Research in VA formally started in the 2000's decade [TC06, KMS<sup>+</sup>08, KAF<sup>+</sup>08] with the aim of augmenting the user's analytical capabilities and promoting analytical reasoning on diverse kinds of data by exploiting the visual pattern recognition mechanisms of the human brain. To achieve these goals, VA tools incorporate elements from other areas of science, such as visualization, perception, decision-making, interface design or data mining into graphical interfaces. VTA employs the same principles as VA but specializes on the processing and presentation of text, which can be structured, semistructured, or unstructured. Originally, many VTA contributions were general-purpose sensemaking tools that were often demonstrated with scientific corpora, among other kinds. These tools combined different text-mining and visualization techniques to display topical commonalities between the distinct components of a corpus, showcasing not only visual representations of the underlying themes, but also adequate interaction techniques that allowed a correct manipulation of the algorithms at play. This is the case, for example, of the lens metaphor, which has been refined and adapted in several studies to different aims [HJQ<sup>+</sup>16, KKP<sup>+</sup>17, BMS17]. The river metaphor is also often seen in these kinds of approaches to depict topic shifts in time [DWCR11, HHKE16].

Also in recent times, some authors have started to incorporate cognitive and sensemaking models specific to the tasks and goals involved in the exploration of scientific texts into their VTA tools. For example, the Action Science Explorer [DSG<sup>+</sup>12], PaperQuest [PEM16], and PaperPoles [HPLC19], among others [WLQ<sup>+</sup>16] mimic the sensemaking process of traditional literature reviews. Guo *et al.* [GL18] propose a two-stage sensemaking framework for the discovery of novel research ideas that is based on previous work by Pirolli and Card [PC05]. In a different approach to the problem, other authors propose visual narratives of a literature review that relate the literature review process to the development of a theater play [PC18]. Following the path set by these authors, in contributions #4 and #5 I explain how I developed an interactive application that adopted a sensemaking loop inspired by

Miller *et al.*'s methodology transfer model.

### 1.3 Objectives

After having introduced the main concepts behind this thesis work in the previous sections, I outline below the main objectives that I aimed to achieve in this thesis. These included the development of state-of-the-art text-mining and graphical representations and interfaces to address open problems in the fields of PDVR and visual document search. Specifically, these objectives were:

1. **To understand current challenges in interdisciplinary visualization research.** PDVR has become an important area of visualization research that involves increasingly larger numbers of researchers every year. Identifying common challenges and goals between different expressions of PDVR is required to propose interactive tools to support these scholars' activities.
2. **To develop a methodology to identify and map novel expressions of PDVR.** As more collaboration areas become available, it is important to develop techniques to quantify their activities to relate them to other areas of interdisciplinary visualization research and frame them in the bigger picture of visualization practice.
3. **To study the properties of the language defined by author-assigned keywords.** Author-assigned keywords adopt a highly formulated and condensed format, modeling a unique and expressive language that deserves further study. As it turns out, the process by which humans extract keywords from academic texts remains mostly unknown. Understanding how authors choose these keywords is necessary to design adequate automatic text summarization and representation methods.
4. **To accelerate knowledge discovery in the document exploration task when performed in an interdisciplinary research context.** There is currently a lack of interactive tools that facilitate importing existing methodologies into new application domains. However, it is well known that certain visual solutions may work equally well in different application contexts, as long as they support the same visual tasks. I argue that tools supporting interdisciplinary researchers should be designed with these principles in mind to maximize their effectiveness.

5. **To address the fragmentation problem in visualization practice.** As more focused communities of practice become available, it is necessary to ensure a sound transference of knowledge between them to avoid the development of redundant solutions for generic problems in visualization research.

## 1.4 Methodology

The work in this thesis required the application of both theoretical and practical methods in different areas of knowledge such as scientometrics, natural language processing, and visualization design. The strategy I adopted adhered to the traditional guidelines of the scientific method, which I adapted to achieve the research aims presented in the previous chapter. Concretely, the following guidelines were employed:

1. **Observation** through the study of a novel area of PDVR known as "visualization for the digital humanities." Here, the particularities and challenges of scientific documents exploration in this research context would be collected. Furthermore, a review of state-of-the-art VTA techniques should be performed, which will shed light on the methods needed to build adequate tools.
2. **Hypothesis formulation:** design of novel linguistic, text mining, and visual methods, models and algorithms that adapt, extend and/or combine previous works in these fields with the aim of improving the document exploration task in a PDVR context. These new artifacts should adhere to standard best practices in their respective domains and should be tested in real-world scenarios.
3. **Observation gathering:** achieved by evaluating the methods developed in the previous stage with data originating in the digital humanities domain. The methods will be progressively refined as more evidence becomes available until they are solid enough to be integrated in an interactive application.
4. **Contrasting the hypothesis:** the methods obtained in previous steps, which were designed and tested in the digital humanities domain, should be exemplified with data obtained in other interdisciplinary domains.
5. **Hypothesis proof or refusal:** Acceptance, rejection or modification, in due case, of the developed techniques as a consequence of the experiments and studies carried out in the previous stages. The steps previously outlined should

be repeated when necessary to obtain the necessary evidence to support or reject the main hypothesis.

6. **Scientific Thesis:** Extraction, structuring and synthesis that allows the appropriate communication of the conclusions obtained at the end of the research process, which are documented in a memory of the thesis.

## 1.5 Summary

In this chapter, I provide arguments to support the coherence between the research papers included in this PhD. thesis. These articles can be grouped into three main themes: in the first one, I include works that focused on creating a dataset of representative publications in a community of interdisciplinary visualization practice (i.e., visualization for the digital humanities). The second one holds papers describing the automatic, text-based knowledge extraction and presentation techniques that are the core of this thesis' work. Finally, in the last group I include a study in which I extended these techniques to cover other interdisciplinary research areas with the aim of resolving the fragmentation problem in visualization practice.

- Data collection:
  - Contribution #1: A. Benito-Santos and R. Therón Sánchez, ‘A Data-Driven Introduction to Authors, Readings and Techniques in Visualization for the Digital Humanities’, IEEE Computer Graphics and Applications, 2020.
  - Contribution #2: A. Benito-Santos and R. Therón, ‘Pilaster: A Collection of Citation Metadata Extracted From Publications on Visualization for the Digital Humanities’, presented at the 5th Workshop on Visualization for the Digital Humanities, colocated to the 2020 IEEE Visualization Conference, 2020.
- Automatic knowledge extraction and representation:
  - Contribution #3: A. Benito-Santos and R. Therón Sánchez, ‘Cross-domain Visual Exploration of Academic Corpora via the Latent Meaning of User-Authored Keywords’, IEEE Access, vol. 7, pp. 98144–98160, 2019.
  - Contribution #4: A. Benito-Santos and R. Therón, ‘GlassViz: Visualizing Automatically-Extracted Entry Points for Exploring Scientific Corpora

in Problem-Driven Visualization Research', presented at the 2020 IEEE Visualization Conference (VIS), Oct. 2020

- Applicability to other interdisciplinary domains:
  - Contribution #5: A. Benito-Santos and R. Therón, 'Defragmenting Research Areas with Knowledge Visualization and Visual Text Analytics', Applied Sciences, 2020

### 1.5.1 Preliminaries

Through my involvement in three interdisciplinary visualization research projects in the fields of sports science [LTB16, BTL<sup>+</sup>18] and digital humanities [BTL<sup>+</sup>17], I was able to experience many of the typical problems of working with domain experts. Despite the very different profiles of the experts I worked with in these projects, I detected several parallelisms and recurrent issues that kept appearing during the development of these studies. In particular, I noticed a lack of methodologies and visual techniques to support the exploration of research papers collections, which is a key task that typically occurs at the initial stages of interdisciplinary visualization research. During this task, visualization and domain experts participating in the project seek visualizations in the literature that can be adapted to the particularities of the problem at hand. This practice has many benefits, since the initial visualization prototypes resulting from adapting existing techniques are often employed as a non-verbal vehicle to leverage the communication issues that I discussed in section 1.2. However, I could notice the literature was scarce on visual text analytics tools that adopted this sensemaking model, which led me to focus on filling this void during my PhD. work. Concretely, my main objective was to combine state-of-the-art text mining and visualization approaches into an interface that allowed knowledge to be transferred between communities of practice. To this end, as I describe below, I needed to start by analyzing in depth one of these communities. Given my familiarity with visualization for the digital humanities, which was the topic of my MSc. thesis [BSTS16], I decided to choose it over other alternatives to get the work in this thesis started.

### 1.5.2 Data Collection and Studies on Keywords

During a preliminary literature review, I noticed several visualization studies employed author-assigned keywords to produce maps and datasets of the discipline

[IIS<sup>+</sup>14a, IIS<sup>+</sup>14b, IIS<sup>+</sup>17, IHK<sup>+</sup>17]. Since keywords contain rich expert knowledge by the authors, and are often used in the search task, I decided to replicate these studies in the visualization for the digital humanities domain. The compilation of these datasets would allow me to understand the challenges of this research area, and also served me to design a data collection methodology that I employed later to construct datasets in other interdisciplinary areas (contribution #5). The publication metadata and insights obtained in the two first studies of this thesis drove the rest of the work, in which I proposed several visual methods and interactive applications to explore keyword datasets with the aim to accelerate language and knowledge acquisition in the document exploration task.

### 1.5.3 Automatic Knowledge Extraction and Representation

**Modeling Keywords Similarity:** The second group of papers revolved around the idea of developing automatic knowledge extraction and representation techniques employing the datasets described previously. My investigations' findings were summarized in a thematic series of two research papers (contributions #3 and #4). In the first case, "*Cross-domain Visual Exploration of Academic Corpora via the Latent Meaning of User-Authored Keywords*," I contributed a distributional similarity model for author-assigned keywords based on the ideas by Levy *et al.* [LG14, LGD15]. Besides, I also proposed a method to cluster a large similarity matrix that adopted LBD principles. The method automatically extracted interesting exploration paths by connecting a- and c-concepts between two disjoint bodies of literature, and was demonstrated in two use cases. Finally, inspired by Chen's previous works on knowledge visualization [Che97, Che99, CKP01], I designed a method to produce a joint visualization of concepts and documents using Kamada-Kawai's force-directed layout algorithm [KK89]. Despite the significant advance this contribution supposed in the thesis's context, the visualization I obtained was static and rather limited in interactivity. Although I considered it was good enough for an initial proof-of-concept, as much of the work had been put into the conception of the similarity model and the graph clustering algorithm, I knew that I would have to apply these software artifacts in a VTA tool of higher fidelity, which I presented in the next contribution of this thesis.

**GlassViz:** In the second paper of this series, "*GlassViz: Visualizing Automatically-Extracted Entry Points for Exploring Scientific Corpora in Problem-Driven Visualization Research*," I devoted my work to two main tasks: first, I aligned my seman-

tic similarity model with recent advances in the field of PDVR. More concretely, I posed it as an automatic solution for the interdisciplinary communication problems defined by Burkhard [Bur04] and Simon *et al.* [SMKS15], and adopted the methodology transfer concept developed by Miller *et al.* in their work on visual musicology [MSK<sup>+</sup>19]. The result was a theoretical model that employed distributional similarity to analyze the interdisciplinary communication channel used by domain and visualization experts, and ultimately drive the exploration of a large corpus of scientific documents. In a second task, I implemented this model in a multiple-view application that supported the qualitative inspection of *quality* a-concepts' neighborhoods (which I called entry points to the dataset). Building on my previous findings, the tool presented these entry points as semantically cohesive groups of keywords in several different force-directed node-link diagrams. Also, the prototype implemented several well-known interaction techniques, among which brushing+linking stands out. This technique allowed the user to select groups of concepts in the node-link diagrams and update two frequency-rank lists that displayed, respectively, a reading order for potentially interesting papers in the entry point and allowed a rapid interpretation of its underlying themes.

#### 1.5.4 Applicability to Other Interdisciplinary Domains

In the last contribution of this thesis, "Defragmenting Research Areas with Knowledge Visualization and Visual Text Analytics," I applied the findings obtained in previous studies to identify knowledge associations and groups of common interests between four different communities of interdisciplinary visualization practice (i.e., biological data visualization, visualization for computer security, sports data visualization, and visualization for the digital humanities). To this end, I built three keyword sets following the same rationale I employed in the first two contributions of this thesis, which served as the input data for an extensive descriptive statistical study of keywords in the four domains. Besides, I also adapted the similarity model to support different sources and calculated overlapping sets of keywords between the distinct sets considered in the study. Then, I evaluated whether the sizes of these overlapping sets influenced the similarity scores found by the modified model, a hypothesis that I could not prove. This finding suggested that the similarity scores of the model are more dependant on how specific low-frequency keywords are combined than on the number of coincident terms between the collections. Lastly, I adapted the GlassViz interface to support the browsing of inter-domain thematic coincidences between the four domains. The results showed that 1. the model and interactive ap-

plication conceived in previous contributions could be successfully applied to other interdisciplinary domains and 2. that automatic text-based, interactive methods are an excellent alternative to foster the transference of knowledge between many loosely-connected communities of practice in an area of research.

## 1.6 Conclusions and Future Work

In this thesis work, I focused on designing a series of methods that allowed knowledge to be transferred between different problem-driven communities of researchers, addressing several important challenges of interdisciplinary research and modern science. To this end, I approached the problem from a novel area of research known as visualization for the digital humanities that I was familiar with. The process I followed to provide a data-driven analysis of the field helped me conceive a methodology to map diffuse research areas that takes into account this sense of community. The dataset I obtained was made publicly available and presented at one of the top venues on visualization and digital humanities in the world. Among other insights, these studies allowed me to confirm that author-assigned keywords are a powerful instrument for text representation. Although I was able to understand some of their properties by the different numerical and qualitative analyses that I proposed, I could not obtain many insights into the process by which humans extract them, which will remain a mystery for now. Using these learnings, I also developed a novel model that used keywords to represent the interdisciplinary communication channel proposed by other authors. I exploited this model to automate the discovery of interesting methodologies and promote a reunifying vision of the field in a VTA tool called GlassViz, which was the first one of its kind to implement a set of design goals and a sensemaking model specially conceived for interdisciplinary research. GlassViz was presented at the top visualization venue in the world. In summary, I can extract the following conclusions from the work performed in this thesis: (1) interdisciplinary science and, particularly, problem-driven visualization research, are becoming essential parts of current scientific practice, and I believe they will continue growing in importance in the future. However, there are still very few approaches that address the fragmentation and communication problems these kinds of practices bring with them. In this thesis, I proved that it is possible to design interfaces to assist interdisciplinary researchers in performing linguistic tasks (e.g., qualitative neighborhood inspection, synonym detection) that currently are highly problematic. (2) The summarization and keyword/keyphrase extraction from scientific documents

currently are underexplored tasks of modern science. In this thesis, I focused on author-assigned keywords because they are a unique feature of scientific texts that has not been reproduced by a machine so far. The assessment of their comprehensibility in comparison to automatic extraction techniques — which seems to be the trend nowadays — should be a priority in the areas of human-computer interaction, psychology, and linguistics. However, I do not think this is currently the case. My opinion is that much of the emphasis in NLP and visualization nowadays is put on building automatic keyword extraction techniques, while little attention is being paid to how humans themselves perform this task in an academic setting. After all, the validity of a vast majority of text summarization and topic modeling algorithms is measured against human performance. However, the mechanics of these mental processes are still largely unknown. Therefore, further work is required to design methods that mimic humans' capabilities and so, they can be better interpreted by humans, too. (3) As I have showed throughout this thesis work, analyzing interdisciplinary research also is a challenging task that requires both deep understanding of the topic itself and the larger body of knowledge they are framed in. Usually, the novelty of these areas makes it impossible to resort to traditional scientific mapping or literature review methodologies to produce holistic visions of them. This is because such methodologies often are oriented towards the analysis of more established fields, and *excessively* rely on online databases and third-party algorithms to produce their results. Thus, they usually neglect the preliminary and diffuse character a discipline surely has in its beginnings. In this thesis, I envisioned a methodology to map such new-born disciplines which departs from the individuals who are part of a community, rather than from results obtained from a database managed by a third party. I think my contribution in this regard is notable: if we are set to take control over the machines, we, the scientists, should make an explicit effort to center our methodologies on the human. This last reflection takes me to my next point, which is the intrinsic value that humanities can bring into the experimental sciences in general, and computer science in particular: as the reader may have guessed already, to me, the separation of humanities and science is deeply nonsensical. While I acknowledge certain differences in their main aims may exist (e.g., the value of philosophy is not to provide definite answers to the questions , but rather learn how to make more and better questions), I believe that a crisp categorization of knowledge, as we see it practiced in our educational systems, in our schools and universities, is not only unnecessary, but also counterproductive. As it turns out, knowledge is deeply intertwined: the field of computer science can benefit from humanities research as

much as the opposite. The separation between humanities and science is an artifice induced by industrialism that I think should be removed for good, and I hope this thesis can contribute towards that aim.

On a different perspective, over the course of this investigation, I noticed several potential lines of work that I could not fit into this thesis and that I briefly discuss hereafter. The first line concerns implementing a joint projection of documents and keywords that allows the user to interact with the underlying linguistic algorithms (e.g., the tokenizer or the matrix factorizer) in reasonable latency times. Although I experimented with this approach in contribution #3, in GlassViz the documents and terms are shown in specific views. Whereas this seemed to work well for a first approach, the ideal representation should display them in the same area. Also, in this thesis I relied heavily on the qualitative inspection and filtering of neighborhood and paths originating at a-concepts, meaning that much information (e.g., the general topology of the network) was missing in the final representations. As I could understand from my experiments, this was mainly because the obtained similarity scores were not interpretable beyond a certain distance, which I found relatively short. This meant that similarity scores between distant terms held no meaning, and thus they could be removed from the analysis. Certain recent dimensionality reduction techniques such as UMAP [MHSG18] can retain global distances and may be worth looking at for future developments in this area. Finally, and once the aforementioned issues are resolved, I think there are many interesting research opportunities related to the representation of hierarchies and sets [LGS<sup>+</sup>14, Ped17] of keywords, which would enhance how a user understands the language implied by the authors of a corpus.



## Chapter 2

# Resumen en Español de las Contribuciones

*¡Que inventen ellos!*

Miguel de Unamuno

### 2.1 Contribución #1

A. Benito-Santos and R. Therón Sánchez, ‘A Data-Driven Introduction to Authors, Readings and Techniques in Visualization for the Digital Humanities’, IEEE Computer Graphics and Applications, 2020.

#### 2.1.1 Resumen

La frontera recientemente redescubierta entre la visualización de datos y las humanidades digitales ha demostrado ser un campo de experimentación apasionante para los académicos de ambas disciplinas. Esta fructífera colaboración está atrayendo a investigadores de otras áreas de la ciencia que interesados en crear nuevas herramientas visuales que faciliten la investigación en humanidades en sus múltiples formas. Sin embargo, a medida que esta colaboración se hace más compleja, la tarea de adentrarse en la disciplina puede resultar intimidante para estos académicos. Para facilitar esta necesaria tarea de inmersión, en este artículo se propuso una introducción *dirigida por los datos* a la visualización aplicada a las humanidades digitales. Para construir un conjunto de datos representativo de la disciplina, se analizaron las citas de un corpus semilla de 300 publicaciones en visualización de las humanidades

obtenida de ediciones recientes del taller *Vis4DH*, la Conferencia *Digital Humanities* de la Alianza de Organizaciones de Humanidades Digitales (ADHO), y la revista especializada *Digital Humanities Quarterly*. De ellas, se extrajeron más de 1900 trabajos referenciados que se analizaron en busca de patrones temáticos, autores relevantes, y otras ideas interesantes. Finalmente, y siguiendo el camino trazado por otros investigadores en las comunidades de visualización e interacción humano-ordenador (HCI), se propuso un análisis de las palabras clave para identificar temas clave y otras oportunidades de investigación en el campo objeto del estudio.

### 2.1.2 Objetivos

En general, los objetivos principales de esta investigación fueron dos:

- Realizar un mapeo sistemático de la literatura relevante en visualización aplicada a las humanidades digitales.
- Construir un conjunto de metadatos obtenido a partir de las publicaciones analizadas que sirviera como base para otros estudios futuros.

En concreto, las preguntas de investigación que se plantearon al inicio fueron las siguientes:

- RQ.1. ¿Cuáles son las publicaciones y autores más influyentes en la disciplina?
- RQ.2. ¿Cuánto tiempo cubre la *memoria collectiva* de la comunidad y como está distribuida temporalmente?
- RQ.3. ¿Qué conceptos generan un número significativamente más alto de publicaciones y citas?
- RQ.4. ¿Cuáles son las temáticas principales en la práctica de la visualización aplicada a las humanidades?
- RQ.5. ¿Cómo se relacionan entre sí estas temáticas?

### 2.1.3 Metodología

Uno de los principales escollos que hubo que salvar al plantear la investigación descrita en este artículo fue la imposibilidad de adoptar técnicas tradicionales para el mapeo de la literatura debido al carácter difuso y multidisciplinar de la disciplina que se quería estudiar. Este hecho hizo que no fuese posible obtener un conjunto

de palabras clave con el que buscar artículos relevantes en bases de datos online y dar comienzo a la investigación. Además, en muchos casos, las publicaciones en el área de humanidades no se encuentran indexadas en dichas bases de datos y, por tanto, de haberse excluido del análisis, el estudio habría resultado incompleto. Por ello, se decidió comenzar la investigación a partir de colectivos de investigadores que se auto-identificasen como involucrados en la práctica de la visualización y las humanidades digitales capturando, de esta manera, ambas partes de la colaboración en una visión unitaria y completa de la disciplina. Gracias a esta modificación, se pudieron identificar alrededor de 300 publicaciones de autores en las áreas de ingeniería y humanidades digitales que concordaban con la descripción aportada, a las que se denominó *conjunto semilla*. Seguidamente, se extrajeron de manera semi-supervisada más de 1900 citas empleadas en dicho conjunto semilla, a lo que se llamó *conjunto de referencias*. Los metadatos obtenidos (título, año de publicación, lista de autores y lista de palabras clave) de ambos conjuntos fueron normalizados, lo que permitió plantear distintos tipos de análisis para resolver las preguntas de investigación planteadas. En concreto, se propusieron cuatro tipos de análisis bien diferenciados para responder las distintas preguntas de investigación formuladas al inicio del estudio: en concreto, se empleó un análisis de la frecuencia de citación usando los metadatos normalizados extraídos de cada cita para responder a la pregunta RQ1. Para la pregunta RQ2, se realizó un análisis temporal empleando el metadato "año de publicación" de cada cita. Además, se empleó un análisis de la frecuencia de citado de cada palabra clave normalizada, lo que permitió responder a las preguntas de investigación tercera y cuarta (RQ.3 y RQ.4). Finalmente, se completó el estudio con un análisis de la correlación entre las palabras clave apoyado en tres artefactos de software principales: 1. clustering jerárquico, 2. diagrama estratégico, y 3. red de correlación, que ayudaron a completar las respuestas para las preguntas de investigación cuarta y quinta (RQ.4 y RQ.5).

#### 2.1.4 Resultados

Además de la construcción y publicación de un conjunto de metadatos de publicaciones en el área obtenido mediante la metodología especificada en el anterior apartado, los resultados principales de esta investigación fueron varios: primero, se obtuvo una visión general del estado de la disciplina que puede ser de utilidad para la comunidad investigadora. Segundo, se propusieron distintos tipos de análisis que permitieron conocer publicaciones relevantes para quien quiera iniciarse en este tipo de práctica de la visualización. Además, se planteó un análisis temporal de los pa-

trones de citación que permitió verificar que existen importantes coincidencias entre las comunidades de humanistas e ingenieros en este aspecto. Finalmente, se pudieron extraer y relacionar entre sí las distintas áreas temáticas que forman la práctica de la visualización para las humanidades digitales, un descubrimiento que fue pionero en este aspecto.

### 2.1.5 Conclusiones

En este trabajo, se presentó un mapeo de la literatura en el campo de la visualización para las humanidades digitales. La combinación de técnicas de análisis bibliométrico con técnicas no supervisadas de análisis de datos permitió capturar el estado actual de la disciplina, a la vez que evitó caer en problemas típicos de metodologías tradicionales.

## 2.2 Contribución #2

A. Benito-Santos and R. Therón, ‘Pilaster: A Collection of Citation Metadata Extracted From Publications on Visualization for the Digital Humanities’, presented at the 5th Workshop on Visualization for the Digital Humanities, colocated to IEEEVIS 2020, Oct. 2020.

### 2.2.1 Resumen

En este artículo se presentó *Pilaster*, una colección de metadatos extraída de publicaciones científicas sobre visualización para las humanidades digitales. La colección se generó a partir de un conjunto semilla de publicaciones relevantes del cual se trajeron trabajos citados que incluyeron artículos presentados en conferencias o publicados en revistas, libros, tesis, o recursos online, entre otros. El trabajo se orientó alrededor de tres puntos fundamentales: primero, el de que la colección sirviese como punto de entrada a la disciplina para humanistas digitales y expertos en visualización sin experiencia previa en el campo. Segundo, se buscó que sirviese como punto de encuentro para humanistas digitales y expertos en visualización más experimentados en este tipo de colaboraciones que busquen colaboraciones para desarrollar nuevas investigaciones. Por último, se pretendió que la colección sirviese también como punto de partida para futuros estudios orientados a comprender las particularidades de la investigación en visualización dirigida por el problema en éste y otros contextos.

### 2.2.2 Objetivos

Los objetivos que se quisieron alcanzar en este trabajo fueron los siguientes:

- Proporcionar una colección de publicaciones relevantes en el ámbito de la visualización para las humanidades digitales para facilitar el proceso de inmersión en la disciplina de investigadoras sin experiencia previa en el campo.
- Acelerar la identificación de nuevas preguntas de investigación y potenciales colaboradores en el ámbito de la visualización para las humanidades digitales.
- Desarrollar una base de código de ejemplo que pueda ser reutilizada por otras investigadoras para plantear sus propios estudios sobre los datos ofrecidos.
- Realizar un estudio comparativo entre las formas de publicación y los patrones de citación empleados por investigadoras en ambos extremos de la colaboración.

### 2.2.3 Metodología

En este estudio, se completaron los resultados obtenidos en la contribución #1 de distintas maneras que se explican a continuación. En primer lugar, se dieron más detalles sobre las causas que motivaron el tipo de análisis de la literatura propuesto. En concreto, se puso de manifiesto la falta de adecuidad de las metodologías de mapeo de la literatura tradicionales, que hicieron imposible su uso en este contexto. Para ello, se propuso una modificación de dichas metodologías que consistió en obtener un conjunto semilla de publicaciones relevantes a través de la identificación de los miembros de la comunidad a estudiar (visualización para las humanidades digitales), en contraposición a la identificación de palabras clave. Además, se emplearon métodos semiautomáticos para normalizar el conjunto de datos, lo que permitió llevar a cabo estudios comparativos entre los patrones de citación empleados por publicaciones centradas en las humanidades o en la ingeniería. Por último, esta normalización también permitió realizar un análisis de coautoría de dichas publicaciones, lo que aportó información sobre cómo se articulan las colaboraciones en este campo.

### 2.2.4 Resultados

Los resultados presentados en este trabajo fueron varios: primero, se creó un sitio web para albergar los datos y el código de ejemplo obtenidos, que se pusieron a disposición de la comunidad. Segundo, se realizaron varios estudios comparativos entre las

publicaciones obtenidas en ambos extremos de la colaboración. Por ejemplo, se observó la tendencia de las publicaciones originadas en el dominio de las humanidades a emplear menos referencias debido a diferencias importantes en los formatos de publicación. Gracias a este tipo de análisis comparativo, se pudieron identificar diferencias también entre las revistas más citadas en uno y otro extremo, lo que permitió identificar áreas para posibles colaboraciones futuras. Finalmente, el análisis de coautoría puso de manifiesto el carácter novel de la disciplina: sólo una mínima parte de los autores identificados participa activamente en conferencias especializadas en ambos dominios (ingeniería y humanidades). Además, se pudo comprobar que el tamaño de los grupos de colaboradores en el ámbito de las humanidades es notablemente más pequeño que en el ámbito de la ingeniería.

### 2.2.5 Conclusiones

En este artículo, se presentó una colección de metadatos de artículos y citaciones originados en el ámbito de la visualización para las humanidades digitales. Para crear dicho conjunto, se hubo de modificar las metodologías existentes para el mapeo de la literatura debido al carácter difuso y novel de la disciplina. Además, se presentó el trabajo a la comunidad de investigadoras en visualización para las humanidades digitales, y se les ofreció también el conjunto de datos y el código empleado para generar todas las figuras del artículo.

## 2.3 Contribución #3

A. Benito-Santos and R. Therón Sánchez, ‘Cross-domain Visual Exploration of Academic Corpora via the Latent Meaning of User-Authored Keywords’, IEEE Access, vol. 7, pp. 98144–98160, 2019.

### 2.3.1 Resumen

Hoy en día, investigadoras de todo el mundo dedican una parte sustancial de su trabajo a la consulta y navegación de colecciones cada vez más grandes de artículos de investigación en Internet. Paralelamente, el reciente surgimiento de nuevos enfoques interdisciplinarios requiere que estas personas adquieran competencias en nuevos campos para los que pueden carecer del vocabulario necesario para formular las consultas adecuadas en dichas bases de datos. Este problema, junto con el problema de la sobrecarga de información, plantea nuevos desafíos en los campos

del procesamiento del lenguaje natural y el diseño en visualización que requieren una respuesta rápida por parte de la comunidad científica. En este sentido, en este artículo se propuso un nuevo esquema de visualización que permite la exploración de colecciones de artículos de investigación a través del análisis de las relaciones de proximidad semántica distribucional encontradas en palabras clave asignadas por los autores a sus propios artículos. El método propuesto permite reemplazar las consultas basadas en cadenas con una bolsa de palabras extraída de un corpus auxiliar generado por el usuario que sirve para capturar la intencionalidad de la investigación. Continuando en la línea establecida por otros autores en los campos del descubrimiento basado en la literatura, procesamiento del lenguaje natural y analítica visual, se combinaron avances recientes en dichos campos con técnicas de análisis visual de redes para ofrecer al usuario una visión del corpus objetivo que se adapta a los intereses particulares de su investigación. Para ejemplificar las ventajas de la propuesta, se llevaron a cabo dos experimentos que emplearon una colección de artículos científicos sobre visualización y una bolsa de palabras extraída de un corpus auxiliar sobre visualización aplicada a las humanidades digitales. En estos ejemplos, se mostró cómo el esquema de visualización propuesto se puede utilizar para maximizar la efectividad de una sesión de navegación mejorando la tarea de adquisición de vocabulario, que en última instancia permite extraer de manera efectiva conocimiento acorde a las expectativas iniciales del usuario.

### 2.3.2 Objetivos

- Desarrollar un modelo para la detección automática de asociaciones interesantes de conocimiento entre dominios basado en el análisis de las palabras clave.
- Identificar formas óptimas de representación de conocimiento en el contexto de la exploración de un corpus de publicaciones científicas por parte de un investigador interdisciplinario en visualización.
- Proveer ejemplos de los métodos obtenidos en el dominio de la visualización de datos para las humanidades digitales.

### 2.3.3 Metodología

En este artículo, que resultó ser clave para el desarrollo de esta tesis doctoral, se hizo hincapié en desarrollar técnicas de análisis semántico de las palabras clave. A través

de un estudio inicial sobre la distribución de las mismas, se pudo proponer un método para cuantificar similitudes distribucionales que dirigieran una hipotética exploración de un corpus. Este método consistió en la generación de un espacio vectorial en el que cada palabra clave fue representada por un vector multidimensional. Siguiendo principios documentados en la literatura del análisis del lenguaje natural, se pudo obtener una matriz de similitudes entre dichos vectores, que se exploró mediante una técnica de escalado psicométrico inspirada en las redes *pathfinder*. La aplicación de la técnica concebida en dicha matriz, generó caminos de exploración que fueron presentados usando representaciones de redes dirigidas por fuerzas, en las que se proyectaron además los documentos para lograr una visualización unitaria y efectiva.

### 2.3.4 Resultados

Los métodos propuestos se exemplificaron empleando palabras clave obtenidas de publicaciones sobre visualización para las humanidades digitales. En concreto, se propusieron dos casos de uso en los que relacionaron temáticas propias de las humanidades (como la visualización de textos de obras de teatro o el análisis visual de eventos históricos geográficamente localizados) con técnicas de visualización originadas en otros dominios. La visualización propuesta motivó un aprendizaje progresivo de dichas técnicas, en las que se partió de conceptos conocidos para el usuario para llegar a los desconocidos de manera incremental y paulatina, lo que tiene un efecto positivo en la adquisición de conocimiento.

### 2.3.5 Conclusiones

En este artículo, se describió un método automático para visualizar un proceso de exploración de documentos dirigido por los principios del descubrimiento científico basado en la literatura. El método propuesto permite a los usuarios explorar palabras clave y documentos relacionados en dos corpus disjuntos de artículos de investigación. La inspección de estructuras locales en datos de proximidad obtenidos de un espacio vectorial generado por las palabras clave, permitió prescindir del uso de las mismas para comenzar la exploración, lo cual supone una solución al problema de la falta de vocabulario en la investigación en visualización dirigida por el problema.

## 2.4 Contribución #4

A. Benito-Santos and R. Therón, ‘GlassViz: Visualizing Automatically-Extracted Entry Points for Exploring Scientific Corpora in Problem-Driven Visualization Research’, presented at the 2020 IEEE Visualization Conference (VIS), Oct. 2020.

### 2.4.1 Resumen

Este artículo describe el desarrollo de un modelo y una prueba de concepto de una herramienta para el análisis visual de textos para apoyar el descubrimiento de textos científicos en el contexto de investigación en visualización dirigida por el problema (PDVR). El modelo propuesto captura el modelo cognitivo adoptado típicamente por investigadores en este área mediante el análisis del canal de comunicación interdisciplinaria representada por palabras clave encontradas en dos conjuntos disjuntos de artículos científicos. La detección de similitudes distribucionales significativas entre las mismas se empleó para construir puntos de entrada para dirigir la exploración de un corpus científico de gran tamaño. La idoneidad del enfoque propuesto se demostró en el contexto de investigación en visualización aplicada a las humanidades digitales.

### 2.4.2 Objetivos

El diseño de la interfaz resultante de este estudio fue dirigido por los siguientes objetivos y preguntas de investigación:

- Motivar una exploración personalizada de un corpus de artículos científicos adaptada a los objetivos de investigación del usuario. *¿Qué clase de conocimiento quiere extraer el usuario del corpus? ¿Qué puede aprender el usuario del corpus que le sea útil para resolver un problema en un dominio particular?*
- Potenciar el descubrimiento de metodologías susceptibles de ser transferidas desde otros espacios del diseño a un dominio de aplicación en particular. *¿Cómo se puede medir el grado de transferibilidad de estas metodologías?*
- Acelerar la comprensión de los contenidos de un corpus y la adquisición de vocabulario por parte de un investigador en visualización dirigida por el problema. *¿Cuáles son los mejores términos para describir un corpus de acuerdo al nivel de experiencia y conocimientos del usuario? ¿Qué temáticas de las*

*contenidas en el corpus resultan más importantes para el usuario? ¿Cómo pueden ser presentadas para mejorar su comprensión?*

- Ofrecer un orden claro de lectura para los documentos descubiertos. *¿Qué documents de la colección son más interesantes para el usuario?*

#### 2.4.3 Metodología

La metodología empleada en esta contribución se basó en la adaptación y refinamiento de los artefactos software obtenido en las investigaciones previas de esta tesis, así como en la puesta en común de las mismas en una aplicación centralizada e interactiva. Para definir los requisitos y objetivos de diseño que orientaron el desarrollo de la misma, se emplearon también los conocimientos adquiridos sobre investigación en visualización dirigida por el problema en el ámbito de la visualización para las humanidades digitales, que se vieron materializados en la creación de un modelo aumentado que explica el problema de comunicación en este tipo de investigaciones interdisciplinares.

#### 2.4.4 Resultados

En este trabajo de investigación, se obtuvieron dos resultados principales: el primero hizo referencia al desarrollo de un modelo de transferencia de metodologías aumentado con conceptos extraídos del descubrimiento científico basado en la literatura, lo que permitió automatizar la detección de metodologías candidatas a ser transferidas, así como también calcular la calidad de las mismas. Para ello, se empleó el modelo de similitud distribucional para palabras clave desarrollado en la anterior investigación de esta tesis. El modelo obtenido sirvió para desarrollar una herramienta de analítica visual de textos orientada a la exploración de documentos científicos que cumpliese con los objetivos mencionados más arriba. Para exemplificar las ventajas de la propuesta, se emplearon palabras clave extraídas del conjunto de publicaciones obtenido en las dos primeras contribuciones de la tesis. El método propuesto se basó en la inspección cualitativa de vecindarios centrados en conceptos encontrados exclusivamente en dicho conjunto, a los que se denominó puntos de entrada. Estos puntos de entrada sirvieron para introducir al usuario a la navegación de un conjunto masivo de artículos científicos sobre visualización (vispubdata) que también se usó en la contribución #3. En total, el método propuesto identificó 12 puntos de entrada que se mostraron en la vista principal de la interfaz. Además se ofrecieron dos vistas auxiliares que permitieron explorar los contextos más comunes de aparición para los

términos de cada punto de entrada, así como documentos relevantes.

#### 2.4.5 Conclusiones

En este artículo, se presentó un modelo y una prueba de concepto para una herramienta de analítica visual de textos científicos orientadas a facilitar la exploración de una base de datos de artículos científicos en visualización de propósito general. Durante el desarrollo de la investigación, se identificaron también ciertas limitaciones que se discuten en el artículo: en concreto, se mencionaron algunas de las desventajas ligadas al uso del algoritmo de estemizado, que producieron algunos falsos positivos difíciles de detectar por medios automáticos. Aunque se consideró que las ventajas aportadas por el uso de dicho algoritmo eran mayores y más importantes que dichas desventajas, se discutieron posibles alternativas para la detección de dichos casos por el usuario. Además, se discutió también la falta de interactividad ofrecida por la interfaz a la hora de manipular los parámetros internos empleados por los algoritmos utilizados, que teóricamente podrían ser resueltos aplicando técnicas de manipulación directa. Finalmente, también se hizo mención a la dificultad de interpretar algunas temáticas debido al proceso de tokenización al que se sometieron las palabras clave. Esto motivó que los usuarios hubieran de reconstruir términos compuestos usando las partes estemizadas de los mismos, algo que se identificó como subóptimo por parte de algunos revisores.

### 2.5 Contribución #5

A. Benito-Santos and R. Therón Sánchez, ‘Defragmenting Research Areas with Knowledge Visualization and Visual Text Analytics’, Applied Sciences, vol. 10, no. 20, Art. no. 20, Jan. 2020.

#### 2.5.1 Resumen

La creciente especialización de la ciencia está motivando la fragmentación de áreas de investigación bien establecidas en comunidades interdisciplinarias centradas en la cooperación con expertos con el objetivo de resolver problemas en una amplia gama de dominios. Este es el caso de la investigación de visualización dirigida por el problema, en la que grupos de académicos utilizan técnicas de visualización en diferentes dominios de aplicación como las humanidades digitales, la bioinformática, las ciencias del deporte o la seguridad informática. En este artículo, se emplearon

algunos de los hallazgos descubiertos durante el desarrollo de una nueva herramienta de análisis visual para la exploración de colecciones de textos científicos, GlassViz, para detectar automáticamente asociaciones de conocimiento interesantes y grupos de intereses comunes entre estas comunidades de práctica. El método propuesto se basa en el modelado estadístico de palabras clave para realizar sus hallazgos, que se demostraron en dos casos de uso. Los resultados muestran que es posible proponer enfoques visuales interactivos semi-supervisados basados en el análisis automático de textos con el objetivo de desfragmentar un área de investigación.

### 2.5.2 Objetivos

- Dar una caracterización de las áreas de aplicación más conocidas de investigación en visualización dirigida por el problema, reparando en sus coincidencias y diferencias.
- Probar y adaptar los métodos concebidos en investigaciones previas para motivar la detección automática de asociaciones de conocimiento entre múltiples subdominios y atajar el problema de fragmentación.
- Generar distintas métricas para evaluar la calidad del modelo de similitud distribucional de investigaciones previas en el contexto del problema descrito.
- Proponer una herramienta para el análisis visual de la fragmentación de un área de conocimiento, que se exemplificará en el dominio de la investigación en visualización.

### 2.5.3 Metodología

Para medir el grado de fragmentación de la disciplina y proponer un método para la detección de interes comunes compartidos por distintas áreas de investigación interdisciplinar en visualización, se construyeron tres conjuntos de palabras clave siguiendo la misma filosofía que se usó en el caso de las humanidades digitales. Concretamente, se obtuvieron tres corpus de publicaciones en los ámbitos de la visualización aplicada a datos biológicos, seguridad informática y ciencias del deporte. Seguidamente, dichos conjuntos fueron tokenizados y estemizados siguiendo un procedimiento análogo al de investigaciones anteriores.

#### 2.5.4 Resultados

Un primer descubrimiento motivado por el análisis cuantitativo de los resultados de este proceso fue que todas las comunidades emplearon conjuntos de palabras clave de similar longitud para resumir sus artículos, que se sitúa entre cuatro y cinco palabras clave por artículo. Para medir el grado de solapamiento entre dichos conjuntos, se calcularon métricas de Jaccard para cada par, y se midieron posibles interacciones entre esta métrica y las similitudes obtenidas entre las palabras de dichos conjuntos. El estudio de esta circunstancia no arrojó datos concluyentes que permitieran probar que existe dicha interacción, indicando posiblemente que las similitudes obtenidas por el modelo dependen más de la manera en la que ciertas palabras son combinadas por los autores que en el grado de solapamiento de los términos. El análisis posterior de la distribución de la longitud de los caminos más cortos encontrados entre las distintas colecciones permitió establecer un punto de corte para considerar sólo las asociaciones a-priori más interesantes en un posterior análisis visual. En dicha exploración visual, se encontraron distintas coincidencias temáticas entre los dominios estudiados, que apoyaron la teoría inicial de que es posible desfragmentar un área de conocimiento con métodos visuales y automáticos basados en texto.

#### 2.5.5 Conclusiones

En este trabajo de investigación, se presentó una propuesta para detectar de manera automática, a través del análisis de palabras clave, intereses comunes entre distintas comunidades de investigación en visualización dirigida por el problema. El enfoque se basó en la construcción y puesta en común de cuatro conjuntos de datos representativos de cada comunidad analizada. Gracias a esto, se pudo demostrar que existe evidencia de que se pueden proponer métodos válidos, automáticos e interactivos basados en texto que tengan el objetivo de reunificar áreas de investigación fragmentadas. A la vista de los resultados, creemos que nuestro enfoque puede ser aplicado a otras áreas distintas a la visualización de datos. Por tanto, esperamos que el trabajo de investigación presentado pueda servir para inspirar futuros estudios que tengan como objetivo reducir el problema de fragmentación en la ciencia moderna.



## Appendix A

# Copy of the Contributions

*The subliminal self, Poincaré said, looks at a large number of solutions to a problem, but only the interesting ones break into the domain of consciousness.*

Robert M. Pirsig - Zen and the Art of Motorcycle Maintenance

### A.1 Contribution #1

A. Benito-Santos and R. Therón Sánchez, ‘A Data-Driven Introduction to Authors, Readings and Techniques in Visualization for the Digital Humanities’, IEEE Computer Graphics and Applications, 2020.

# A Data-Driven Introduction to Authors, Readings, and Techniques in Visualization for the Digital Humanities

Alejandro Benito-Santos and Roberto Therón Sánchez  
University of Salamanca

**Abstract—The newly rediscovered frontier between data visualization and the digital humanities has proven to be an exciting field of experimentation for scholars from both disciplines. This fruitful collaboration is attracting researchers from other areas of science who may be willing to create visual analysis tools that promote humanities research in its many forms. However, as the collaboration grows in complexity, it may become intimidating for these scholars to get engaged in the discipline. To facilitate this task, we have built an introduction to visualization for the digital humanities that sits on a data-driven stance adopted by the authors. In order to construct a dataset representative of the discipline, we analyze citations from a core corpus on 300 publications in visualization for the humanities obtained from recent editions of the InfoVis Vis4DH workshop, the ADHO Digital Humanities Conference, and the specialized digital humanities journal *Digital Humanities Quarterly*. From here, we extract referenced works**

*Digital Object Identifier 10.1109/MCG.2020.2973945*

*Date of publication 14 February 2020; date of current version  
28 April 2020.*

**and analyze more than 1900 publications in search of citation patterns, prominent authors in the field, and other interesting insights. Finally, following the path set by other researchers in the visualization and Human–Computer Interaction (HCI) communities, we analyze paper keywords to identify significant themes and research opportunities in the field.**

■ **THE COLLABORATION BETWEEN** the digital humanities (DH) and the data visualization communities has grown larger in recent years. This fact is attracting scholars from both areas of knowledge who are keen on designing tools that can reveal insight on humanistic data in an increasingly broader range of disciplines.

However, precisely due to its novelty and inherent interdisciplinary character, this collaboration often is hard to articulate, as it poses very particular challenges in the visualization design process and the construction of shared design spaces.<sup>1,2</sup> For these reasons, the scholars' initial excitement may soon become disenchantment if these challenges are not addressed from the very initial stages of the collaboration. In this work, we attempt to provide new researchers with an interest in the field with a series of recommended readings, authors, and terminology derived from a meta-analysis of the discipline's current state in a very concise yet effective manner.

In this regard, we hope our work succeeds at the task of orienting interdisciplinary visualization researchers, and that the contents of this article can lead them to examples, best practices, and resources to ease the production of future quality research on visualization for the humanities.

In order to produce a critical summarization of a scholarly field, it is a recurring first step in mappings studies, surveys, and literature reviews to invest time in clearly defining what exactly is to be considered in the study. Once a definition of the subject of the study has been reached, the researcher employs it to systematically retrieve publications from a selection of sources (e.g., online scientific databases or search engines) that are further analyzed at later stages. However, defining the DH is a challenging task that inevitably builds on rather shaky epistemological grounds, and therefore it has been (and still is) the subject of important discussions in the community. For example, the 2012 edition of *Debates in the Digital Humanities* accounted for 21 definitions of the DH alone.<sup>3</sup> Indeed, some authors argue that this

continuous process of questioning the self-identify is one of the core values of the DH, and therefore this question may never be resolved. For these reasons, producing a definition of "visualization in the DH," that satisfied both humanities and visualization scholars at the same time seemed overwhelming to us. Not only this, but capturing this definition into a textual query string that could be used to query an online scientific database to retrieve relevant publications was something that we wanted to avoid.

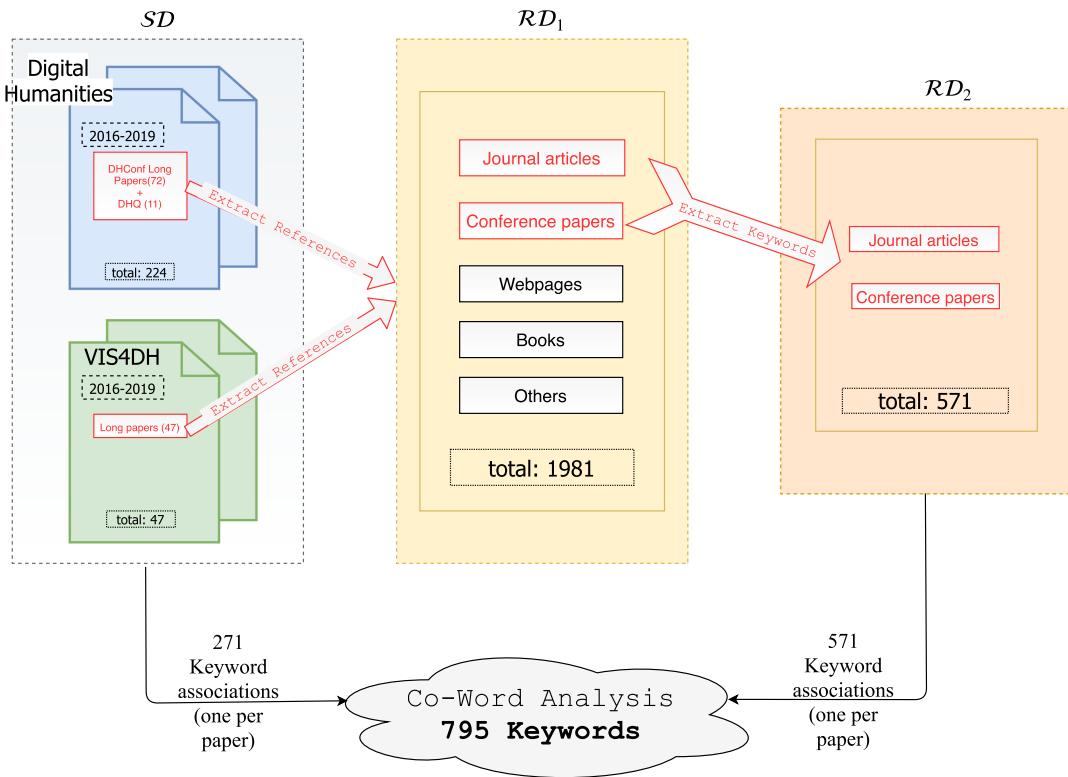
Instead, we decided to adopt a more practical stance to address this issue, which brought us to rely on data-driven, quantitative techniques that supported the foundations of this work. To this end, we analyzed visualization contributions to two core venues intimately related to visualization for the humanities: the Vis4DH InfoVis workshop\* and the ADHO Digital Humanities Conference†. From these works, we extracted referenced publications to construct a dataset of more than 1900 journal articles, conference submissions, books, and web pages (see Figure 1) to which we applied several bibliometric techniques to answer a set of research questions that we outline as follows:

- *RQ.1. What are the most influential...*
  - *RQ.1.1. publications?*
  - *RQ.1.2. authors?*
- *RQ.2. How long is the community's collective memory and how is it distributed in time?*
- *RQ.3. What are the concepts generating a more significant number of publications?*
- *RQ.4. What are the main themes in the DH Visualization practice?*
- *RQ.5. How do these themes relate to each other?*

## RELATED WORK

To answer the research questions that were proposed at the beginning of the study, and to gain insight into the discipline of visualization in the DH,

\*<http://www.vis4dh.org/>  
†<http://adho.org>



**Figure 1.** Construction process of the keywords dataset  $\mathcal{K}$  and the intermediate publication datasets originating at  $SD$ . The final result is a dataset of 1942 unique keywords related to the DH visualization practice.

our study relies on previous works in the visualization, human-computer interaction (HCI), and bibliometric domains that we introduce in this section.

#### Mapping Visualization

Many scholars have attempted to map the scientific landscape of visualization in different knowledge domains employing keywords.<sup>4</sup> Moreover, the task of understanding vast amounts of research papers is a longstanding HCI problem that has produced significant contributions in the past.<sup>5</sup> Regarding the mapping of the discipline, an important recent advance is the work by Isenberg *et al.*,<sup>6</sup> who compiled a dataset of visualization research papers presented at IEEE VIS (VisWeek) in the period 1990–2018. Since its publication, different visual solutions to explore the dataset have been proposed, ranging from the visualization of co-citation and co-authorship patterns<sup>7</sup> to the visualization of topic models,<sup>8</sup> or a combination of approaches based on network analysis and natural language processing (NLP) techniques.<sup>9</sup> More related to our study, the authors of the dataset performed co-word

analysis on the research paper keywords<sup>10</sup> that has greatly inspired our work.

#### Surveying Visualization in the DH

There exist notable previous attempts to produce reviews on visualization for the DH. For example, Jänicke *et al.*<sup>11</sup> evaluate past visualization approaches to support distant and close reading tasks on a variety of textual data. This review was later extended by Jänicke *et al.*<sup>12</sup> to include other kinds of text visualization. More recently, Windhager *et al.* review visual solutions to explore cultural heritage collections.<sup>13</sup> As opposed to our study, these works focus on specific subdomains of the DH practice, and therefore are not able to offer a complete view of the discipline.

#### Co-Word Analysis

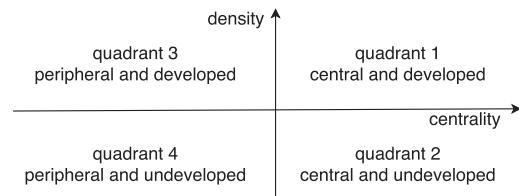
Co-word analysis is a bibliometric quantitative technique that is rooted in the idea that a paper's keywords are able to describe its contents correctly. Therefore, it can be assumed that the co-occurrence of keywords in a publication denotes a kind of implicit conceptual link between the ideas represented by such terms. The study of the

frequency patterns emerging from these links has been long employed to measure the development of science in a wide variety of knowledge domains such as chemistry, software engineering, consumer behavior, patent analysis, ubiquitous computing, library and information science, to name a few. In particular, the works by Liu *et al.*,<sup>14</sup> who successfully analyzed publication keywords in 20 editions of the CHI conference, and Isenberg *et al.*, with *vispubdata*<sup>6</sup> and their study on visualization keywords,<sup>10</sup> are good examples of the validity of co-word analysis as a tool to produce comprehensive studies on these areas.

### Strategic Diagrams

Strategic diagrams combine co-word with network and clustering techniques and have been typically employed to produce maps of the intellectual structure of a discipline in a variety of topics.<sup>10,14</sup> The process to generate these diagrams is straightforward: First, a network of keywords is generated employing different methods, which can be simple co-occurrence (two keywords are connected if they appear on the same paper) or correlation (two keywords are connected if they are positively correlated).<sup>10</sup> The network is then partitioned into clusters (or subnetworks), usually making use of unsupervised hierarchical clustering algorithms. For each of the resulting clusters, two key measures are calculated: *density* and *centrality*. The first measure, density, “characterizes the strength of the links that tie the words making up the cluster together”<sup>15</sup> and depicts the ability of a cluster to constitute a coherent and integrated whole, which can be understood as a measure of the theme’s development. Therefore, the higher the density of the links of the cluster, the more likely it is to contain inseparable expressions. The second measure, centrality, measures the strength and number of interactions of the cluster with other parts of the network and it is employed to quantify the importance of a theme in the research field under study. The more and stronger connections a cluster has, the more central the theme is in respect to the whole network.

The combination of these two concepts is then plotted in the *strategic diagram*, a two-dimensional representation of density (y-axis) and centrality (x-axis). The space is usually divided into four quadrants with separation lines corresponding to



**Figure 2.** Strategic diagram with its four main quadrants explained. The location of a cluster in the diagram characterizes the theme it represents in the context of the discipline.

the median density and centrality values of all the previously calculated clusters. This disposition is presented in Figure 2.

Below we provide details on how these areas usually are interpreted in the study of a given research field.

- *Quadrant 1* (see top-right of Figure 2): Internally coherent (high density) and central (strongly connected to other subnetworks) themes to the research network. These clusters are considered to be the “motor themes” of the discipline. They are dealt with systematically and over a long period, probably by a well-defined group of researchers.
- *Quadrant 2* (see bottom-right of Figure 2): Clusters in quadrant two are strongly connected to other clusters, but the density of their internal links is low. They are interpreted as connectors of other clusters or emerging themes that are starting to become central but have not yet been the object of a significant number of contributions.
- *Quadrant 3* (see top-left of Figure 2): These clusters are not well communicated with other parts of the network, but the strength of their internal links denotes research problems whose study is already well-developed. It is often the case that these clusters were central in the past, but their relative importance has decayed in recent times.
- *Quadrant 4* (see bottom-left of Figure 2): Within this category fall the clusters that are peripheral and underdeveloped. They are considered marginal in the global research network.

### DATASETS

A critical step of bibliometric studies is the selection of publications to consider. To this end,

researchers usually rely on online scientific databases from which this information is extracted via query strings. Query strings result from the application of a search strategy that is aligned with the aim of the study. There exist different methods to construct a query string in a systematic manner, although some authors have noted that they might be difficult to apply when the subject of the study is hard to define.<sup>16</sup> DH-specific publications are hard to find in scientific databases since many of them are not indexed (e.g., DH conference papers). All datasets collected in this study can be consulted in the supplementary materials, which are available in the IEEE Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/MCG.2020.2973945>.

Our study is based on a core set of publications in visualization for the humanities (*SD* for “Seed Dataset”) which, in turn, is twofold. First, it contains publications from the VisWeek Vis4DH workshop that represent the engineering/visualization community. Second, it also contains publications from the ADHO DH, which is completed with the addition of selected works from Digital Humanities Quarterly (DHQ). These two sources are meant to represent the humanities side of the community. The collection process is outlined as follows:

1) *Engineering Dataset:*

- a) *Vis4DH Workshop:* This workshop is a co-located event with IEEE VIS conferences that kicked off in 2016. The workshop, initially supported by visualization researchers with common experience in the DH, attracted stakeholders from different academic backgrounds in the humanities and science, promoting a series of publications, debates, and panel discussions framed under the particular interdisciplinary collaboration setting that is characteristic of the discipline.<sup>17</sup> Initially, a total of 38 publications published in the 2016 (17), 2017 (10) and 2018 (11) and 2019 (9) editions of the workshop were included in *SD*.

2) *Humanities Dataset:*

- a) *ADHO DH Conference:* The DH Conference is an annual event organized by the umbrella organization known as the Association of Digital Humanities Organizations

(ADHO). Due to the popularization and increasing availability of visualization techniques in recent years, there has been a great surge in the number of papers of this kind submitted to the conference.<sup>1</sup> In order to select a sufficient number of papers, we employed the following strategy: first, we downloaded the conference abstracts in the period 2016–2019 (4 editions, to overlap in time with the years the Vis4DH workshop has taken place). Then, we included all papers matching the regular expression “[Vv]isua \*” in their title, list of keywords, or list of topics. This resulted in 214 candidate contributions (see Figure in the supplementary materials, available online).

- b) *Digital Humanities Quarterly:* Following the same rationale as we did with publications on the DH conference, we included works from the DH-centric journal DHQ in our seed dataset. We included 15 extra works with this procedure for a total of 229 papers representing the humanities side in our seed dataset.

Before moving on to other sections of the paper, here we acknowledge some limitations related to the research methodology that was adopted in this work. For example, it is worth noting that, due to limiting publications in *SD* to only those appearing in the Vis4DH workshop, DH Conference and DHQ journal, we might have left out certain works that should have been initially included. Although we believe the citations dataset can (and should be enhanced by the community in the future: see for example the work by Isenberg *et al.*<sup>6</sup>), we believe the citation analysis captures a majority of relevant works for the DH practice that are good enough to propose an initial analysis.

By composing *SD* of a mix of humanities-related publications presented in a visualization conference and visualization-related papers presented in a DH venue, we ensured enough representative works of scholars pertaining to both areas of knowledge were included in the study while avoiding to employ a query search string, which would have been very difficult to construct, given the problematic previously presented in this article.

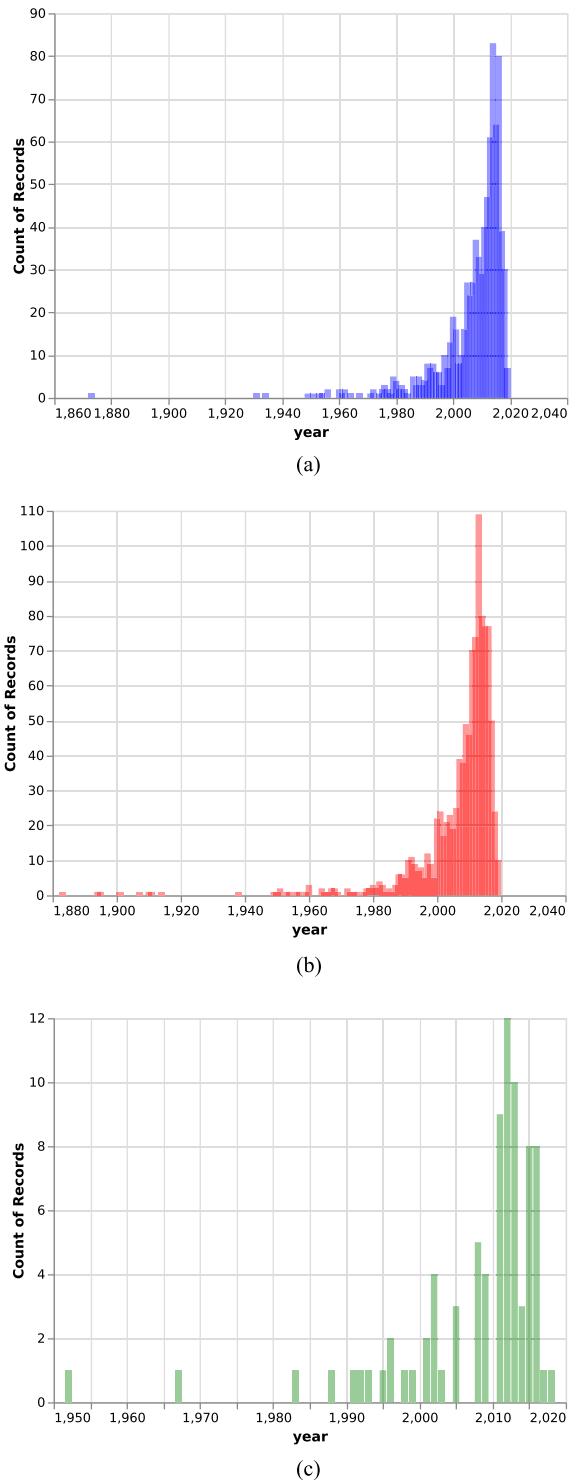
Additionally, we extracted all references found in long papers in *SD* to construct a new dataset,

$\mathcal{RD}_1$ , which originally contained 1981 referenced works (excluding self-references) including journal publications, conference papers, books/book chapters, webpages, blogposts, and others information. In total, we obtained 830 citations from works in the humanities subset of the seed dataset, whereas 1068 could be traced to works in contributions to any of the Vis4DH workshops. Eighty-three publications (4% of the total) were referenced from both subsets of publications. From this list of publications, we extracted author keywords (when applicable, note that some works in  $\mathcal{RD}_1$  are books or blog posts that do not contain author-assigned keywords), forming  $\mathcal{RD}_2$ , obtaining 571 papers or keyword associations (mainly from journal and conference papers). These 571 keywords were merged with those from the seed dataset (224), to base our co-word analysis in a total of 795 keywords.

#### Insights on Cited Publications

An analysis of the temporal distribution of the publications cited by works published on VIS or DH venues reveals very similar citation patterns, including works that go as far back as the last decades of the nineteenth century (see Figure 3). In Table 1 the top cited papers, up to rank 5, are displayed. The main themes found in these key works are: text visualization/distant reading, poetry visualization, graphs/network visualization, and visualization design theory and best practices.

Following a similar process, from  $\mathcal{RD}_1$  we extracted the most cited books (listed, up to rank 4, in Table 2). Unsurprisingly, in this publications set we can find Franco Moretti's pioneer works on distant reading *Graphs, Maps, Trees: Abstract Models for a Literary History* and *Distant Reading*, which supposed a turning point in the modern development of the DH. Also worth noting is Johanna Drucker's *Graphesis*, in which the author critically comments on different aspects of the DH from visualization and design theory perspective. Also, two classic data/information visualization books are shown at the bottom of the table, Card's *Readings in Information Visualization: Using Vision to Think* and Tufte's *The Visual Display of Quantitative Information*. Interestingly, these two volumes, which may sound more familiar to data visualization practitioners,



**Figure 3.** Temporal distribution of works cited from (a) vis/engineering publications in the seed dataset (blue), (b) humanities publications in the seed dataset (red), and (c) both (in green). The collective memory of both communities seems to follow a similar pattern in both cases. Interestingly, the oldest work cited by publications in the vis and humanities seed datasets is “The origins of intelligence in children” (1952) by J. Piaget.

**Table 1.** Top cited papers in dataset  $\mathcal{RD}_1$ .

#	Author	Title	Venue	Year
15	S. Jänicke <i>et al.</i>	On close and distant reading in digital humanities: A survey and future challenges	EuroVis	2015
10	B. Shneiderman	The eyes have it: A task by data type taxonomy for information visualizations	InfoVIS	1996
	J. Drucker	Humanities approaches to graphical display	DHQ	2011
8	A. Thudt <i>et al.</i>	The Bohemian bookshelf: Supporting serendipitous book discoveries through information visualization	CHI	2012
	N. McCurdy <i>et al.</i>	Poemage: Visualizing the sonic topology of a poem	TVCG	2016
6	A. Gibbs and T. Owens	Building better digital humanities tools: Toward broader audiences and user-centered designs	DHQ	2012
5	M. Whitelaw	Generous interfaces for digital cultural collections	DHQ	2015
	M. Dörk <i>et al.</i>	Critical InfoVis: Exploring the politics of visualization	CHI	2013
	U. Hinrichs	In defense of sandcastles: Research thinking through visualization	DH	2015
	S. Jänicke	Visual text analysis in digital humanities	EuroVis	2016

are the oldest in the listing. We hypothesize this fact may be due to a certain degree of stagnation in the DH visualization community and may be indicative of the need for novel techniques resulting from renovated visualization design processes conceived for the DH practice.

#### Keywords Dataset

As it has been explained before, the dataset of keywords that was used to perform the co-word analyses contains author-assigned keywords from publications in the seed and  $\mathcal{RD}_2$  datasets. As it is usual in this kind of approaches, we removed domain stopwords using the following regular expression: “(data—information).?visuali[sz]ation [s]?,” “visual analytics” and “digital humanities.” After removal, the 795 papers containing author-assigned keywords yielded a total of 2511 unique keywords, occurring 4015 times (5.05 author keywords per paper).

#### ANALYSIS PROCESS

In this section, we provide details on the calculations and algorithms that were applied to the keywords dataset obtained in the previous step in order to create the strategic diagram and the keywords network. All code was implemented in a *Jupyter Python* environment employing the libraries *nltk*, *pandas*, *bokeh*, and *networkx*.

#### Preprocessing of Keywords

To group keywords of similar themes, some authors have relied in the past on an expert coding of the keywords.<sup>10</sup> To accelerate the analytic process, we, instead, designed an automatic method that yielded similar quality results and also worked well in a smaller corpus such as ours. The procedure, which is well known in the NLP literature, involved the tokenization and stemming of keywords, in which the multiterm words are split into their constituent parts and reduced to their root form. We employed Porter’s stemming algorithm as it yielded the most satisfactory results. In a similar manner as we did with the original

**Table 2.** Top cited books in dataset  $\mathcal{RD}_1$ .

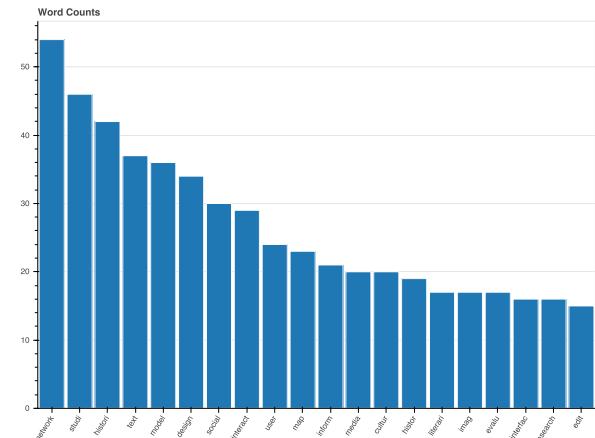
#	Author	Title	Year
13	F. Moretti	Graphs, maps, trees: Abstract models for a literary history	2005
7	F. Moretti	Distant reading	2013
6	M.L. Jockers	Macroanalysis: Digital methods and literary history	2014
5	J. Drucker	Graphesis: Visual forms of knowledge production	2014
4	E. R. Tufte and P. Graves-Morris	Graphesis: Visual forms of knowledge production	2012
	S. Rücke <i>et al.</i>	Visual interface design for digital cultural heritage	2011

**Table 3.** Three examples of hierarchy groups resulting from the stemming of keywords. In bold, tokens that were matched to an upper element of the hierarchy.

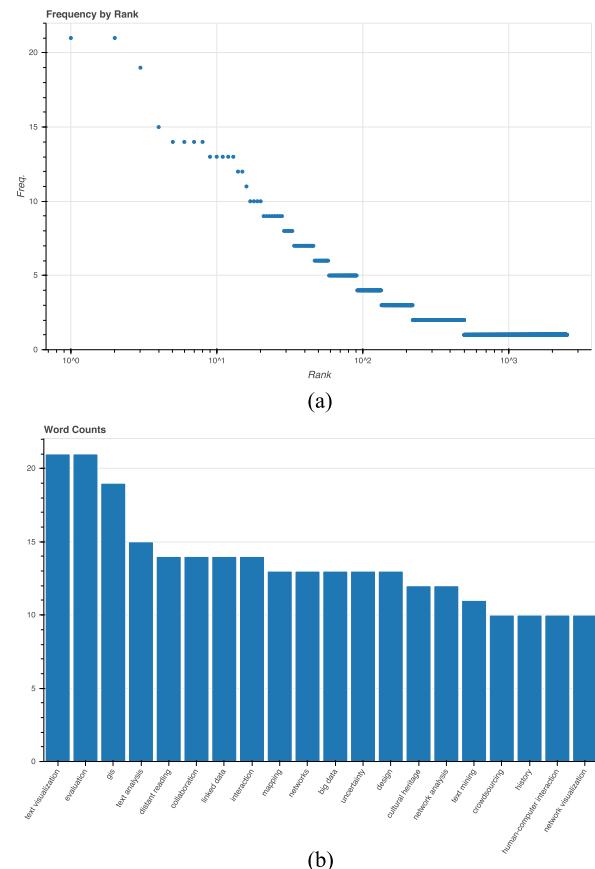
stem	keywords
american	<b>american culture, american history, american television, c19 american literature, nineteenth century american, wright american fiction corpus</b>
corpu	<b>cbeta corpus, corpus analysis, corpus analysis tool, corpus examples, corpus linguistics, corpus studies, corpus visualization, corpus workbench, diachronic corpus, n-gram corpus, wright american fiction corpus</b>
cultur	<b>american culture, cultural artefacts, cultural collections, cultural differences, cultural heritage/history, cultural probe, cultural studies, digital cultural heritage, online cultural heritage, personalized access to cultural heritage, popular culture, virtual cultural heritage, visual culture</b>

keywords, the following stems were also removed from the analysis: “analysi,” “digit,” “human,” “visual,” “analyt,” “dh,” “data,” “algorithm,” “comput” as they referred to generic elements of computer science and the domain under study. Furthermore, uninformative keyword stems, appearing less than six times were also removed. At the end of this process, all individual keyword tokens had been translated into their correspondent root form (see Table 3), yielding a total of 106 tokens (dataset  $\mathcal{K}$ ) that were employed to construct a correlation matrix of keywords.

Ultimately, the tokenization and stemming of the keywords modified the distributional model of the keywords (see Figure 4) in the corpus by organizing them in a *hierarchy*. This change is key to reveal insights that could not be reached from studying the distribution function from the previous situation (see Figure 5). For example, in the new situation it can be seen that the particle “network” has been promoted to the first position in the new distribution. However, in the previous case the occurrences of the term appeared in the 5th (“network analysis”), 8th (“networks”), and 13th (“network visualization”). This grouping promoted the term to the #1 frequency rank in the new distribution, highlighting the key role of “networks” as a transversal theme in the discipline. In a similar effect, the term “gis” is removed from the list of top words after preprocessing, giving way to the more general concept “map” that is now placed at rank #6.



**Figure 4.** 20 most common roots after stemming of the keywords. The stemming effectively changed the distributional model of the keywords, revealing different patterns to what could be observed in the prestemming situation (shown in Figure 5).



**Figure 5.** (a) Keywords frequency by rank (log-scaled), (b) 20 most common keywords. The observed distributional model seems to be in line with findings from similar studies.<sup>10,14</sup>

**Table 4.** Hierarchical cluster results for the  $\mathcal{K}$  dataset. Members are sorted by frequency, with the two most popular terms in bold.

ID	Members	N	#	cw#	centr.	dens.
1	<b>user, inform</b> , interfac, retriev, search	9	10.222	1.639	0.544	0.163
2	<b>languag, process</b> , natur	3	7	3.333	0.483	0.485
3	<b>imag, annot</b> , graphic, tool	4	10.5	2.5	0.682	0.239
4	<b>semant, link</b> , web	3	11.333	3.333	0.723	0.278
5	<b>studi, literari</b> , literatur, linguist, corpu	5	17.6	2.8	0.764	0.132
6	<b>recognit, relat</b> , extract, featur, name	5	4.2	1.4	0.225	0.325
7	<b>evalu, graph</b> , chart, multipl	4	7.5	1.333	0.413	0.322
8	<b>cultur, collect</b> , heritag, explor	4	12.75	3.167	0.749	0.222
9	<b>histor, ontolog</b> , place, servic, event	5	6.2	1	0.306	0.199
10	<b>text, mine</b> , vector, word	4	12.5	2.167	0.530	0.199
11	<b>model, edit</b> , topic, scholarli	4	15	2.833	0.698	0.181
12	<b>manag, databas</b> , plan, architectur, project	5	4.6	0.8	0.299	0.177
13	<b>design, research</b> , scienc, knowledg, technolog	7	11	1.14286	0.533	0.116
14	<b>mediev, align</b> , dynam, program	4	4.75	1.167	0.330	0.241
15	<b>histori, collabor</b> , art, archiv, learn	8	12.25	1.149	0.625	0.0796
16	<b>map, media</b> , spatial, 3 d, archeolog	9	10.111	0.889	0.585	0.069
17	<b>represent, classif</b> , narr, detect	4	3.75	0.667	0.181	0.182
18	<b>network, social</b> , commun, critic, cartographi	7	16	1.429	0.642	0.069
19	<b>interact, video</b> , uncertainti, document, method	16	6.25	0.142	0.356	0.011

## Correlation Matrix and Clustering

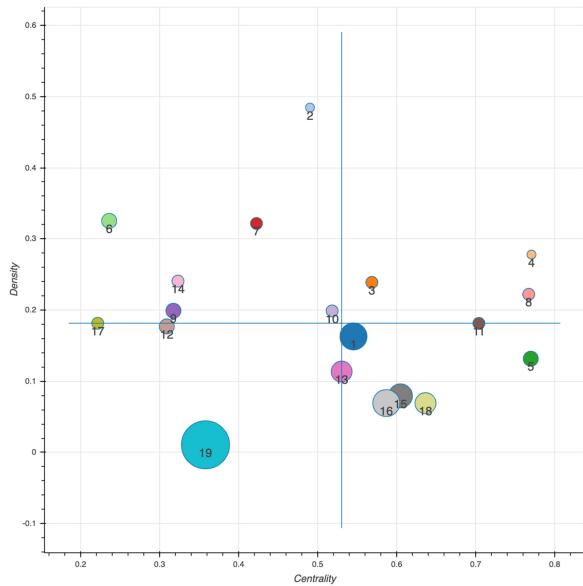
After we applied the preprocessing step outlined in the last section, we constructed a boolean document-term matrix in which we annotated when a certain token was contained in a document. After, we used this matrix to calculate a correlation matrix on the keywords. Finally, the keywords were hierarchically clustered using Ward's method and a squared Euclidean metric. Instead of relying in a predefined number  $k$ , we employed a maximum distance criterion to form clusters. Under this assumption, any two observations in a cluster shall not have cophenetic distance greater than 95% of the maximum total distance between two any two pairs in the dataset.

## RESULTS

In this section, we discuss, in light of research questions *RQ.4* and *RQ.5*, the keyword clusters,

network, and strategic diagram that were built following the procedure introduced in previous sections. In Table 4, we display the results of the semisupervised hierarchical clustering process that was applied to the keywords. For each cluster, we show the following.

- *Members*: The set of keyword stems that form the cluster. The two top keywords of each clusters are written in bold.
- *Size (N)*: The number of keywords that are in the cluster.
- *Frequency (F)*: Average frequency for all terms in the cluster.
- *Co-Word Frequency (CW-F)*: Average number of times any two given keywords of the cluster can be seen together in the documents collection.
- *Centrality*: Degree of the interaction of the cluster with any other parts of the network.



**Figure 6.** Strategic diagram for the 19 hierarchical clusters that were created. Blue lines indicate the medians.

It measures how well communicated the topic is with other themes. We calculated it as the average betweenness centrality using the standard value two for the  $K$ -step reach.

- **Density:** Topic's degree of internal cohesion. It measures how strong the connections between members of the same cluster are.<sup>15</sup> This is calculated as the average correlation between all member pairs in the cluster.

The 19 clusters are plotted in the strategic diagram of Figure 6, according to their centrality and density measures. Additionally, we complement this information with the keywords network (see Figure 7), which aims to highlight structural patterns and other interesting information not easily identifiable in the strategic diagram. The network links depict positively correlated pairs of keywords, which were obtained from the correlation matrix. Correlations  $\leq 0.20$  were omitted. The visualizations were created using the Python library Bokeh<sup>‡</sup>. The network layout employs *networkx*'s implementation of the Kamada–Kawai graph layout algorithm. In the network visualization, the circle sizes and edge thickness follow a logarithmic scale that is dependant on the term's frequency and correlation strength, respectively.

<sup>‡</sup><https://bokeh.pydata.org>

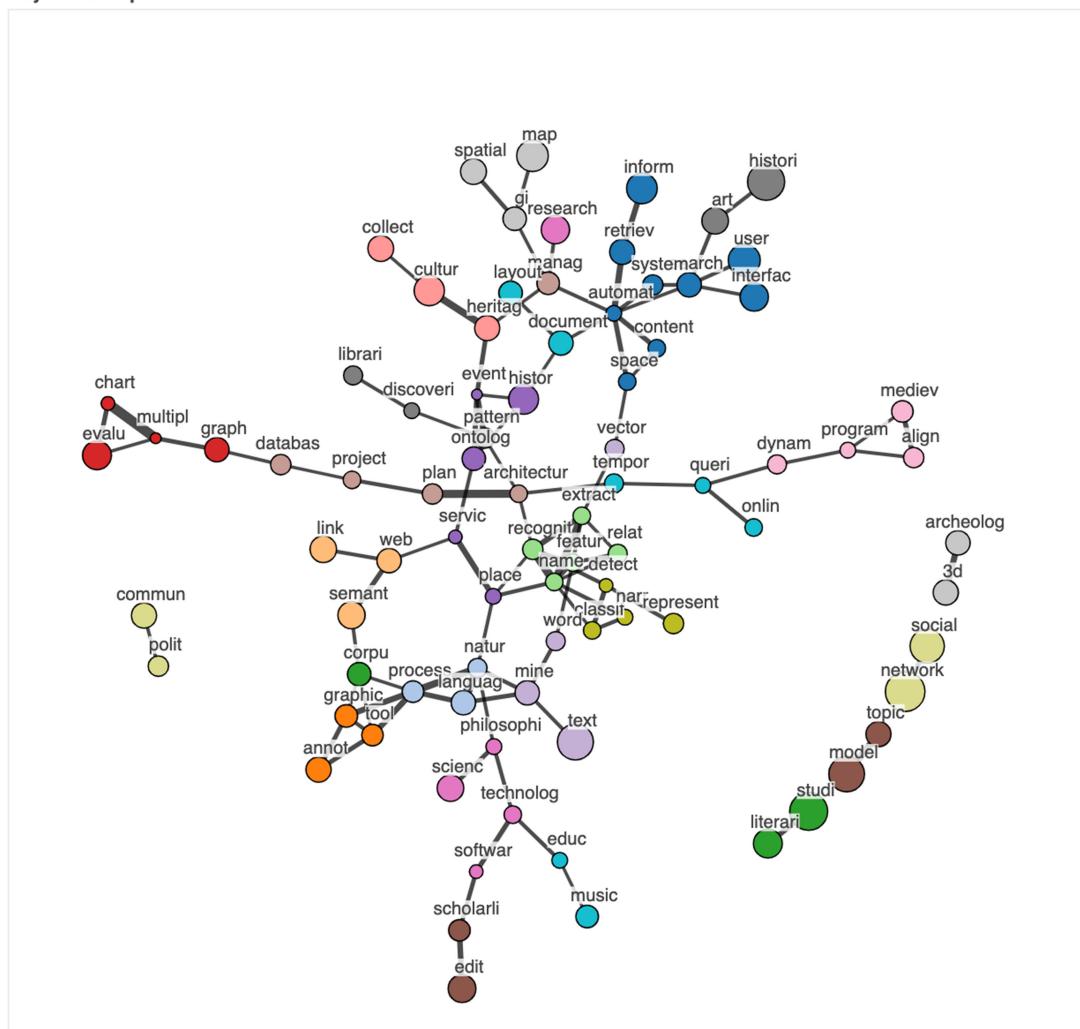
## DISCUSSION

The algorithm successfully organizes keywords in 19 main themes that were found in the corpus. Remarkably, the smaller clusters showing higher densities (and therefore appearing on the upper side of the strategic diagram) are easily interpretable. This can be observed for example in cluster 2 (“natural language processing”), a cluster that from its position in the graph seems to be of major importance in the discipline. In a similar manner, we can find cluster 4 (“semantic web” and “open linked data”) in the first quadrant. Cluster 10, that is placed right at the crossing of the medians can be easily interpreted as text analysis based on word embeddings, a discipline that has attracted much interest from the community due to its recent popularization. Perhaps in the short future, we will see novel techniques in the DH practice that employ more modern and powerful linguistic models beyond word2vec in a variety of DH research contexts beyond text summarization, such as translation of ancient languages or others. Cluster 19, the largest of it all appearing in quadrant 4, contains terms that are more difficult to relate. Interestingly, it catches our attention the word “uncertainty,” which is becoming a hot topic among data visualization practitioners in recent years. As it happens, two out of nine papers submitted to Vis4DH 2019 contained themes related to the management and display of uncertainty in visualization for the humanities, a trend that we are expecting to continue in forthcoming years.

Looking at the right of the chart, cluster 11 (topic models and scholarly editing software) appears to be a central and well-established theme in the discipline by looking at their position in the diagram. In Figure 7, it can be seen how topic models do not seem to be particularly attached to any other themes of the discipline, which means they maintain a relatively constant high correlation with other terms shown (at least 0.2). Therefore, it is reasonable to think that topic models are employed in a broad range of DH applications due to their summarization capabilities and close relationship to distant reading. We invite the reader to explore the dataset<sup>§</sup> using the visualization notebook<sup>¶</sup> set up for the purpose.

<sup>§</sup>[https://docs.google.com/spreadsheets/d/1TCnElbyow7s7\\_qnl\\_KZs4cUZjrt4bpz5C8VJLe-XIA/](https://docs.google.com/spreadsheets/d/1TCnElbyow7s7_qnl_KZs4cUZjrt4bpz5C8VJLe-XIA/)  
<sup>¶</sup><https://github.com/ale0xb/vis4dh-analysis>

Keywords Map



**Figure 7.** Keyword stems map. Edges represent a correlation strength of  $\geq 0.20$  between two nodes. Thicker links depict higher correlation values. Absolute keyword frequency was encoded in circle size.

## LIMITATIONS

Regarding the preprocessing step that was applied to the keywords, we acknowledge its capacity to explain the themes may be suboptimal when compared to other approaches employing an expert coding of the keywords. Therefore, although its analytical capacity may be inferior, it presents other advantages, such as a minimum time investment or unsupervised character, that may make it a good fit in a broad range of research contexts. Also, in relation to this technique, we are aware that the stemming algorithm that was applied to the data may introduce noise as nonexistent connections between concepts. This issue is hard to resolve since it is related to the disambiguation of terms that share the same root but they are semantically and etymologically different. While some authors in the NLP literature address

this by including part-of-speech tagging in the analysis, this can be very hard to achieve in case of stand alone expressions such as keywords. Therefore, it might be worth looking into alternative language/co-word models or even move to full-text kinds of analysis.<sup>18</sup>

Finally, some publications had to be excluded from the co-word analysis because the authors did not add keywords to their work, rendering this kind of analysis inadequate for these publications. Although automatic keyphrase/keyword extraction techniques exist in the literature, substantial more work is required to understand the implications of substituting human-generated keywords with their machine-generated counterparts, especially in the task of extracting knowledge from vast amounts of scientific literature.

## CONCLUSION

In this article, we have presented a systematic, data-driven approach to provide an introduction to an uncharted interdisciplinary research field as visualization for the DH. By combining numerical overviews with unsupervised data science and bibliometric techniques, we were able to capture the discipline's current state while avoiding common pitfalls of more traditional analysis workflows, which would have been hard to apply in this context. Furthermore, we share our dataset (accessible at [https://docs.google.com/spreadsheets/d/1TCnElfbyow7s7\\_qnl\\_KZs4cUZjrt4bpz5C8VJLe-XIA/](https://docs.google.com/spreadsheets/d/1TCnElfbyow7s7_qnl_KZs4cUZjrt4bpz5C8VJLe-XIA/)) with researchers who might be willing to use it and expand it in future research. Ultimately, we hope the findings of this research may be of help to humanities and visualization scholars stepping on such a vibrant and interesting discipline in the future.

## ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their helpful suggestions in improving the paper and Ines Holiday for the proofreading. This work was supported by the CHIST-ERA programme under National (MINECO-Spain) Grant PCIN-2017-064. This article has supplementary downloadable material at <http://ieeexplore.ieee.org>, provided by the authors.

## ■ REFERENCES

1. S. Jänicke, "Valuable research for visualization and digital humanities: A balancing act," in *Proc. 1st Workshop Vis. Digit. Humanities*, 2016, pp. 1–5.
2. K. Coles, "Show ambiguity: Collaboration, anxiety, and the pleasures of unknowing," in *Proc. 1st Workshop Vis. Digit. Humanities*, 2016, pp. 38–42.
3. M. K. Gold, *Debates in the Digital Humanities*. Minneapolis, MN, USA: Univ. Minnesota Press, 2012.
4. M. Meyer, T. Munzner, and M. Sedlmair, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2431–2440, Dec. 2012.
5. C. Dunne, B. Schneiderman, R. Gove, J. Klavans, and B. Dorr, "Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 12, pp. 2351–2369, 2012.
6. P. Isenberg *et al.*, "Vispubdata.org: A metadata collection about IEEE visualization (VIS) publications," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 9, pp. 2199–2206, Sep. 2017.
7. R. Vuillemot and C. Perin, "Investigating the direct manipulation of ranking tables for time navigation," in *Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst.*, Seoul, South Korea, 2015, pp. 2703–2706.
8. M. Abdelaal, F. Heimerl, and S. Koch, "CoTop: Visual topic-based analysis of scientific community structure," in *Proc. Int. Symp. Big Data Vis. Analytics*, Nov. 2017, pp. 1–8.
9. Z. Zhou, C. Shi, M. Hu, and Y. Liu, "Visual ranking of academic influence via paper citation," *J. Vis. Lang. Comput.*, vol. 48, pp. 134–143, Oct. 2018.
10. P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, "Visualization as seen through its research paper keywords," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 771–780, Jan. 2017.
11. S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann, "On Close and distant reading in digital humanities: A survey and future challenges," in *Eurographics Conference on Visualization (EuroVis)—STARs*, R. Borgo, F. Ganovelli, and I. Viola, Eds. Norrköping, Sweden: The Eurographics Association, 2015.
12. S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann, "Visual text analysis in digital humanities," *Comput. Graph. Forum*, vol. 36, no. 6, pp. 226–250, Sep. 2017.
13. F. Windhager *et al.*, "Visualization of cultural heritage collection data: State of the art and future challenges," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 6, pp. 2311–2330, Jun. 2019.
14. Y. Liu *et al.*, "CHI 1994–2013: Mapping two decades of intellectual progress through co-word analysis," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, New York, NY, USA, 2014, pp. 3553–3562.
15. M. Callon, J. P. Courtial, and F. Laville, "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry," *Scientometrics*, vol. 22, no. 1, pp. 155–205, Sep. 1991.
16. K. El-Arini and C. Guestrin, "Beyond keyword search: Discovering relevant scientific literature," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2011, pp. 439–447.
17. A. J. Bradley *et al.*, "Visualization and the digital humanities;," *IEEE Comput. Graph. Appl.*, vol. 38, no. 6, pp. 26–38, Nov. 2018.

18. A. Benito-Santos and R. Therón Sánchez, "Cross-domain visual exploration of academic corpora via the latent meaning of user-authored keywords," *IEEE Access*, vol. 7, pp. 98144–98160, 2019.

**Alejandro Benito-Santos** is currently a Research Assistant and Lecturer with the Department of Computer Science and Automation, University of Salamanca, Salamanca, Spain, which he joined in 2016. He received the B.Sc. degree in computer engineering and the M.Sc. degree in intelligent systems in 2016 from the University of Salamanca. He is a member of the Visual Analytics Group VisUSAL (within the Recognized Research Group GRIAL), where he is currently working toward the Ph.D. degree under the supervision of Dr. Roberto Therón Sánchez. 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 lie in the areas of human-computer interaction, design, statistics, and education. He has taught HCI and Introduction to Python Programming for Statisticians with the Faculty of Sciences of Salamanca in the past. He is a Student Member of IEEE since 2018. Contact him at abenito@usal.es.

**Roberto Therón Sánchez** is currently the Manager of the VisUSAL Group (within the Recognized Research Group GRIAL), University of Salamanca, Salamanca, Spain, which focusses on the combination of approaches from computer science, statistics, graphic design, and information visualization to obtain an adequate understanding of complex datasets. He received the Diploma degree in computer science from the University of Salamanca, the B.S. degree from the University of A Coruña, the B.S. degree in communication studies and the B.A. degree in 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 of the configuration space for redundant robots. He has authored more than 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 paleoclimate 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 was the recipient of the Extraordinary Doctoral Award for his Ph.D. thesis. Contact him at theron@usal.es.

## A.2 Contribution #2

A. Benito-Santos and R. Therón, ‘Pilaster: A Collection of Citation Metadata Extracted From Publications on Visualization for the Digital Humanities’, presented at the 5th Workshop on Visualization for the Digital Humanities, colocated to IEEE VIS 2020, Oct. 2020.

# Pilaster: A Collection of Citation Metadata Extracted From Publications on Visualization for the Digital Humanities

Alejandro Benito-Santos and Roberto Therón

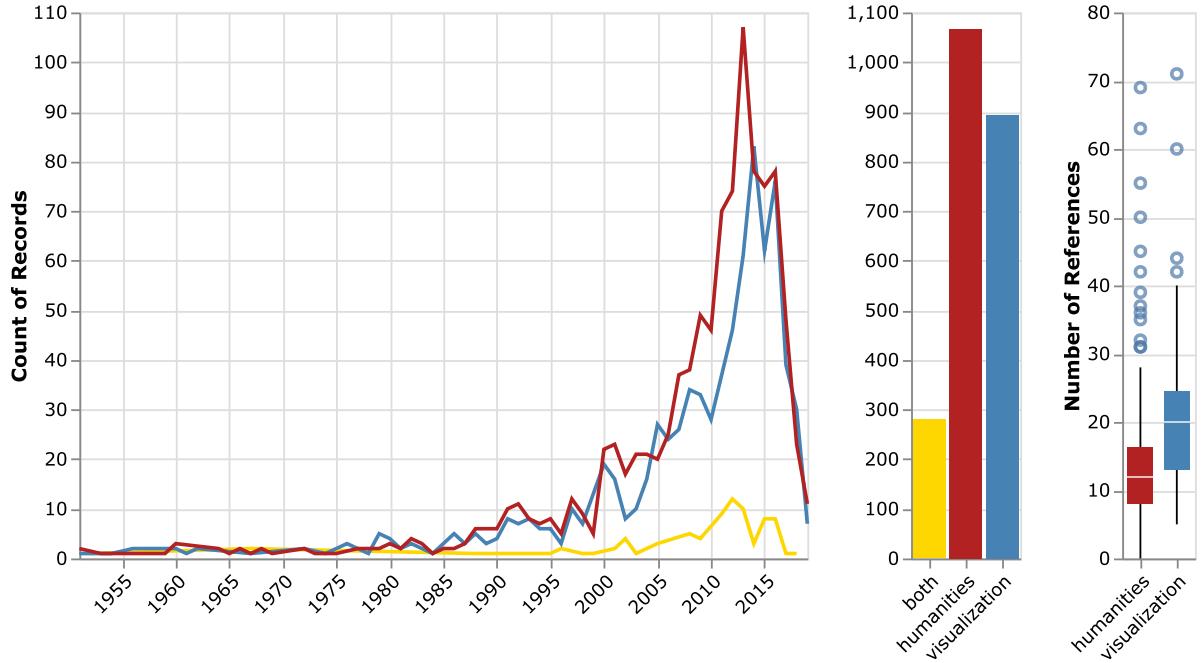


Fig. 1: Three visualizations depicting different features of the citations dataset. From left to right, the first chart shows a portion (1950-2019) of the temporal distribution of resources in the citations dataset originating in the VIS community (blue), in the digital humanities community (red), or in both (yellow). The second chart shows total record counts in the citations dataset for each of these three categories. Finally, the whisker plots on the right display the distribution of reference list lengths by publication venue for records in the seed dataset.

**Abstract**— In this paper, we present Pilaster, a collection of citation metadata extracted from publications in visualization for the digital humanities. The collection is generated from a seed set of relevant publications from which we extracted cited works, including journal and conference papers, books, theses, or blog posts, among other resources. The main aim of this work revolves around three main points: first, the collection may serve as an *entry point* to the discipline for digital humanists and visualization scholars without previous experience in the field. Second, Pilaster can be regarded as a *meeting point* for more established visualization or humanities scholars seeking to collaborate in the development of novel research ideas and related visualization design studies in the context of the humanities. Third, and given the large amount of visualization design spaces that were captured, we believe the dataset has the potential to become the *starting point* for future studies aimed at understanding the particularities of problem-driven visualization research in this and other contexts.

**Index Terms**—collaboration, dataset, digital humanities, visualization, citation analysis, scientometrics

## 1 INTRODUCTION

The collaboration between computer scientists and humanities scholars presents a highly interesting field of experimentation that has produced

important learning outcomes in the past and continues to do so until today. In general, and as it has occurred in other disciplines of science, applying computational methods to humanities research workflows has helped accelerate knowledge discovery and enhance the overall quality of results in humanistic research. A significant part of these combined efforts has typically focused on the application of data visualization techniques aimed at leveraging the interaction between humanities scholars and the said computational methods, producing interesting results in different conventional areas of humanistic research, such as discourse [5], literary [8], or poetry analysis [1, 11, 12] or the browsing and sensemaking of cultural collections [20, 21].

However, the building and organization of interdisciplinary teams of experts that can produce valuable research outcomes in both the visualization and humanities domains seldom are problem-free [10, 18]. Thus, this calls for special considerations to be taken into account by

• Alejandro Benito-Santos is with VisUSAL, University of Salamanca, Spain. E-mail: abenito@usal.es.  
• Roberto Therón is with VisUSAL, University of Salamanca, Spain. E-mail: theron@usal.es.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

all the involved parties [14]. This is particularly the case for visualization researchers new to the field whose previous experience may lie in other areas of visualization practice, and who may rapidly become overwhelmed by the complexities of the collaboration. Analogously, humanities scholars without previous or little experience in participating in visualization design studies may also encounter problems when trying to specify requirements and tasks due to their lack of visualization literacy [15].

The results presented in this paper constitute an extension of our recent work in the field [3, 4] that aims at supporting the immersion process [7] of interdisciplinary researchers in visualization design studies within a digital humanities context, among other goals that are described throughout the paper. To this end, we employ a metadata collection of works on visualization for the digital humanities that we started building in 2019 and that we have kept curating and refining since then. The resulting dataset comprises almost 2,000 resources related to the practice of visualization in the context of digital humanities derived from an extensive analysis of the citations in a core set of 119 papers published at three different venues identified at the beginning of the study. In the following sections, we discuss the rationale we followed to build the dataset, and some of the problems we found in the process and which we could not fit into our previous contribution [4] due to space limitations. Later, we present a description of the data fields and provide several descriptive statistics derived from the data that also offer new insight into the collection. Finally, we exemplify potential applications of *Pilaster* with two simple use cases that others may find useful for carrying their own studies on the dataset. The first use case aims to capture our latest work on normalizing publication aggregation names, an effort that yielded new interesting insights into the commonalities and differences in venues commonly cited by DH and VIS researchers. In a second use case, we shed new light on how collaborations in the field are articulated, which suggests a lack of overlap between the two communities.

## 2 SURVEYING VIS4DH

"Let's be honest—there is no definition of digital humanities, if by definition we mean a consistent set of theoretical concerns and research methods that might be aligned with a given discipline [...] How else to characterize the meaning of an expression that has nearly as many definitions as affiliates? It is a social category, not an ontological one."

R.C. Alvarado in *The Digital Humanities Situation* (2011) [2]

As mentioned in the previous section, the dataset arises from citations found in a core set of publications on visualization for the digital humanities. According to established literature review methodologies [13], the process of literature review starts by defining the scope of the study, ("visualization for the digital humanities" in our case). In the next step, the scope is condensed in a series of textual queries that are launched against online literature databases to obtain relevant publications. Then, these publications are analyzed, summarized, and discussed according to the classification dimensions and other traits derived by the authors of the survey [13]. Finally, the results of the review are wrapped up and prepared for dissemination to the scientific community. Whereas the process is seemingly straightforward, and we knew of similar methodologies that had been successfully applied to conduct surveys on specific sub-fields of the DH visualization practice [20], it presents several issues that rendered it unfit for our purpose of capturing the different DH areas in which the visualization practice *mostly* occurs. Besides, much of the work in digital humanities is presented exclusively at annual conferences (although some notable journals exist) whose proceedings are not indexed in the main online scientific databases. If, as it was our initial intention, our work should be aimed at interdisciplinary visualization practitioners, completely excluding all these works from an initial analysis seemed clearly counterproductive. Still, and beyond these considerations, we had to provide a sensible definition of the digital humanities to commence the survey. Here, we were facing a recurrent problem of the digital humanities that has been at the center of many academic debates. we resorted to the literature looking for a working definition of digital humanities that we

could put to use, but we could not find any. *How were we supposed to survey a topic that cannot be defined?* [6]

As some authors like Alvarado have pointed out, the answer for the question of what the digital humanities are cannot rely on conventional conceptions of what a discipline should be [2]. Rather, he claims, it is more useful to see the digital humanities as a *social category* that relates a collective of researchers who are involved in different, probably distant disciplines, and who call themselves "digital humanists." This statement was the cornerstone on which the methodology we adopted to generate the collection was built, and allowed us to move on to the data collection stage without the need to provide a definition of the digital humanities that would have stood on very shaky epistemological grounds, let alone a query that translated this definition into something that could be understood by a search engine. In such circumstances, we decided to adopt an utilitarian stance that focused instead on identifying the group of scholars who call themselves digital humanists *and* practice visualization. Taking this reasoning further forward, it seemed obvious that this group must be composed of visualization practitioners interested in digital humanities, and also of digital humanists who have shown an interest for visualization. As we discuss in the next section, we looked for specific academic collectives whose members matched any of these two conditions.

## 3 DATA COLLECTION

In this section, we detail how we built a seed dataset of publications from which the citations were extracted at a later stage. The methodology that we followed to build the dataset is inspired by other recent works in visualization research [9, 13] that were adapted to cope with the diffuse character of digital humanities, as we explain in Sect. 2.

### 3.1 Sampling VIS authors

As explained before, the construction of the seed dataset involved the sampling of publications in both ends of the humanities-visualization collaboration. To find the components of the first group, we considered participants in the last editions of the VIS4DH workshop which is, to the best of our knowledge, the only space devoted to the task of "bringing together researchers and practitioners from the fields of visualization and the humanities to discuss new research directions at the intersection of visualization and (digital) humanities research<sup>1</sup>." Although we knew of more research papers published at visualization conferences that could probably have been included in the seed dataset, we decided not do so due to the aforementioned impossibility of establishing a well-defined boundary between what qualifies as digital humanities and what not. At any rate, we assumed relevant papers would eventually appear during the analysis of the citations and therefore we preferred to keep the seed dataset as well scoped as possible. The inspection of the proceedings of the four first editions (2016–2019) of the workshop left a balance of 47 papers and 136 authors making up the first sample to be part of the seed dataset.

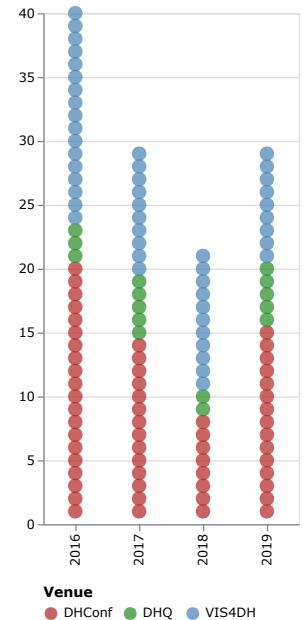


Fig. 2: Dot plot showing the distribution by year and publication venue of papers in the seed dataset. A total of 119 papers were analyzed in a first stage.

<sup>1</sup><https://vis4dh.dbvis.de/>

### 3.2 Sampling DH authors

To obtain representative publications in the humanities side, we decided to inspect the proceedings of the last 4 editions (2016–2019) of the joint annual conference of the Alliance of Digital Humanities Organizations (ADHO) and its peer journal Digital Humanities Quarterly (DHQ). However, and unlike the previous case, we could not find a similar event to the VIS4DH in the DH Conference, and rather visualization practice seems to be spread across different areas such as geohumanities, linked open data, or audiovisuals<sup>2</sup>. Given that we wanted to capture all works that employed visualization techniques regardless of their area of application, we opted for capturing long presentations and papers in the two venues that were related to visualization as tagged by their own authors. Concretely, we captured publications whose title, user-authored keywords or list topics (topics are chosen by the authors from a list of keywords compiled by ADHO) matched the regular expression “[Vv]isua\*”. The search yielded a total of 72 publications (57 long presentations from the conference proceedings and 15 long papers from the journal) which constituted the “humanities” part of the seed dataset. The final composition of the seed dataset is shown in Fig. 2.

## 4 DATA PROCESSING

Publications in the first group were downloaded in PDF format from the workshop’s homepage and their respective reference lists extracted with the pdftotext<sup>3</sup> library and stored for later processing. Reference lists of the second group were obtained by parsing the TEI-XML files in which the documents were encoded. The TEI files of publications in the DH Conference proceedings and the DHQ journal were obtained from the ADHO’s GitHub repository<sup>4</sup> and from the journal’s website, respectively. The TEI files of the 2019 edition of the DH Conference had to be directly scraped from the conference website as they were missing from the repository. The bibliography sections of each paper in the seed dataset were analyzed with the Neural-ParsCit suite [16], which automatically extracts diverse metadata from text lines in a paper’s reference list. The metadata includes but it is not limited to the title, publication year and venue, authors list, DOI and URL. For each of the extracted works, we completed their metadata with information obtained from the Elsevier API<sup>5</sup> by matching their name with existing records in the database. Finally, author names and publication venues were normalized by following a semi-supervised iterative procedure that consisted in visually inspecting pairs displaying short edit distances. Whenever the names were found to refer to the same entity (author or venue), they were unified under their most common form. This process was repeated until no similar pairs were left. At the end of the extraction process, we obtained 2238 references of works that were cited from the seed dataset. They were resolved to 1934 different works of which 23 were publications originally included in the seed dataset.

## 5 DATASET DESCRIPTION

In this section, we describe the data fields that compose the entries and provide general descriptive statistics of the values they take. Below, we list data attributes that are common to items found in the seed or citations datasets:

- **key:** An automatically generated random key that identifies a given resource.
- **title:** The resource title obtained
- **authors:** The item’s list of authors separated by semicolons. The complete list of authors comprises 3499 names of which 185 can also be found in the seed dataset.
- **aggregation:** The normalized name of the aggregation in which the item can be found (e.g., a conference names or journal/book titles). We identified 1148 different aggregation types holding 1783 items in the citations dataset.

<sup>2</sup><https://adho.org/special-interest-groups-sigs>

<sup>3</sup><https://github.com/jalan/pdftotext>

<sup>4</sup><https://github.com/ADHO>

<sup>5</sup><https://dev.elsevier.com/>

- **year:** The year in which the item was created. It was obtained by parsing the reference or from the Elsevier API.

- **source\_theme:** Denotes the provenance of the record. For items in the seed dataset, this field takes two values (“visualization”, “humanities”) depending on the type of the sample that included them, as described in Sect. 3. The value is inherited by items in the citations dataset to annotate their provenance. Items cited from both parts of the seed dataset have this value set to “both”.

Additionally, items in the **seed** dataset contain the following three extra fields:

- **publication\_short\_title:** An abbreviated form of publication\_title.
- **author\_keywords:** Keywords list given by the items’ authors.
- **n\_references:** Length of the reference list that can be found at the end of the paper.

Finally, data attributes exclusive to items in the **citations** dataset are listed below:

- **cited\_by:** A list of foreign keys pointing to papers in the seed dataset that cite the item.
- **cited\_by\_venue:** Venue (VIS4DH, DH Conference, DHQ) of the paper(s) citing the item.
- **cited\_by\_count:** Number of papers in the seed dataset that cite a given item excluding self-references. We considered a citation to be a self-reference when the set intersection between the authors of the citing work and the authors of the cited work was not the empty set.
- **type:** In cases where the publication could not be matched again an Elsevier record, we derived its type (e.g., conference paper, journal article, book) from other publications in the same venue that could be found. In total we identified 20 different cited work types (Fig. 3).
- **aggregation\_type:** The type of the aggregation, if existent, in which the item can be found (e.g., journal, conference proceeding, or book).
- **link:** Web links extracted from the original reference that were parsed by means of a regular expression.

In Fig. 1, we present some descriptive statistics that give an idea of the composition of the citations dataset according to its different dimensions. The first chart on the left shows how both communities follow similar temporal citation patterns with similar mean (2006) and median (2011) values. The next chart shows how the cited resources can be divided into three groups according to the community their citing counterparts belong to. As it can be seen in the figure, we obtained 280 resources that were referenced from VIS4DH and DHConference/DHQ papers, which in turn are among the most cited in the dataset: 82 out of the 100 most cited works belong to this category, which were cited a total of 267 times (11.93% of all citations by papers in the seed dataset). Publications in this category are highly relevant because they represent the intersection point between the visualization and DH communities and therefore, they describe a shared communication channel [18] between visualization and domain experts in DH research that we believe it is worth studying in greater depth. The seed and citations dataset were stored in a public spreadsheet<sup>6</sup> for ease of use by other researchers.

## 6 USE CASES

In this section we propose two simple use cases of the dataset that can help to illustrate the potential use cases for the dataset. The first use case employs the citations dataset to explore commonly cited venues. Besides, we show other venues that are cited exclusively by researchers in one of the two sides. The second use case provides some insights on how interdisciplinary teams are conformed and how the collaborations are organized. The figures in this and the other sections of the paper

<sup>6</sup>[https://docs.google.com/spreadsheets/d/1Z8aMhxpa1510hkusVAFW6L4QyQfPPvUnv8IjKuF2\\_Jo/](https://docs.google.com/spreadsheets/d/1Z8aMhxpa1510hkusVAFW6L4QyQfPPvUnv8IjKuF2_Jo/)

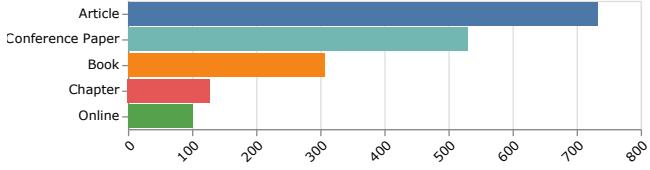


Fig. 3: Five most common resource types in the citations dataset. A majority of the cited items (1,263, 65.30%) belong to one of the two top categories, although there are also references to books (307), book chapters (128) or online resources (100), such as blog posts or datasets.

were generated in Python code<sup>7</sup> using the Vega-Lite grammar [17] and Altair [19].

### 6.1 Studying publication aggregations

In this first use case, we are interested in exploring what venues are cited most often from what kinds of sources in the seed dataset. The stacked bar chart in Fig.4.a shows aggregations above the 95th percentile by number of times cited. From this visualization, some information can be decoded: for example, the two tallest bars in the chart depict the top two most cited venues, which are *IEEE TVCG* and the *DH Conference proceedings*. Moving to the right of the chart, other venues typically associated with visualization research appear, such as the *Conference on Human Factors in Computing* (CHI), *Computer Graphics Forum*, and the VIS4DH workshop, all of which are cited more or less evenly from the two categories of the seed dataset. A similar effect happens with other venues typical of DH research, among which we can find *Digital Humanities Quarterly*, *Digital Scholarship in the Humanities* and its previous title, *Literary and Linguistic Computing*. As opposed to VIS venues, there seems to be a larger imbalance between the categories of items citing DH venues, which are mostly from works in the DH seed dataset. Closer to the tail of the distribution, we can detect other special venues that are exclusively cited by publications originating in the DH domain, such as the *International Conference on Document Analysis and Recognition* (ICDAR) or the *Annual Meeting of the Association for Information Science and Technology* (ASIS&T). We capture this idea in more detail in the chart of Fig.4.b, which represents venues cited exclusively by at least two publications in one of the two domains. These venues, we argue, may be indicative of current knowledge gaps in both sides of the visualization practice that could point to potential new areas for collaboration.

### 6.2 Exploring the authors graph

In this second use case, we obtain insight into the size and structure of collaborations by means of a social network analysis of co-authorship relationships found in the seed dataset. The node-link diagram of Fig. 5 depicts collaborations in both areas. By looking at the color of nodes in the chart, it can be seen that the number of authors who published papers in both categories is fairly (2.76%) low, meaning that interactions between the two communities still are scarce, a fact that may be linked to certain issues pointed by other authors in the past [10]. Attending to the topology of the graph, author communities in the VIS side appear to be larger than their counterparts in the DH side, which may be partially due to differences in the average number of authors per paper in the two groups (DH:  $2.92 \pm 1.51$  vs VIS  $4.15 \pm 2.10$ ) but probably also to other factors that may deserve further study.

## 7 LIMITATIONS AND FUTURE WORK

### 7.1 Data Collection Methodology

In Sect. 3, we described the rationale that we followed to sample publications in both sides of the collaboration. Although we made an explicit effort to obtain a set of publications that was representative of the discipline, we are aware that, due to certain characteristics of the

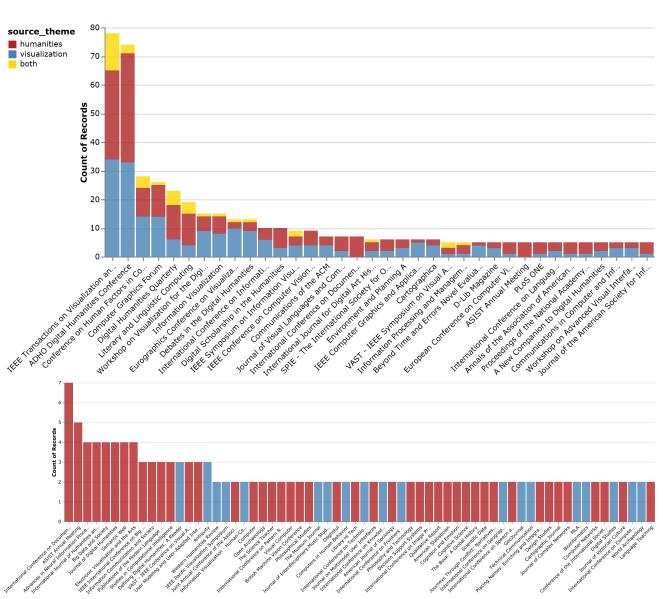


Fig. 4: (a) Top 5% most cited aggregations in the citations dataset. IEEE Transactions on Visualization and Computer Graphics and the ADHO DH Conference are the most cited. (b) Most popular aggregations that are referenced exclusively by at least two different works in the VIS or DH seed datasets.

employed methodology, we might have missed previous work that could have been part of the seed dataset. For example, this could happen with VIS authors working on DH topics who have not participated in the VIS4DH workshop. A similar effect could happen with DH practitioners who decided not to include any terms matching the regular expression “[Vv]isua\*” in their abstracts. In this respect, we expect to receive suggestions from the community of potential new sources that can be included as part of the collection in future developments to make it more complete.

### 7.2 Differences in Publication Formats

The distribution of citations according to their provenance is skewed towards the humanities side, a phenomenon that can be traced to differences between the publication formats typically used on each domain. For example, whereas long presentations at the DH Conference are submitted as abstracts of maximum 750-1000 words, submissions to the VIS4DH workshop adopt the short paper form of 4+1 pages, which usually yield around 3500-4000 words ( $\approx 3x$  longer). Although this difference in length is not translated into a similar difference in the average number of citations per paper between the two categories (right of Fig. 1), humanities papers consistently generated less citations on average than their VIS counterparts. However, they represent a thematically richer set of publications. Although we believe this fact is just representative of the reality of the field and it is not a drawback in itself, it is important to take it into account before extracting any conclusions from the dataset.

### 7.3 Head or Tails

In this paper, we tried to provide an overview of the collection by focusing on the heads of the rank-frequency distributions of, for example, resource types (Fig. 3) or publication aggregations (Fig. 4). Whereas we believe this kind of analysis serves well the objective of describing the dataset, we are aware that this practice may also have unintended side effects: for example, it could happen that these rank-frequency distributions may be interpreted as importance rankings that go beyond the purpose of providing an entry point to the dataset, a practice which we have argued against in the past [3]. By looking *only* at top-ranked items while disregarding the rest, other vital information for advancing

<sup>7</sup><https://colab.research.google.com/drive/15cNprIDxsN1WMa660lo-ApimMib8vdth>

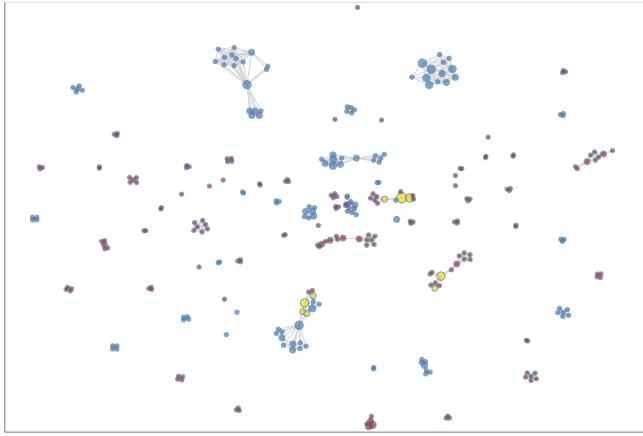


Fig. 5: A node-link diagram depicting co-authorship relationships between authors in the seed dataset. Only 9 out of 328 (2.76%, in yellow) of individuals authored publications in both the VIS and the DH datasets.

the field may be missed, a practice that also dangerously contributes toward perpetuating prestige bias (among other biases) in academia. Rather, we recommend potential users of the collection to repair on items found at the tails of the distributions, for example by performing searches on specific terms that could unveil highly-interesting but lowly-cited, underrepresented themes, works, venues, or authors.

## 8 CONCLUSION

In this paper, we presented *Pilaster*, a metadata collection of papers and related citations employed by scholars working at the intersection of visualization and digital humanities. By departing from a representative sample of publications in the field, we aimed at capturing the different perspectives of scholars at both ends of the collaboration. Furthermore, we exemplified how insight into the discipline can be obtained by means of two use cases that can be easily adapted by other researchers to cover more complex interactions and usage scenarios. In addition, the resulting spreadsheet and code used to generate the figures in this paper were put in the public domain and can be consulted online. Beyond serving as an entry point to the discipline for novel researchers to the field, the results of our work are also aimed at more established scholars who may find them useful for detecting potential future collaborations or novel research ideas, as we illustrated in Sect. 6. Although we plan to continue updating the dataset as new publications become available, we encourage other researchers to send us feedback or suggestions of other use cases that we may not have covered here.

## REFERENCES

- [1] A. Abdul-Rahman, J. Lein, K. Coles, E. Maguire, M. Meyer, M. Wynne, C. R. Johnson, A. Trefethen, and M. Chen. Rule-based Visual Mappings – with a Case Study on Poetry Visualization. *Computer Graphics Forum*, 32(3pt4):381–390, June 2013. doi: 10.1111/cgf.12125
- [2] R. C. Alvarado. The digital humanities situation. *Debates in the digital humanities*, pp. 50–55, 2012.
- [3] A. Benito-Santos and R. Therón Sánchez. Cross-domain Visual Exploration of Academic Corpora via the Latent Meaning of User-authored Keywords. *IEEE Access*, 7:98144–98160, 2019. doi: 10.1109/ACCESS.2019.2929754
- [4] A. Benito-Santos and R. Therón Sánchez. A Data-Driven Introduction to Authors, Readings and Techniques in Visualization for the Digital Humanities. *IEEE Computer Graphics and Applications*, pp. 1–1, 2020. doi: 10.1109/MCG.2020.2973945
- [5] M. El-Assady, V. Gold, A. Hautli-Janisz, W. Jentner, M. Butt, K. Holzinger, and D. A. Keim. VisArgue : A Visual Text Analytics Framework for the Study of Deliberative Communication. In *PolText 2016 - The International Conference on the Advances in Computational Analysis of Political Text*, pp. 31–36, 2016.
- [6] M. K. Gold. Day of dh: Defining the digital humanities. *Debates in the Digital Humanities*, pp. 69–71, 2012.
- [7] K. W. Hall, A. J. Bradley, U. Hinrichs, S. Huron, J. Wood, C. Collins, and S. Carpendale. Design by Immersion: A Transdisciplinary Approach to Problem-Driven Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):109–118, Jan. 2020. doi: 10.1109/TVCG.2019.2934790
- [8] U. Hinrichs, S. Forlini, and B. Moynihan. Speculative Practices: Utilizing InfoVis to Explore Untapped Literary Collections. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):429–438, Jan. 2016. doi: 10.1109/TVCG.2015.2467452
- [9] 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 Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, Sept. 2017. doi: 10.1109/TVCG.2016.2615308
- [10] S. Jänicke. Valuable Research for Visualization and Digital Humanities: A Balancing Act. In *Proc. 1st Workshop on Visualization for the Digital Humanities (VIS4DH)*, 2016.
- [11] S. Jänicke and D. J. Wrisley. On Alignment of Medieval Poetry. In *Digital Humanities 2018 Book of Abstracts*, 2018.
- [12] N. McCurdy, J. Lein, K. Coles, and M. Meyer. Poemage: Visualizing the Sonic Topology of a Poem. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):439–448, Jan. 2016. doi: 10.1109/TVCG.2015.2467811
- [13] L. McNabb and R. S. Laramee. How to Write a Visualization Survey Paper: A Starting Point. In M. Tarini and E. Galin, eds., *Eurographics 2019 - Education Papers*. The Eurographics Association, 2019. doi: 10.2312/eged.20191026
- [14] M. Miller, H. Schäfer, M. Kraus, M. Leman, D. A. Keim, and M. El-Assady. Framing Visual Musicology through Methodology Transfer. *Proceedings of the Workshop on Visualization for the Digital Humanities (VIS4DH) at IEEE VIS 2019*, Oct. 2019.
- [15] J. Pereda, P. A. Murrieta Flores, D. R. Panagiotis, and J. C. Roberts. Tangible User Interfaces as a Pathway for Information Visualisation for Low Digital Literacy in the Digital Humanities. In *Proc. 2nd Workshop on Visualization for the Digital Humanities (VIS4DH)*, 2017.
- [16] A. Prasad, M. Kaur, and M.-Y. Kan. Neural parscit: A deep learning based reference string parser. *International Journal on Digital Libraries*, 19:323–337, 2018.
- [17] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, Jan. 2017. doi: 10.1109/TVCG.2016.2599030
- [18] S. Simon, S. Mittelstädt, D. A. Keim, and M. Sedlmair. Bridging the gap of domain and visualization experts with a Liaison. In *Accepted at the Eurographics Conference on Visualization (EuroVis)*, vol. 2015. The Eurographics Association, Cagliari, Italy, 2015.
- [19] J. VanderPlas, B. Granger, J. Heer, D. Moritz, K. Wongsuphasawat, E. Lees, I. Timofeev, B. Welsh, and S. Sievert. Altair: Interactive Statistical Visualizations for Python. *The Journal of Open Source Software*, 3:1057, Dec. 2018. doi: 10.21105/joss.01057
- [20] F. Windhager, P. Federico, G. Schreder, K. Glinka, M. Dörk, S. Miksch, and E. Mayr. Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2018. doi: 10.1109/TVCG.2018.2830759
- [21] F. Windhager, S. Salisu, and E. Mayr. Exhibiting Uncertainty: Visualizing Data Quality Indicators for Cultural Collections. *Informatics*, 6(3):29, Sept. 2019. doi: 10.3390/informatics6030029

### A.3 Contribution #3

A. Benito-Santos and R. Therón Sánchez, ‘Cross-domain Visual Exploration of Academic Corpora via the Latent Meaning of User-Authored Keywords’, IEEE Access, vol. 7, pp. 98144–98160, 2019.

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<sup>✉</sup>, (Member, IEEE), AND ROBERTO THERÓN SÁNCHEZ<sup>✉</sup>

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.

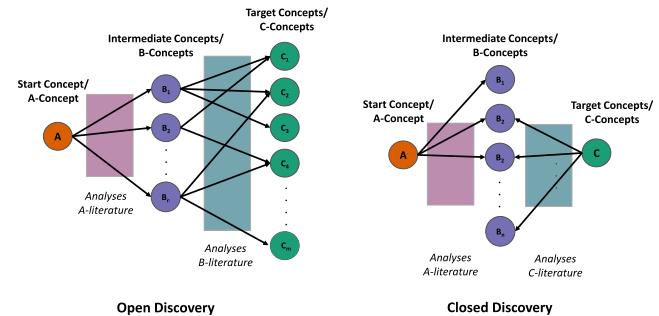
## II. RELATED WORK

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 dendograms 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 ( $U^T U = V^T V = I$ ) and  $\Sigma$  is a diagonal matrix of sorted singular values of the same rank  $r$  as the input matrix. Let  $\Sigma_k$ , where  $k < r$ , be the diagonal formed by the  $k$  first singular values of  $\Sigma$  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),<sup>1</sup> 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.

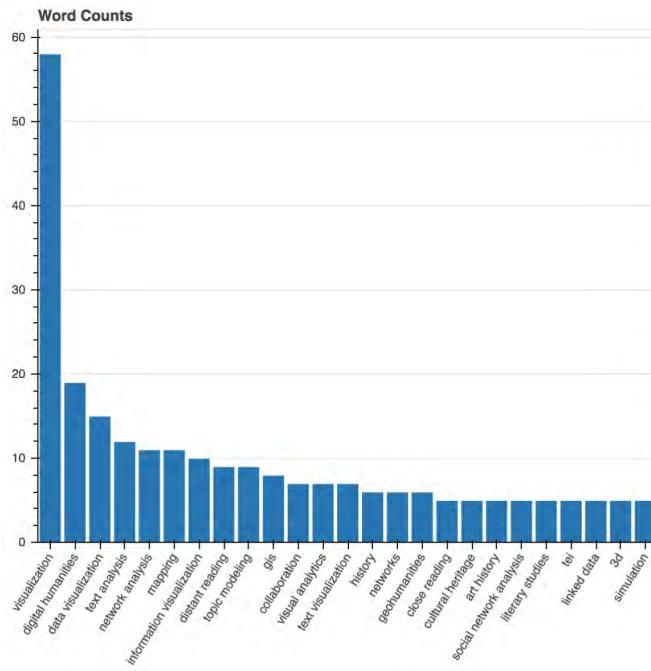
According to the context presented in the previous paragraphs, the first dataset was constructed from metadata describing papers published in the DH conferences between years the years of 2015 and 2018 [47]–[49]. Given the broad range of themes present in this conference, we limited our search to papers 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<sup>2</sup> and actively maintained and updated by its authors.

<sup>1</sup><http://vis4dh.dbvis.de/>

<sup>2</sup><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  $C_q = \{d_{q_1}, d_{q_2}, \dots, d_{q_m}\}$  to a larger target corpus  $C_t = \{d_{t_1}, d_{t_2}, \dots, d_{t_n}\}$  with  $n \gg m$ . In our scheme, any given document is assumed to have a variable number  $j$  of author-assigned keywords  $d_a = \{k_1, \dots, k_j\}$

##### 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

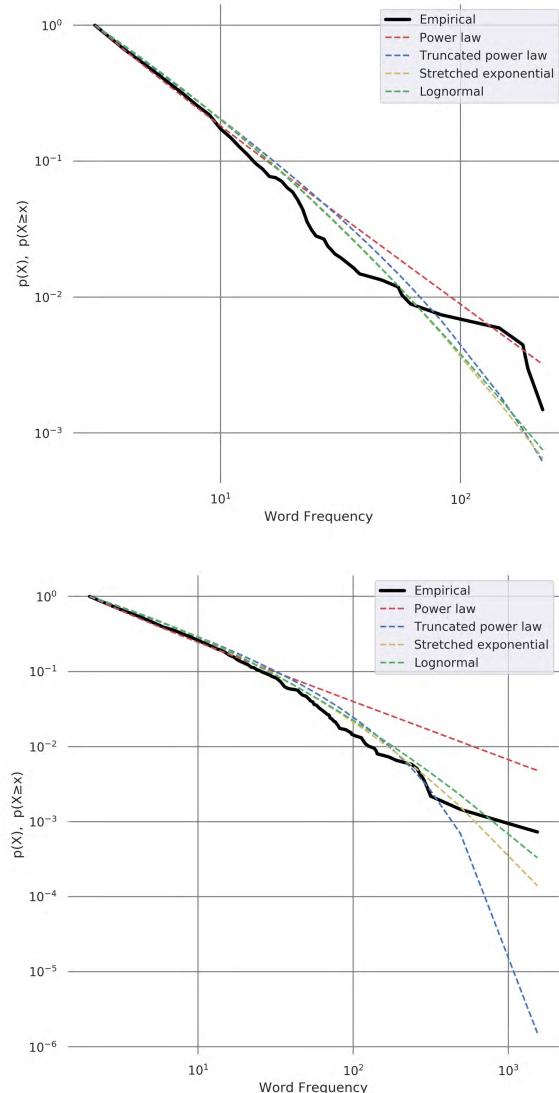
of keywords; second, the new distributional model of the keywords was better captured by LSA, which assumes a Gaussian distribution [51]. Although previous studies [18], [52] employ a power-law distribution to explain the phenomena of author-assigned keywords, recent studies also show this kind of distribution 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 “power-law” [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 ( $|V_q| = 257$ ,  $|V_t| = 2143$ ,  $|V_l| = 320$ ).

## 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



**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 ( $p_a = 0.054$ ,  $gof = 0.0311$ ) 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.

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  $B(t, d) = \{1 \text{ if } t \text{ occurs in } d \text{ or } 0 \text{ otherwise}\}$ .

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

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  $\alpha$  to 0.95 yielded the best results in the similarity task, which is in line with observations from other studies [7].

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

where the smoothed unigram distribution of the context is:

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

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  $PMI(w, c) = \log 0 = -\infty$  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:

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

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  $M$  in the following manner:

$$M_i = \begin{cases} 0 \leq i < |V_q| & \iff M_i \in V_q \\ |V_q| \leq i < |V_q| + |V_l| & \iff M_i \in V_l \\ |V_q| + |V_l| \leq i < |V_g| & \iff M_i \in V_t \end{cases} \quad (5)$$

### 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  $M^{PPMI} \approx \hat{M}^{PPMI} = U_k \Sigma_k U_k^T$ . 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  $D$ :

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

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

$$S(x, y) = 1 - D(x, y) \quad (7)$$

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

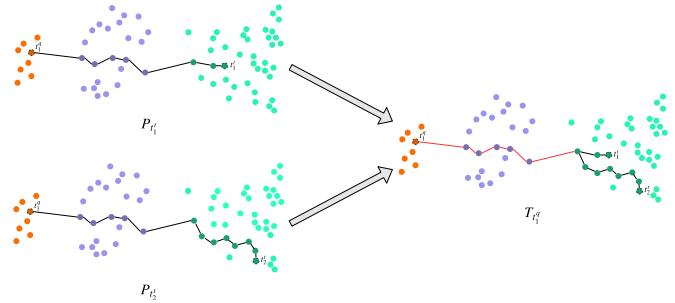
The second stage of our method focuses on exploring the similarity matrix  $S$  that was obtained in the last step. To overcome the conceptual distance between the query and target corpora, we look for structural patterns in the similarity relationships between keywords in the query vocabulary and those 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  $(t_j^r, t_{k1}^r), (t_{k1}^r, t_{k2}^r), \dots, (t_{kl}^r, t_i^q)$  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  $t_j^r$  and every other token in  $V_q$ . Therefore, a shortest path  $P$  exists in  $P'$ , connecting the node  $t_j^r$  to another node  $t_i^q$  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  $S(t_j, t_i)$ .

$$sim(t_j^r, t_i^q) = 1 - \min_{P \in P'_j} \left\{ \sum_{k=1}^l dist(t_k, t_{k+1}) \mid (t_k, t_{k+1}) \in P \right\} \quad (8)$$

By (8), the path  $P_{t_j^r}$  that maximizes the similarity score  $sim$  is a *significant* path of the target token  $t_j^r$  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  $P_1, P_2 \dots P_n$  is a spanning tree of its nodes in  $G$ . This procedure is illustrated in Figure 4. On the left, two shortest paths for tokens  $t_1^r$  and  $t_2^r$  are shown. As  $|V_t| \gg |V_q|$ , some paths will share at least a common destination token in  $V_q$ ,  $t_1^q$  in the example. Input paths are ultimately merged into the same tree  $T_{t_1^q}$ .

After merging paths with common elements, we obtain a set of trees  $T = \{T_1, T_2 \dots T_i\}$  for each token  $t_i^q \in V_q$  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  $t_q^j$



**FIGURE 4.** Shortest paths  $P(t_1^r)$  (left, top) and  $P(t_2^r)$  (left, bottom) connecting tokens  $t_1^r$  and  $t_2^r$  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.

not present in  $T$  to the MST of its nearest neighbor, given that there are not any other shorter paths connecting  $t_q^j$  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.

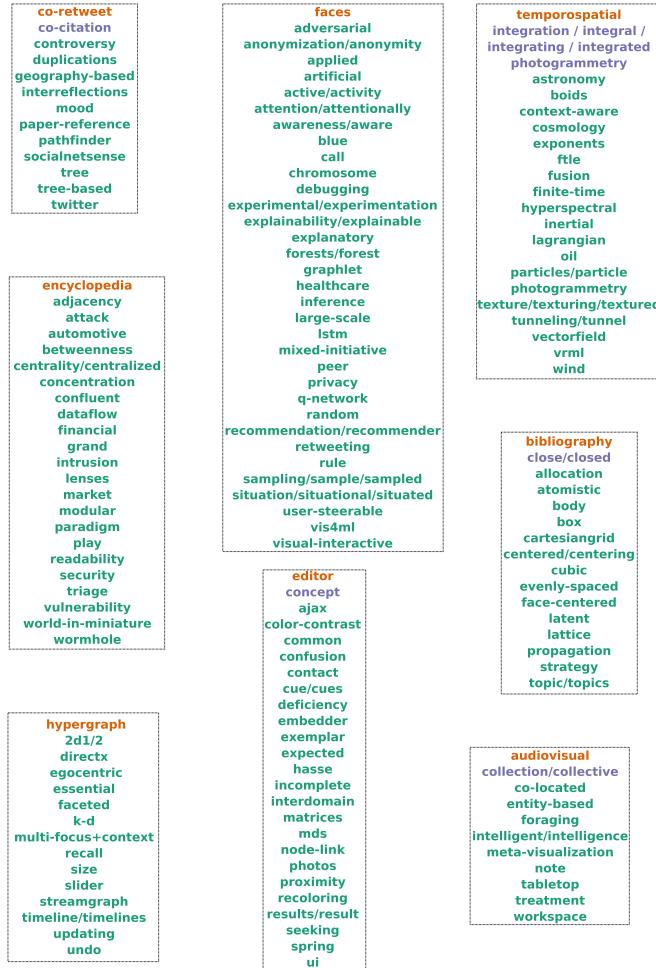
During our experiments, we were able to generate paths for 138 distinct query tokens. On these paths, a total of 1,745 target tokens were represented, along with 85 other link tokens.

## COMBINING SIGNIFICANT PATHS

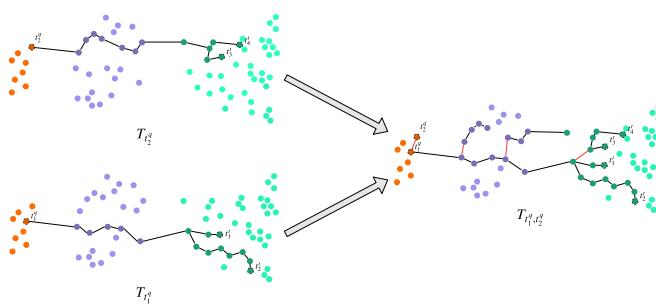
Apart from the visualization of a single tree, our visualization scheme also supports the combination of two or more query terms to represent related keywords and documents. Given that by definition all trees in  $T$  are disjoint subgraphs of  $G$ , we can find an MST in  $G$  that contains all vertices in  $T_1, T_2, \dots, T_n$  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,  $T_{t_1^q}$ , with another tree  $T_{t_2^q}$ . 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.

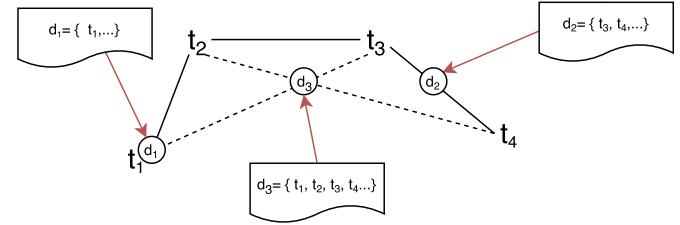


**FIGURE 5.** Keyword components (query, link, target) of some of the trees obtained by our method. Tokens were translated into their original keyword forms for clarity's sake. Each tree can be interpreted as a topic formed by a group of keywords that are highly related to the same element in  $V_q$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



**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).

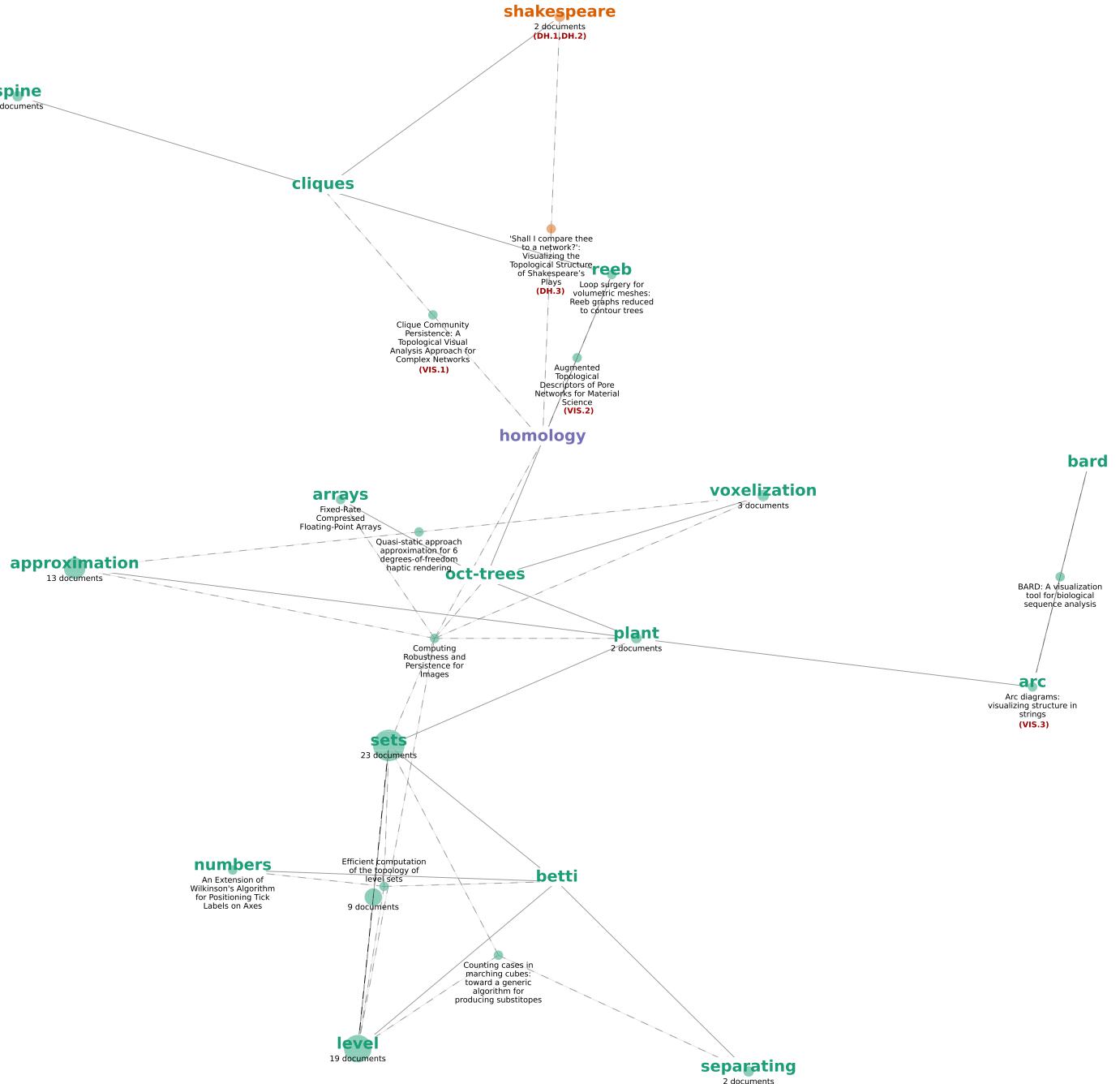
Documents are represented as dots in the visualization and follow the same color scheme as keywords: documents in the query corpus are shown in orange, whereas those appearing in the target dataset are shown in green. Whenever two or more documents share the same position in the plane, they are aggregated 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

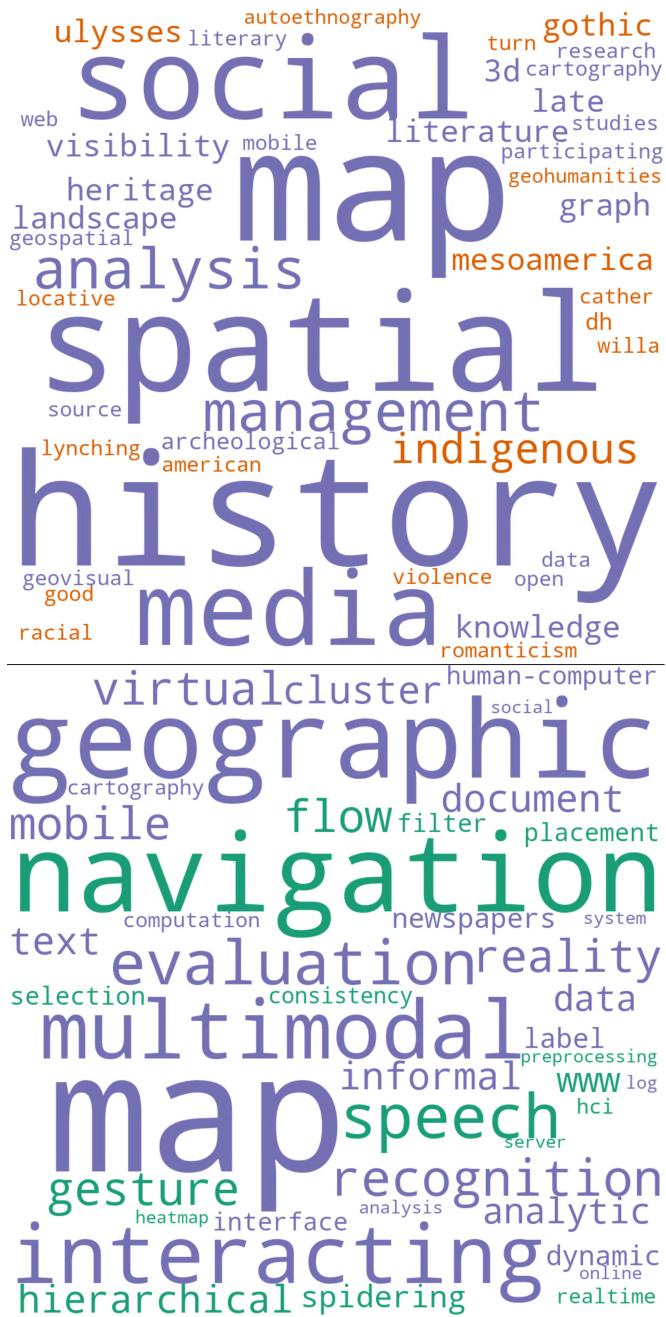
features of data, which is another interesting term as found by our model. Documents “Clique Community Persistence” (VIS.1) and “Augmented Topological Descriptors of Pore Networks” (VIS.2) treat this matter in the context of *graph cliques* and *reeb graphs*, respectively. Interestingly, it can be observed 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



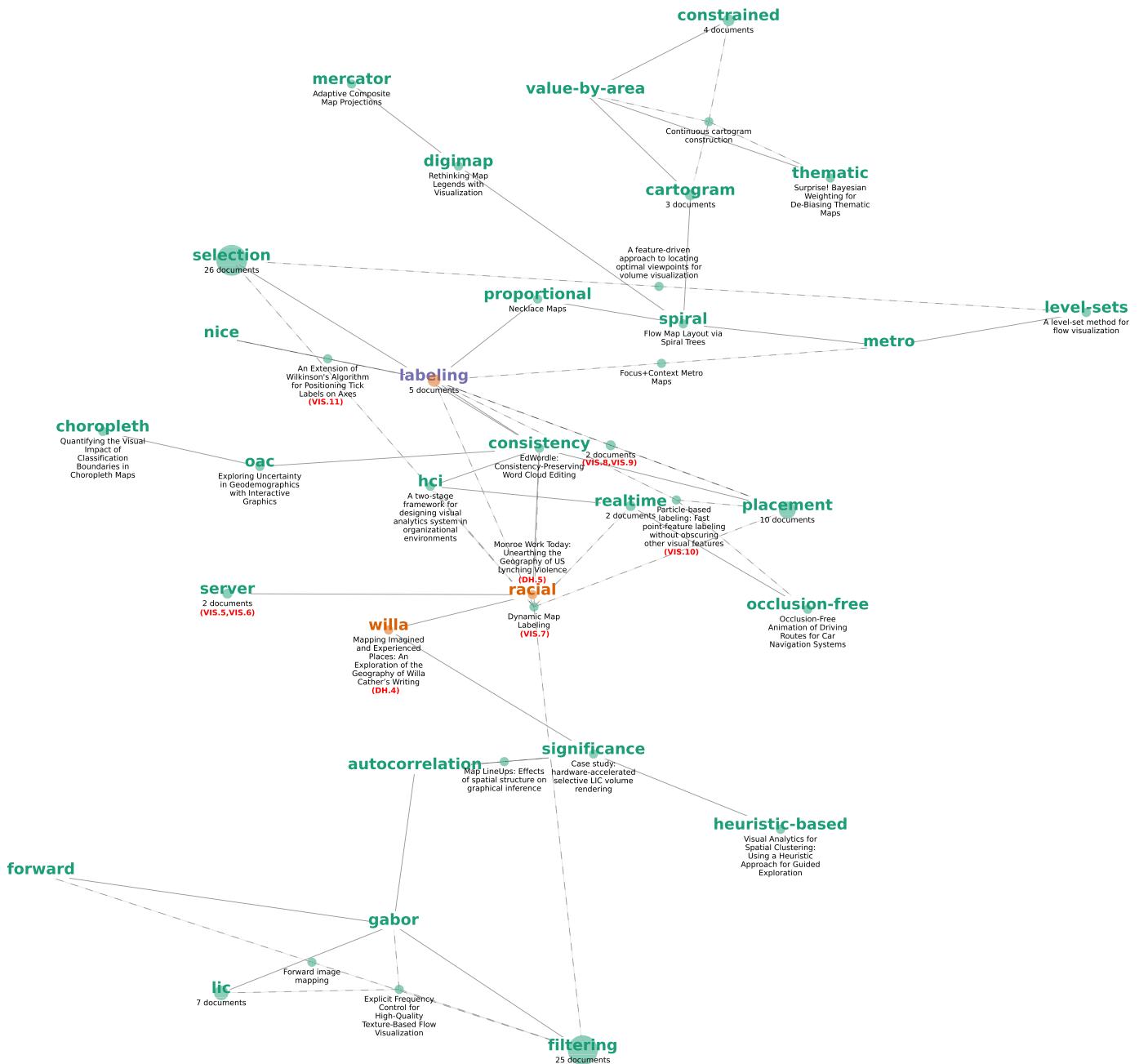
**FIGURE 9.** Word clouds showing the context of the link keyword “gis” in the query (top) and target (bottom) datasets. The SPPMI statistics matrix, in combination with LSA, is able to identify recurrent context terms such as “map” or “spatial,” favoring the establishment of fine-grained affinities 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 sub-theme 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 assess the validity of the showcased techniques and to identify further requirements for future works along this line.

## VII. CONCLUSION

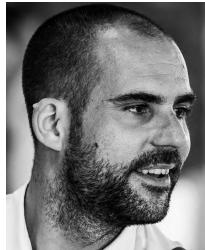
In this paper, we described an automatic method to enhance the open LBD process by visual means. The proposed method allows users to explore author-assigned keywords and related documents in two disjoint bodies of scientific literature which can accelerate the discovery of visualization techniques appropriate 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. Thirunarayanan, 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 survisu," *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, "EpiphaneNet: 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. Trumper and T. E. Goldsmith, "Structural enhancement of learning," *Contemp. Educ. Psychol.*, vol. 29, no. 4, pp. 426–446, Oct. 2004.
- [31] S. Verfíssimo, 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. Dumais, "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 citizenship/digital 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](https://doi.org/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](https://doi.org/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](https://doi.org/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.

• • •

## A.4 Contribution #4

A. Benito-Santos and R. Therón, ‘GlassViz: Visualizing Automatically-Extracted Entry Points for Exploring Scientific Corpora in Problem-Driven Visualization Research’, presented at the 2020 IEEE Visualization Conference (VIS), Oct. 2020.

# GlassViz: Visualizing Automatically-Extracted Entry Points for Exploring Scientific Corpora in Problem-Driven Visualization Research

Alejandro Benito-Santos\*

Roberto Therón†

VisUSAL Research Group. Universidad de Salamanca, Spain

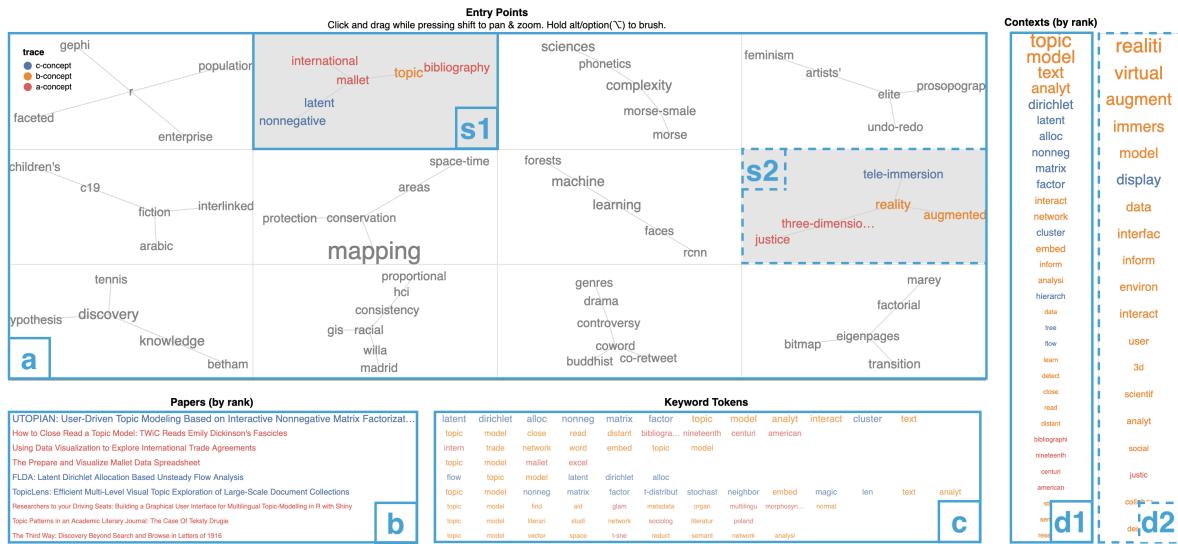


Figure 1: GlassViz interface showing entry points to a corpus of visualization research papers along with related documents and keywords: (a) Quality neighborhoods representing entry points as connected keyword groups. (b) List of documents (showing only the first nine) sorted by number of keyword tokens matching those in selection s.1. (c) Keyword tokens for each document in view b. (d.1) Ranked list of tokens appearing in view c informing of the composition of topics in the entry point selected in s.1. (d.2) State of view d.1 when the selection in view "a" is changed to s.2.

## ABSTRACT

In this paper, we report the development of a model and a proof-of-concept visual text analytics (VTA) tool to enhance document discovery in a problem-driven visualization research (PDVR) context. The proposed model captures the cognitive model followed by domain and visualization experts by analyzing the interdisciplinary communication channel as represented by keywords found in two disjoint collections of research papers. High distributional inter-collection similarities are employed to build informative keyword associations that serve as entry points to drive the exploration of a large document corpus. Our approach is demonstrated in the context of research on visualization for the digital humanities.

**Keywords:** visual text analytics, literature-based discovery, visualization of scientific corpora, distributional similarity, sensemaking, methodology transfer, digital humanities

## 1 INTRODUCTION

Problem-driven visualization research (PDVR) [28] requires intensive collaboration between visualization and domain experts to solve problems in a specific academic discipline such as biology, sports science, computer security, or the humanities. Motivated by the increas-

ing specialization and difficulty of said problems, this collaboration usually materializes in the celebration of workshops, parallel events and micro-conferences (e.g., BioVis, Vis4DH, CityVis, VizSec) and related specialized publication datasets. Setting aside each domain's particularities, these communities generally have to deal with the same typical problems of visualization practice (e.g., dimensionality reduction, hierarchy visualization, or color perception). To obtain insight on these topics and generate novel research ideas [14], researchers perform literature reviews on other larger datasets of visualization publications in search of techniques conceived in other domains that may assist them in solving specific problems of their own domains. This transference of knowledge between communities of practice is known in human-computer interaction (HCI) and visualization as "methodology transfer" (MT), that is, "*the action of utilizing available models that provide solutions to existing and unsolved problems*" [4, 24]. For example, under this paradigm, a digital humanist focusing on the analysis of digital editions may find interesting a visual algorithm conceived for the analysis of genetic data or vice-versa (e.g., an Arc Diagram [33]). However, the arrival at this kind of findings is seldom straightforward. A first hurdle is related to the lack of linguistic competences [28] to formulate queries that serve as entry points [14] to the dataset. To illustrate this situation, take the example of the same digital humanist willing to explore a large corpus of visualization research papers. From previous experience, she knows that the analysis of digital editions is typically related to the concepts of "network analysis" and "graph theory", which are her *entry points* to the dataset. However, she might not be familiar yet with other more specific techniques that

---

\*e-mail: abenito@usal.es

<sup>†</sup>e-mail: theron@usal.es

could be useful in this context, such as "graph clique" or "persistent homology." Conversely, the authors of papers containing these specific terms might not have chosen to include the more general terms "network analysis" or "graph theory" in their keyword selections for being too obvious and thus uninteresting for the audience addressed initially in their works. Therefore, these publications are effectively invisible to the digital humanist's eyes because she has not yet acquired the necessary vocabulary to formulate an adequate query for this dataset. Irremediably, in a typical setup she will have to start the search by typing keyword(s) she is familiar with, initiating an iterative sensemaking process [19, 25], that will be followed by a faceted browsing of the dataset according to its different dimensions (e.g., authors, keywords, or citations). The situation depicted in this example presents further problems: firstly, searching by general terms will return large document lists with varying degrees of relevance that the researcher needs to inspect and filter individually. Second, the subsequent browsing is performed by manual means following a chain of first-order co-occurrences of metadata items, which may rapidly become a frustrating experience for the user, especially when the data volumes are large. To overcome these issues, we propose a distributional model and a related proof-of-concept (POC) tool that aim to capture similarities between keywords in different domains and to automate the generation of meaningful entry points to a corpus of research papers that needs to be explored. The model and tool are demonstrated in the context of a researcher working at the intersection of visualization and the humanities.

## 2 RELATED WORK

**Problem-Driven Visualization Research (PDVR):** PDVR brings together domain and visualization experts that collaborate to solve specific, *inherently complex* domain problems. Beyond technical expertise in both domains, some authors have stressed the importance of language to success in interdisciplinary research [3]. In this regard, Simon et al. explain collaborations in PDVR with a communication model [28] in which domain experts generate the *problem space* by providing *data* and *driving problems*, and visualization experts contribute *exploratory data analysis* and *visualization* techniques defining the *design space*. Following this reasoning, solutions are mappings between the problem and design spaces, and their number is defined by the breadth (or richness) of the communication channel shared by the two teams. More recently, Miller et al. [24] developed these concepts further in their Methodology Transfer Model (MTM). The MTM incorporates the notions of *similarity* and *alignment* to identify potential MTs between different knowledge domains. Our work employs distributional similarity to extend these theoretical models with other concepts drawn from information science (see next paragraph). **Literature-Based Discovery (LBD):** LBD is a knowledge extraction technique that "*generates discoveries, or hypotheses, by combining what is already known in the literature.*" [31] The concept was introduced in the 1980s by Don R. Swanson, an information scientist known for coining the first form of LBD, the *ABC model* [29]. The *ABC model* employs *transitive inference* to unveil non-trivial implicit associations between two disjoint bodies of scientific literature (source and target). It utilizes a simple yet powerful syllogism to pair knowledge fragments: If a term/concept (a-concept) is related to the intermediate term/concept (b-concept) which appears in both the source and target literatures, and the b-concept is related to a c-concept which only appears in the target literature, then we can find a relation, characterized by the b-concept, between the a-concept (which the user is familiar with) and the c-concept (which is new to user). Specifically, we look upon recent work by Thilakaratne et al. [30], who employ word embeddings to find interesting cross-disciplinary affinities in online paper databases. As opposed to the authors, who employ paper abstracts to generate neural embeddings using the *word2vec* model [23], our work relies on author-assigned keywords (hereinafter simply "key-

words"), which are descriptive words assigned by the authors to their research papers and have been successfully employed in the past by other researchers to "facilitate the process of understanding differences and commonalities of the various research sub-fields in visualization." [18]. Also, and despite recent efforts [8], the process by which humans extract keywords from academic texts remains mostly unknown [20]. Therefore, keywords model a unique and highly expressive language that serves as the starting point for our study. **Visual Text Analytics (VTA) of Scientific Literature:** In recent times, some authors have started to incorporate linguistic and sensemaking models into their VTA tools to replicate the typical tasks and goals of exploring scientific texts [12]. For example, the Action Science Explorer [10] and PaperPoles [15] mimic the sensemaking process of traditional literature reviews. Concretely, PaperPoles supports the browsing of publications in a context-aware environment by requesting positive or negative queries from the user as the application workflow progresses. PaperQuest [26] employs a relevance algorithm to rank papers according to the sensemaking process of literature reviews. PaperQuest assumes that the user has one or more *seed papers* at her disposal to start the exploration, a concept that we implemented in GlassViz. Guo et al. [14] propose a two-stage sensemaking framework to discover novel research ideas based on previous work by Pirolli and Card [25]. Wang et al. implement two different logic flows in their system (author-based and citation-based) to mirror the traditional literature review process [32]. To the best of our knowledge, GlassViz is the first VTA tool to incorporate the sensemaking model followed by interdisciplinary visualization researchers using an LBD workflow.

## 3 DATA PROCESSING

We selected two research paper collections as the S and T literatures in our LBD setup. T-Literature (VIS4DH dataset), representing the target domain that solutions need to be imported to, comprises 221 papers on visualization for the Digital Humanities (DH) [2]. S-Literature (VIS dataset) is a set of 2117 visualization publications that have appeared at the IEEE Visualization set of conferences: InfoVis, SciVis, VAST and Vis between the years 1991–2018 [17]. Keywords were extracted from each document, tokenized and translated into their American English forms when necessary. Tokens matching NLTK's list of English stop words (e.g., "and" or "of") were removed from further analysis, which yielded a total of 3403 different tokens. Next, each token was light-stemmed using the Porter algorithm. Given that keywords are a very sparse feature of scientific papers, the stemming procedure had the positive effect of compressing the input vocabulary (from 3403 to 2720 tokens) by linking redundant forms together under the same root (e.g., "filtering," "filters" and "filtered" under "filter"). In addition, and despite certain limitations that we discuss in Sect. 6, the stemming algorithm helped relate documents referring to the same high-level concepts requiring minimal human intervention (e.g., a manual classification [17]). Finally, we removed uninteresting tokens with inverse document frequency (IDF) of less than 1.0, resulting in only one token ("visual") being discarded. Each document was treated as a bag-of-(key)word tokens defining a vocabulary composed of three disjoint sets as per Swanson's ABC model:  $V_a$  (a-concepts, or tokens appearing exclusively in the VIS4DH dataset),  $V_c$  (c-concepts, or tokens appearing exclusively in the VIS dataset), and  $V_b$  (b-concepts, or tokens appearing in both datasets). In the end, the vocabulary sizes obtained were:  $|V_a| = 259$ ,  $|V_b| = 302$ , and  $|V_c| = 2159$ .

## 4 SYSTEM DESIGN

### 4.1 Tasks and Design Goals

Our approach relies on the extraction of entry points to guide the exploration of a scientific corpus. The extraction of the entry points is based on the following assumptions: at the beginning of this study, we observed that researchers participating in PDVR internally

follow an MTM that is mainly driven by their previous experience in other projects and domains. Here, the expert initially analyzes the problem and breaks into its constituent parts, leading to a set of themes that are matched against previous grounded knowledge. In this mental process, candidate solutions are detached from the original problem's domain and matched against the new domain in search of viable solutions. The most *similar* solutions are then implemented to obtain preliminary insight into the data, which is often necessary to promote discussions between stakeholders and advance the project at its early stages. Later in the design process, the team may decide to modify and/or combine these initial solutions to provide a visualization that aligns better with the data and tasks of the problem at hand [24]. Motivated by the presented situation, we extracted the following design goals and related questions at the beginning of the study, which ultimately drove the design of our distributional model and POC tool: **DG.1:** Motivate a personalized exploration of scientific corpora that is tailored to the user's research aims. "What kind of knowledge does the user want to extract from a dataset?", "What can a user learn from the dataset that is useful for solving a particular domain problem?" **DG.2:** Potentiate the discovery of methodologies that could potentially be transferred from other existing design spaces to the source domain. "How can we measure the degree of transferability of solutions conceived in other knowledge domains?" **DG.3:** Accelerate sensemaking and language acquisition in the context of PDVR. "What are the most informative terms that best describe a dataset according to the user's level of expertise and grounded knowledge?", "What themes are especially interesting for the user?", "How can they be presented in the best possible manner to augment their comprehensibility?" **DG.4:** Provide a reading order for discovered documents. What documents are the most important in the collection for the user?

## 4.2 Theoretical Model

Our theoretical model (Fig. 2) combines Swanson's and Miller et al.'s models to build automatic entry points that resemble the researchers' sensemaking model and assist them in the task of mapping problem and design spaces in different domains and bodies of literature. According to Simon et al. [28], the problem space is defined by application domains and their data, whereas the design space comprises analytical tasks and visualizations. In the diagram, we depict the idea that valid MTs consist of a series of concepts specific to each domain (a- and c-concepts) and a variable number of techniques that address a generic, high-level problem in the visualization domain (b-concepts). Thus, as per Swanson's model, solutions (or papers) in the T-literature link problems and designs containing only a- and b-concepts, while those in the S-literature contain only c- and b-concepts. Then, it should be theoretically possible to deduct recurrent terms of potential solutions by analyzing the distribution of terms in existing solutions documented in the literature in other domains and relating them to the problem(s) at hand using high-order co-occurrence. This idea is depicted in the Venn diagram at the center of the image. At the intersection of the four sets, the core terms of the elements in the four spaces meet, giving clues about the descriptions of potential solutions. Besides, more potential solutions could be found by following chains of co-occurrence that led to peripheral intersection spaces. As we explain in the next section, our proposed model captures high-order co-occurrence of concepts for enhancing the document exploration process.

## 4.3 Keyword Embeddings

We rely on the generation of keyword embeddings for detecting distributional similarities between problems, data, tasks, or visualizations in the S- and T-literatures. These embeddings were generated by following the method proposed by Levy et al. [21], which requires minimal hyper-parameter tuning and they are known to excel at word similarity tasks [21, 22]. Initially, the method relies on

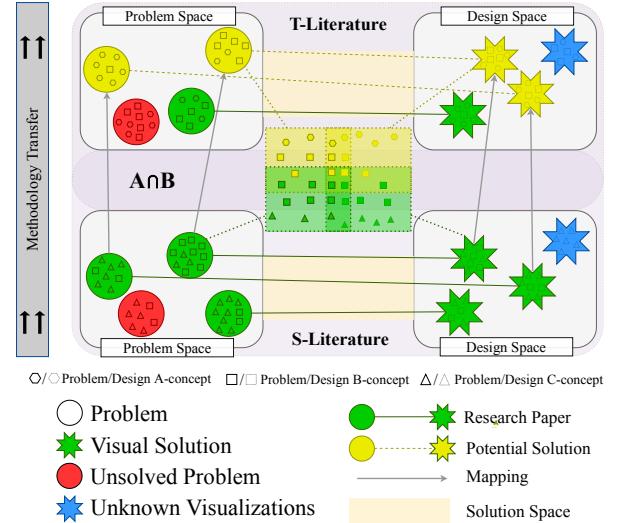


Figure 2: Methodology Transfer Model (MTM) adapted from Miller et al. [24]. The model maps problems and designs found in two disjoint bodies of literature and it is augmented with concepts drawn from Swanson's ABC Model for Literature-Based Discovery to automate the discovery of candidate MTs and to provide the user with informative entry points to the S-Literature.

an initial pointwise mutual information (PMI) matrix that encodes the probability for a pair of keyword tokens to be seen together in a document with respect to the probability of seeing those two same tokens in the union of the two corpora (see Equation 1). For all keyword token pairs in the S and T literatures, each cell  $M_{i,j}$  represents the log odds ratio of  $w_i$  (a keyword) and  $c_j$  (any other keyword appearing with  $w$  in a document D, its context) joint probability and the product of their marginal probabilities. The marginal probabilities were empirically obtained from the corpora by counting the number of occurrences of each token divided by the union size of the document collections.

$$PMI(w, c) = \log \frac{P(w, c)}{P(w)P(c)} = \log \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \quad (1)$$

Given that PMI can be  $-inf$  for pairs of tokens that were never seen in the corpus, it is customary to use the positive version of the PMI matrix that is defined as:

$$PPMI(w, c) = \max(PMI(w, c), 0) \quad (2)$$

Following recommendations in the literature [1, 22], we applied a light smoothing with  $\alpha = 0.95$  (Equation 4) to counterbalance the PMI bias towards infrequent events (note that the alpha factor is a corpus-dependent parameter and was manually adjusted).

$$SPPMI(w, c) = \log \frac{\hat{P}(w, c)}{\hat{P}(w)\hat{P}_\alpha(c)} \quad (3)$$

where the smoothed unigram distribution of the context is:

$$\hat{P}_\alpha(c) = \frac{\#(c)^\alpha}{\sum_c \#(c)^\alpha} \quad (4)$$

To capture high-order co-occurrence and to generate dense keyword vectors from the sparse SPPMI matrix was factorized into the product of three matrices by applying a non-parametric algebraic method, SVD, which was popularized in the NLP community with Latent

Semantic Analysis (LSA) [9, 21]. If the SPPMI matrix is the matrix  $M$ , SVD decomposes  $M$  into the product of three matrices  $U\Sigma V^T$ , where  $U$  and  $V$  are orthonormal ( $U^T U = V^T V = I$ ) and  $\Sigma$  is a diagonal matrix of sorted singular values of the same rank  $r$  as the input matrix. Then, our resulting vector space model (VSM) is formed by dense keyword embeddings resulting from keeping only the first  $k$  columns in  $U$  ( $k = 50$  in our case).

## 5 GLASSVIZ

In this section, we describe the design decisions that drove the development of our prototype tool by carrying an experiment using the datasets introduced in Sect. 3. Our approach is centered around the qualitative inspection of *quality* local neighborhoods of a-concepts that were derived using a *cosine metric* [30]. According to the literature, it is customary to select between 3 and 5 nearest neighbors for this task (see [16], section 4.1.1) Thus, we decided to extract the 4 nearest neighbors for each a-concept  $t_a$  in  $V_a$ . Tokens represented by very similar vectors ( $\text{sim}(t_a, t_b) \leq 0.01$ ) and thus displaying identical nearest neighbors were considered redundant for the purpose of this task and thus were removed (488 in total). Quality neighborhoods were defined as those containing significant similarities between a- and c- concepts and were identified by two criteria: (1) the neighborhood included at least one c-concept, and (2) when criterion 1 was met, the similarity between the a-concept and its nearest c-concept in the neighborhood fell within the first quartile of all highest similarities ( $\text{dist}(t_a, t_c) \leq Q_1$ , with  $Q_1 = 0.2451$ ), which yielded 15 quality neighborhoods. To relate neighborhoods representing similar themes, neighborhoods with common terms were merged, resulting in 12 distinct entry points. Finally, we wanted to display the neighborhood's embedding subspaces defined by each entry point in the best possible manner to motivate a gradual transition from familiar a-concepts to interesting, possibly unknown c-concepts. This implied representing not only similarities between the nearest neighbors and the a-concept originating the neighborhood but also showing similarities among neighbors. To this end, we relied on a graph scaling technique, pathfinder networks (PFNETs) [27] that was applied to the complete similarity subgraphs formed by terms on each entry-point. PFNETs are well-known in the visualization and information theory literature [1, 5, 6] for their suitability to capture underlying knowledge structures (**DG.1**) and for motivating a fast vocabulary learning (**DG.3**) with a minimal cognitive gap. This is achieved by pruning graph edges that are not on shortest paths according to two parameters  $q$  (the number of indirect proximities considered to build the PFNET) and  $r$  (the metric used to compute pairwise similarities) [5, 6, 27]. Concretely, we calculated minimum spanning trees (MSTs), the most concise form of a PFNET ( $q = n - 1, r = \infty$ ), for the 12 complete subgraphs. Following recommendations in the literature [5], each PFNET was plotted using a force-directed algorithm [13] that placed nodes displaying high pairwise cosine similarities closer in the chart. In this representation, the nodes depict a-, b-, or c-concepts as per Swanson's ABC model. Each MST portrays an exploration path (or entry point) to the VIS dataset that can be inspected individually in the designated areas of view 1.a (see Fig. 1). A total of 29 a-concepts (red), 16 b-concepts (yellow) and 19 c-concepts (blue) were captured. Each node shows a text label containing the most common form of the corresponding token and whose size encodes the token(s)'s absolute frequency in the union of the two literatures. In view 1.b, documents in the current selection are listed in descending order of number of keyword tokens matching those in the current selection (**DG.4**). The number of documents containing any of the terms captured by the entry points was 69 for the T-Literature A and 297 for the S-Literature (31.22% and 14.03% coverage, respectively). To the right of view 1.b., view 1.c shows keyword tokens for each document shown in view 1.b., whereas view 1.d1 (and 1.d2 for selection s2) aggregates and presents these tokens in a rank frequency list.

By visually inspecting each of the 12 entry points in view 1.a (**DG.1**), the user can recognize interesting inter-collection distributional similarities between concepts describing application areas, domain problems, analytical tasks/techniques and visualizations as per the model introduced in Sect. 4.2. The entry points can be further inspected using a brushing+linking interaction technique. For example, when brushing the entry point in selection s1 of Fig. 1, views 1.b, 1.c and 1.d1 are updated. Here, the entry point introduces two c-concepts ("nonnegative" and "latent") that are related by their distributional similarity to two DH-specific problems ("bibliography" and "international") and a technique ("mallet") depicted by the a-concepts in red in the diagram. To get a better understanding of the entry point's underlying theme and concepts (**DG.3**), the user could look at view 1.d1 to discover the most frequent tokens ("topic," "model," "text," "analyt," "dirichlet," etc.) found in documents matching any of the entry point's concepts shown in s1, allowing a first rapid interpretation of the theme. By interacting with the items in view 1.b, the user could retrieve in a pop-up view multiple related metadata to a document; i.e., title, authors list, publication year/venue and number of matching keywords with the entry point. In the same view, it can be observed that the three a-concepts in s1 can be traced to three documents in the VIS4DH dataset describing two domain problems (the analysis of international trade agreements and bibliographic works, respectively) and a domain-specific analysis tool, a wrangling Excel script for a popular NLP toolkit among DH practitioners. Similarly, the two c-concepts "nonnegative" and "latent" can be traced to three other documents in the VIS dataset and reconstructed by the user to "nonnegative matrix factorization" and "latent dirichlet analysis," introducing potentially interesting analysis techniques (**DG.2**). The same workflow could be applied to any other entry point of view 1.a, for example to the one depicted in selection s2. This entry point relates the domain problem of "social justice" to the a-concept "tele-immersion" under the background theme of virtual and augmented reality (view 1.d2).

## 6 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

We have presented a model and a related VTA prototype aimed at accelerating the process of knowledge and language acquisition in PDVR. By modeling the distribution of keywords defining the interdisciplinary communication channel as documented by research papers found in two disjoint bodies of literature, we were able to generate entry points that motivated a personal exploration of a corpus of visualization papers according to the researcher's particular needs and expectations and that required minimal user intervention. However, we identified the existence of certain limitations in our approach that are discussed hereafter: firstly, the stemming algorithm employed to compress the input data produced some false positives that are difficult to avoid by automatic means. Concretely, this side effect could be observed in cases where keywords with different meanings were reduced to the same lexical form, for example in the tokens *factory* (from "smart factory") and *factorial* (from "factorial analysis"). Also, *GlassViz* does not allow the interactive tuning of certain parameters such as the number  $k$  of singular values, the smoothing alpha factor  $\alpha$ , or the similarity thresholds set to detect redundant vectors and quality neighbors. To resolve these and other issues, we plan to incorporate direct manipulation techniques [11] in the future. Furthermore, the tokenization of keywords increased the difficulty in interpreting the entry points' background themes, a limitation that could be resolved by employing auxiliary n-gram statistics [7] to assist the user in reconstructing the original phrases.

## ACKNOWLEDGMENTS

The authors want to thank the three anonymous reviewers for their helpful comments. This work was supported by a grant from the Spanish Ministry of Economic Affairs and Digital Transformation under the EU CHIST-ERA agreement (PCIN-2017-064).

## REFERENCES

- [1] A. Benito-Santos and R. Therón Sánchez. Cross-domain visual exploration of academic corpora via the latent meaning of user-authored keywords. *IEEE Access*, 7:98144–98160, 2019.
- [2] A. Benito-Santos and R. Therón Sánchez. A Data-Driven Introduction to Authors, Readings and Techniques in Visualization for the Digital Humanities. *IEEE Computer Graphics and Applications*, pp. 1–1, 2020.
- [3] L. J. Bracken and E. A. Oughton. ‘What do you mean?’ The importance of language in developing interdisciplinary research. *Transactions of the Institute of British Geographers*, 31(3):371–382, July 2006. doi: 10.1111/j.1475-5661.2006.00218.x
- [4] R. Burkhard. Learning from architects: The difference between knowledge visualization and information visualization. In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, pp. 519–524, July 2004. doi: 10.1109/IV.2004.1320194
- [5] C. Chen. Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management*, 35(3):401–420, May 1999. doi: 10.1016/S0306-4573(98)00068-5
- [6] C. Chen, J. Kuljis, and R. J. Paul. Visualizing latent domain knowledge. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 31(4):518–529, Nov. 2001. doi: 10.1109/5326.983935
- [7] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization Techniques for Assessing Textual Topic Models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI ’12, pp. 74–77. ACM, New York, NY, USA, 2012. doi: 10.1145/2254556.2254572
- [8] J. Chuang, C. D. Manning, and J. Heer. “Without the Clutter of Unimportant Words”: Descriptive Keyphrases for Text Visualization. *ACM Trans. Comput.-Hum. Interact.*, 19(3):19:1–19:29, Oct. 2012. doi: 10.1145/2362364.2362367
- [9] 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*, 41(6):391–407, 1990. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
- [10] C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12):2351–2369, 2012. doi: 10.1002/asi.22652
- [11] M. El-Assady, R. Kehlbeck, C. Collins, D. Keim, and O. Deussen. Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1001–1011, Jan. 2020. doi: 10.1109/TVCG.2019.2934654
- [12] P. Federico, F. Heimerl, S. Koch, and S. Miksch. A Survey on Visual Approaches for Analyzing Scientific Literature and Patents. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2179–2198, Sept. 2017. doi: 10.1109/TVCG.2016.2610422
- [13] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991. doi: 10.1002/spe.4380211102
- [14] H. Guo and D. H. Laidlaw. Topic-based Exploration and Embedded Visualizations for Research Idea Generation. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2018. doi: 10.1109/TVCG.2018.2873011
- [15] J. He, Q. Ping, W. Lou, and C. Chen. PaperPoles: Facilitating adaptive visual exploration of scientific publications by citation links. *Journal of the Association for Information Science and Technology*, 70(8):843–857, 2019. doi: 10.1002/asi.24171
- [16] F. Heimerl and M. Gleicher. Interactive Analysis of Word Vector Embeddings. *Computer Graphics Forum*, 37(3):253–265, June 2018. doi: 10.1111/cgf.13417
- [17] 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 Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, Sept. 2017. doi: 10.1109/TVCG.2016.2615308
- [18] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller. Visualization as Seen through its Research Paper Keywords. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):771–780, Jan. 2017. doi: 10.1109/TVCG.2016.2598827
- [19] G. Klein, B. Moon, and R. R. Hoffman. Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems*, 21(5):88–92, Sept. 2006. doi: 10.1109/MIS.2006.100
- [20] S. Lahiri. Replication of the Keyword Extraction part of the paper “Without the Clutter of Unimportant Words”: Descriptive Keyphrases for Text Visualization”. *arXiv e-prints*, 1908:arXiv:1908.07818, Aug. 2019.
- [21] O. Levy and Y. Goldberg. Neural Word Embedding as Implicit Matrix Factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 27*, pp. 2177–2185. Curran Associates, Inc., 2014.
- [22] O. Levy, Y. Goldberg, and I. Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3(0):211–225, May 2015.
- [23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.
- [24] M. Miller, H. Schäfer, M. Kraus, M. Leman, D. A. Keim, and M. El-Assady. Framing Visual Musicology through Methodology Transfer. In *Proc. 4th Workshop on Visualization for the Digital Humanities (VIS4DH)*, 2019.
- [25] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, pp. 2–4, 2005.
- [26] A. Ponsard, F. Escalona, and T. Munzner. PaperQuest: A Visualization Tool to Support Literature Review. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA ’16, pp. 2264–2271. Association for Computing Machinery, San Jose, California, USA, May 2016. doi: 10.1145/2851581.2892334
- [27] R. W. Schwaneveldt, ed. *Pathfinder Associative Networks: Studies in Knowledge Organization*. Pathfinder Associative Networks: Studies in Knowledge Organization. Ablex Publishing, Westport, CT, US, 1990.
- [28] S. Simon, S. Mittelstädt, D. A. Keim, and M. Sedlmair. Bridging the gap of domain and visualization experts with a Liaison. In *Proceedings of the Eurographics Conference on Visualization (EuroVis)*, vol. 2015. The Eurographics Association, Cagliari, Italy, 2015.
- [29] D. R. Swanson. Fish Oil, Raynaud’s Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986. doi: 10.1353/pbm.1986.0087
- [30] M. Thilakaratne, K. Falkner, and T. Atapattu. Automatic Detection of Cross-Disciplinary Knowledge Associations. In *Proceedings of ACL 2018, Student Research Workshop*, pp. 45–51, July 2018.
- [31] M. Thilakaratne, K. Falkner, and T. Atapattu. A Systematic Review on Literature-based Discovery. *ACM Computing Surveys (CSUR)*, Dec. 2019.
- [32] Y. Wang, D. Liu, H. Qu, Q. Luo, and X. Ma. A Guided Tour of Literature Review: Facilitating Academic Paper Reading with Narrative Visualization. In *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction*, VINCI ’16, pp. 17–24. ACM, New York, NY, USA, 2016. doi: 10.1145/2968220.2968242
- [33] M. Wattenberg. Arc Diagrams: Visualizing Structure in Strings. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, pp. 110–116, Oct. 2002. doi: 10.1109/INFVIS.2002.1173155

## A.5 Contribution #5

A. Benito-Santos and R. Therón Sánchez, ‘Defragmenting Research Areas with Knowledge Visualization and Visual Text Analytics’, *Applied Sciences*, vol. 10, no. 20, Art. no. 20, Jan. 2020.

Article

# Defragmenting Research Areas with Knowledge Visualization and Visual Text Analytics

Alejandro Benito-Santos \* and Roberto Therón Sánchez \*

VisUSAL Research Group, Universidad de Salamanca, 37008 Salamanca, Spain

\* Correspondence: abenito@usal.es (A.B.-S.); theron@usal.es (R.T.S.)

Received: 27 September 2020; Accepted: 14 October 2020; Published: 16 October 2020

**Abstract:** The increasing specialization of science is motivating the fragmentation of traditional and well-established research areas into interdisciplinary communities of practice that focus on cooperation between experts to solve problems in a wide range of domains. This is the case of problem-driven visualization research (PDVR), in which groups of scholars use visualization techniques in different application domains such as the digital humanities, bioinformatics, sports science, or computer security. In this paper, we employ the findings obtained during the development of a novel visual text analytics tool we built in previous studies, *GlassViz*, to automatically detect interesting knowledge associations and groups of common interests between these communities of practice. Our proposed method relies on the statistical modeling of author-assigned keywords to make its findings, which are demonstrated in two use cases. The results show that it is possible to propose interactive, semisupervised visual approaches that aim at defragmenting a body of research using text-based, automatic literature analysis methods.

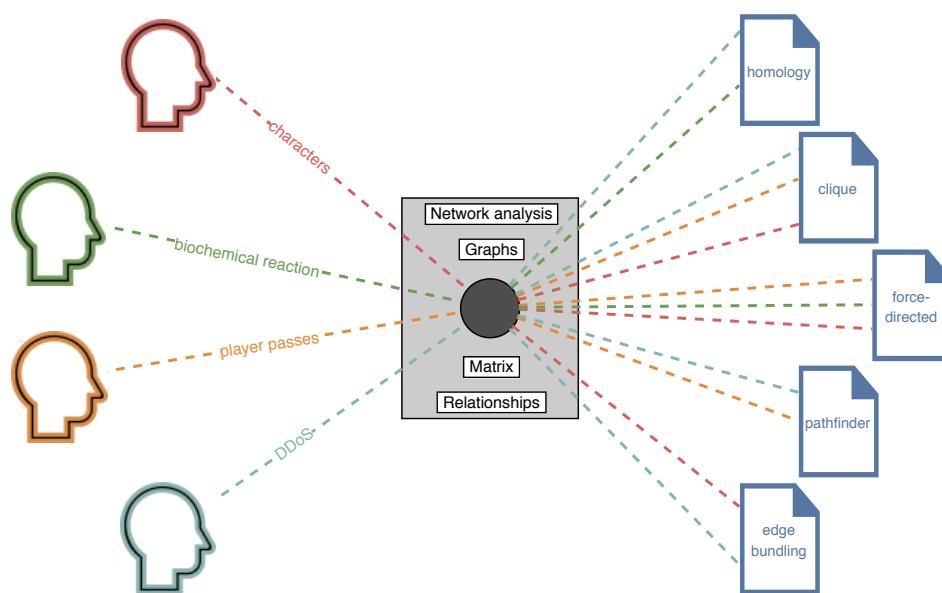
**Keywords:** visual text analytics; problem-driven visualization research; methodology transfer; author-assigned keywords; distributional similarity; knowledge visualization

---

## 1. Introduction

The increasing specialization of science has motivated the surge of different novel interdisciplinary collaborations between research communities in a wide range of domains. This is particularly the case for problem-driven visualization research (PDVR) [1], a type of interdisciplinary practice that connects domain and visualization experts to solve non-trivial, specific domain problems in diverse areas such as biology, city planning, or sports science. In this regard, it is usual that scholars involved in these kinds of collaborations gather in workshops and micro-conferences to discuss each area's particularities, fragmenting visualization research into *communities of practice*. Resulting from their activity, these communities often produce reference publication datasets in a wide variety of focused areas, a fact that reflects the need of these visualization practitioners to obtain information that is tailored to their particular research aims. However, and despite the absolute utility value of these collections, they may also be indicative of the creation of isolated communities within the visualization practice, a fact that could lead to an excess of redundant visualization solutions for generic, domain-agnostic tasks (establishing comparisons, creating summaries, or searching for specific elements) that are replicated across collaborations [2]. Thus, this risk calls for novel approaches that allow a fluid exchange of ideas among practitioners from different knowledge domains to avoid wasting time and human resources that is potentially harming visualization research. To this aim, in recent times, certain authors have started to introduce proposals to facilitate this desirable transfer of knowledge across communities [3], which is known in HCI and visualization research as methodology transfer (MT) [4]. In our recent work, *GlassViz* [5], we contributed a visual text analytics (VTA) tool (built in the Vega-Lite grammar [6] and its Python API Altair [7]) that aims at supporting problem-driven visualization researchers in the

task of exploring large collections of scientific papers by visualizing automatically extracted candidate MTs fit to the researcher's interests, which are represented by an auxiliary collection. To achieve this aim, *GlassViz* finds entry points, which are groups of distributionally related keywords that introduce the user to the corpus, offering a reading order of discovered documents, among other advantages, effectively reducing the cognitive gap involved in the exploration task. This is based on the idea that typically, interdisciplinary researchers employ general keywords to browse a collection of papers to find unknown, potentially interesting techniques that can be used to solve domain-specific problems, which are generally novel and cannot be found in the collection. This idea is depicted in Figure 1, in which four interdisciplinary visualization researchers seek visualization solutions for four different problems in their respective domains. To achieve their goal, these researchers will employ intermediate general terms that these problems are commonly related to (e.g., "network analysis", "graphs", "matrix", or "relationships"). An extensive search employing these terms might unveil visualizations and algorithms in the target collection that they may use to solve the problems at hand.



**Figure 1.** Document exploration model for four different interdisciplinary visualization researchers in the humanities (red), bioinformatics (green), sports science (orange), and security (teal) domains accessing a large collection of research papers containing unknown, potentially interesting techniques (right, in blue) that could be applied to solve their domain-specific problems (e.g., characters, biochemical reaction). Each problem is typically related to the same intermediate concepts (shown at the center), which are the users' *entry points* to the collection. This idea is employed in our study to detect similarities between the different communities of practice and bring them together.

Reflecting on the theory for an interdisciplinary search methodology that we introduced in our previous studies [5,8], we formed the new hypothesis that these groups of shared interests could be automatically detected by analyzing domain-specific literatures on each of the implied areas of knowledge. This hypothesis is the starting point of the work described in this paper, whose main contributions are (1) a set of domain-specific metadata datasets of research papers in three typical kinds of PDVR. These collections are combined with two others compiled by us [9] and other researchers [10] in previous studies in the field and are the input data of our study (described in Section 3); (2) an analysis based on keywords that measures coincidences and differences between keyword sets extracted from the aforementioned collections (Section 4.1); (3) a method to measure distributional similarity between these keywords (Sections 4.2 and 4.3); (4) an analysis of inter-collection similarities as found by our method (Section 4.4); and (5) an enhanced version of our VTA tool, *GlassViz*, to allow the interactive navigation of distributionally-affine sets of domain-specific terms (Sections 5 and 6).

## 2. Related Work

Our contribution is inspired by other works in visualization design, visual analytics, information science, and text mining that we introduce in this section.

### 2.1. Visual Text Analytics of Research Paper Collections

Visual text analytics (VTA) is a novel, text-centered specialization of a broader research discipline known as visual analytics (VA) [11,12] that aims at augmenting the user's analytical capabilities and promoting analytical reasoning on textual data by exploiting the visual pattern recognition mechanisms of the human brain. Concretely, VTA tools deal with unstructured or semistructured text, and they have been typically demonstrated using diverse collections of research papers. In this regard, many authors have combined visualization of multivariate research paper metadata with text mining techniques applied to the papers' contents to create browsable spatializations of a collection. For example, this is the case of Berger et al. [13], who modified the popular word2vec model to provide a joint bidimensional representation of keywords and documents based on citation contexts. In a similar approach to ours, Fried and Kobourov in Maps of Computer Science [14] and Shahaf et al. in Metro Maps of Science [15] employ different graph-based clustering techniques and force-directed layout algorithms to explore a similarity matrix obtained from comparing vectors derived from paper titles and abstracts in the DBLP database.

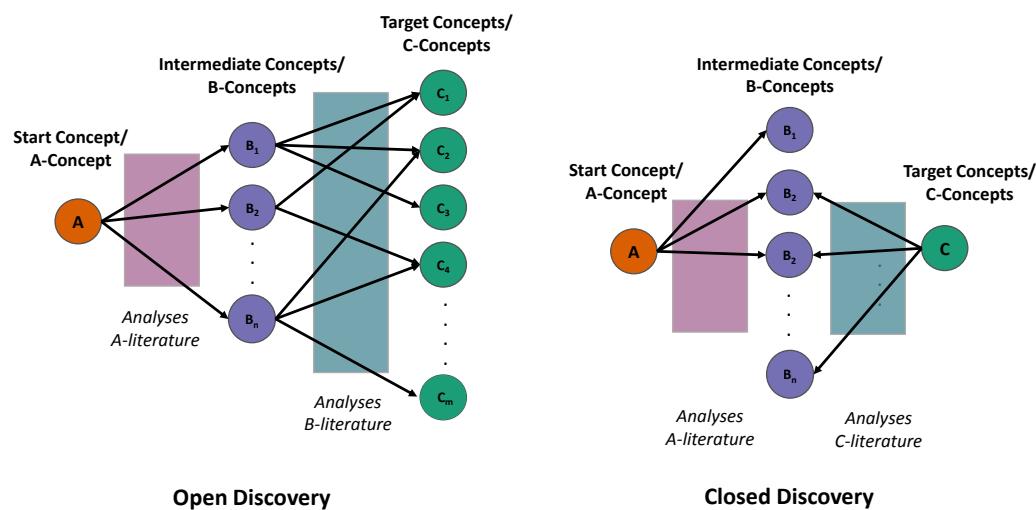
### 2.2. Cognitive Data Visualization

Cognitive data visualization refers to the area of multidisciplinary visualization research that aims to augment the capabilities of the human cognitive system [16]. Researchers in this field employ knowledge originating in cartography, statistics, neuroscience, and ergonomics to design visualizations that accelerate knowledge acquisition. Examples of contributions in this field can typically be found in the proceedings of the *IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* [17], among other venues. There are several examples in the literature of cognitive data visualizations aiming to support the sense-making process of scientific documents collections worth mentioning (see, e.g., in [18]). Specifically, we rely on previous work by Chen, who employs a psychometric graph scaling method, pathfinder networks (PFNETs), to represent semantic spaces found in a collection of conference proceedings [19,20] that were derived from a similarity matrix of co-authorship occurrences using latent semantic analysis (LSA) and singular value decomposition (SVD). This approach was further explored in more recent contributions [21,22] to build complete systems aiming to support the literature review process. In our past work [8], we employed SVD to propose a similar analysis based on author-assigned keywords that we adapted to this study (see Section 4). In particular, we used shortest paths to partition a graph of distances between vector representations of keywords, which in turn served to detect cross-domain affinities between two collections of research papers. The solution is extended and complemented with our recent findings in *GlassViz* [5] to propose an interactive system that enables the exploration of affinities between several communities of practice, as we show in Section 5.

### 2.3. Literature-Based Discovery

Literature-Based Discovery (LBD) is a knowledge extraction technique that aims at making scientific discoveries by connecting what is already available in the literature [23]. The term was coined in the 1980s by Don R. Swanson, an information scientist who followed this method to unveil a relationship between dietary fish oil and Raynaud's disease, a circulatory disorder [24]. To this end, he used the *ABC model*, a method that follows a syllogism to connect terms in two disjoint bodies of literature A and B: if a concept A, exclusive to Literature A, is related to an intermediate concept B that appears in both Literatures A and B, and in turn, this concept B is related to another concept C which is exclusive to literature C, then there is a relationship between the concept A (known by the user)

and the concept C (new to the user) that is characterized by the concept B. The ABC model supports two modes of discovery (open and closed, see Figure 2), and it is currently in the process of being applied to other domains, such as computer science, employing word embeddings [25] extracted from online databases of scientific documents [26]. Our method extends and adapts these ideas to obtain distributional embeddings from author-assigned keywords (see next section), which are used to detect affinities between terms found exclusively on one of the interdisciplinary visualization research areas that we centered our study around (see Section 3).

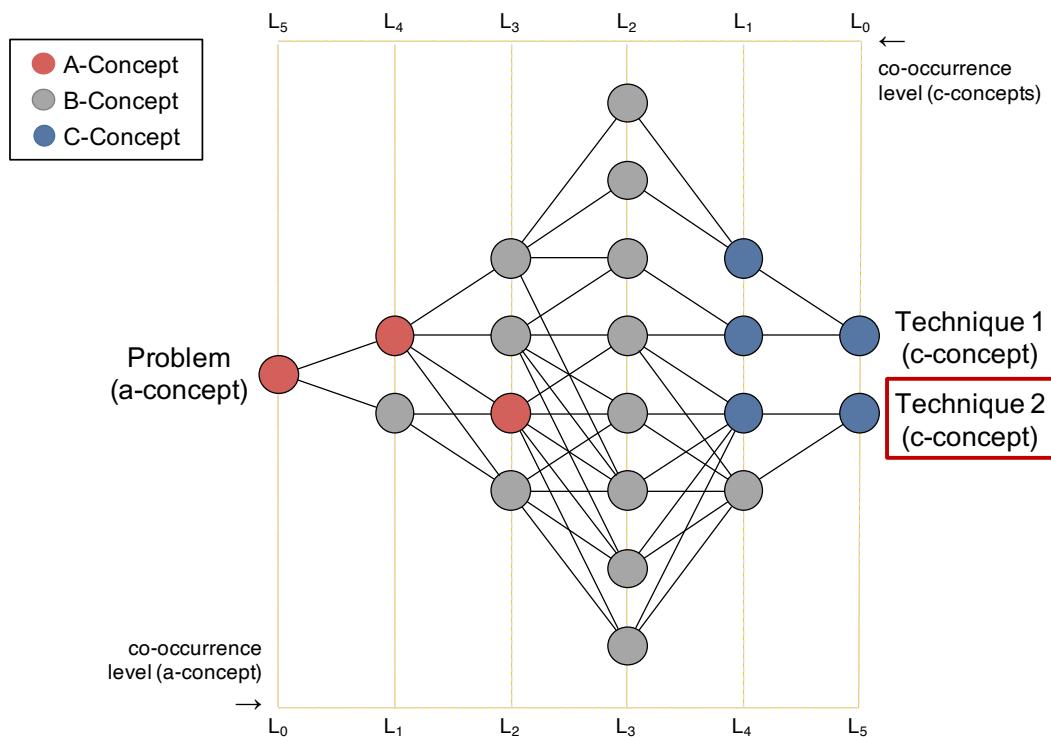


**Figure 2.** Swanson's LBD ABC Model in its open (left) and closed (right) discovery modes (figure from our previous work [8]). The model relates concepts found in two disjoint bodies of scientific literature by identifying connecting b-concepts found in both collections. In the *open discovery* mode, an initial, user-provided term is used to find related c-concepts. This mode is purely exploratory, and it is typically employed in the task of hypothesis generation. Differently, in the *closed discovery* mode, the user provides an a-concept and a c-concept to detect intermediate related b-concepts, generally to validate hypotheses. Our approach employs both modes to explore inter-domain affinities between elements in different bodies of literature.

#### 2.4. Distributional Similarity

Distributional similarity refers to the idea that linguistic items presenting similar distributions in a corpus, which usually appear in the same contexts, have similar meanings [27]. This concept is implemented in different available vector space models that produce vector representations of the words in a corpus. The obtained representations are usually employed to conduct different linguistic tasks, such as similarity or analogy detection and evaluation, or classification. Concerning LBD, similarity evaluation with word embeddings has been tested to automate the LBD workflows presented in the previous section [5,8,23]. Given that word embeddings can capture high order co-occurrence, they seem to be an excellent alternative to discover hidden connections in the scientific literature. This idea is exemplified in Figure 3. The chart shows a starting concept describing a problem found in a body of literature. This concept is connected to two techniques described in a different body of literature by co-occurrence relationships (represented by the edges) between these three concepts and a number of intermediate concepts. By looking at the distribution of the edges in the network, it is easy to see that the c-concept describing the second technique is more related to the problem described by the a-concept than the other c-concept, as the number of high-level co-occurrences between these two concepts is higher for the second c-concept than it is for the first one. Theoretically, this kind of similarity could be detected with arithmetic operations performed on the implied concepts' vector representations of a model trained with this data. Our embeddings are inspired by the proposal by Levy et al. [28], which requires minimal hyperparameter tuning, and it is known to excel at word

similarity tasks [28,29]. To detect affinities between concepts in different areas of interdisciplinary research, we rely on a cosine metric, which is the preferred option for conducting similarity-based tasks, as seen in other works [30,31].



**Figure 3.** Co-occurrence network between a problem found in literature A and two potentially interesting techniques found in literature B. High-order co-occurrence of terms found in the two literatures can be measured to obtain evidence that supports technique 2 as a better option for solving the problem at hand, according to the ABC model.

## 2.5. Methodology Transfer

Methodology Transfer (MT) refers to the practice of reusing available models to provide solutions for novel, unsolved problems. The practice was first introduced into the visualization domain by Burkhard in 2004 [4], who, inspired by previous work by Eppler [32] and standard practices in the domain of architecture, advocated for transferring knowledge between different stakeholders and communities of practice. To this end, he defined a framework of knowledge visualization (as opposed to information visualization), which is defined as "*the use of visual representations to improve the transfer of knowledge between at least two persons or groups of persons*". Since then, the own Burkhard and others have applied the framework to different areas of interdisciplinary visualization practice such as urban planning [33], decision-making support in the medical domain [34], or education [35]. In more recent times, Miller et al. elaborated on Burkhard's ideas to frame a novel research field (visual musicology) as per the principles of methodology transfer [3] using their methodology transfer model (MTM). In our previous work [9], we augmented this model with concepts drawn from LBD to automate the discovery of potential MTs in PDVR [5]. In Section 4, we explain how this model is evolved and adapted to detect concept associations between different areas of interdisciplinary visualization research.

## 3. Data Description

To demonstrate the advantages of our approach, we employ five different collections of research papers in the context of four typical areas of PDVR (VIS4DH, BioVis, SportsVis, and VizSec). The main collection, VIS, is the body of literature that connects the other four collections by providing a large set

of keyword associations that augments and extends those found on each of the domain-specific collections. Specific details about each collection are provided below, which can be found as supplemental materials to this paper.

### 3.1. Domain-Specific Literatures

#### 3.1.1. VIS4DH

This domain-specific collection comprises 221 papers on visualization for the Digital Humanities (VIS4DH) between the years 2016–2019 that were compiled in our previous study [9]. The publications were obtained from two primary sources: The first one, the VIS4DH workshop, is a collocated event with the IEEE VIS set of conferences that gathers researchers working at the intersection of visualization and the humanities to discuss new research directions in visualization and digital humanities research (<https://vis4dh.dbvis.de/>). The second source was obtained from visualization papers located at the humanities side of the collaboration, namely, those published in the ADHO (<https://adho.org/>) Digital Humanities Conference and its peer journal Digital Humanities Quarterly (DHQ) (<http://www.digitalhumanities.org/dhq/>).

#### 3.1.2. BioVis

The second domain-specific collection was specifically compiled for this study and holds publications by researchers interested in biological data visualization. The symposium's main aims are “to educate, inspire, and engage visualization researchers in problems in biological data visualization, as well as bioinformatics and biology researchers in state-of-the-art visualization research”. The workshop started in 2011 as a parallel event with the IEEE Visualization conference but has since then moved to other venues as well: currently, it is a dual meeting taking place at the IEEE VIS Conference and the Conference on Intelligent Systems for Molecular Biology (ISMB). In total, we obtained 69 publications presented at the BioVis (<http://biovis.net/>) symposium between the years 2011–2019.

#### 3.1.3. SportsVis

We wanted to include another important typical area of PDVR in this study: sports data visualization. Although this type of collaboration is also well-established in the visualization practice, and as opposed to the approach we followed to collect the previous datasets that drew publications from discipline-specific venues, such gathering did not exist in this case. The only attempt to hold an event on sports data visualization occurred in 2013 during the IEEE VIS conference in Atlanta with the celebration of the 1st IEEE VIS Workshop on Sports Data Visualization (<http://workshop.sportvis.com/>). Unfortunately, this was the only edition of the event, which did not continue since. Instead, to build a representative dataset of the discipline, we relied on previous work by Perin et al. [36], who created a survey of the state of the art of sports data visualization in 2018. The authors also built a website as companion material of the paper (<https://sportsdaviz.github.io/>) in which they list all works cited in the survey and keep updating regularly. Thus, we built the fourth dataset with all papers appearing in this website that contained author-assigned keywords, which were completed with works presented at the first edition of the VIS Workshop on Sports Data Visualization. In the end, we could collect 59 documents related to this specialty.

### 3.1.4. VizSec

Finally, and following a similar method as in the first two cases, the fourth collection represents publications in visualization for cybersecurity, which is also a long-established area of interdisciplinary visualization research. The main venue that has been regularly capturing contributions in this field since 2004 is the International Workshop/Symposium on Visualization for Cyber Security (VizSec) (<https://vizsec.org/>), from which we obtained 175 papers presented at all its past editions (2004–2019).

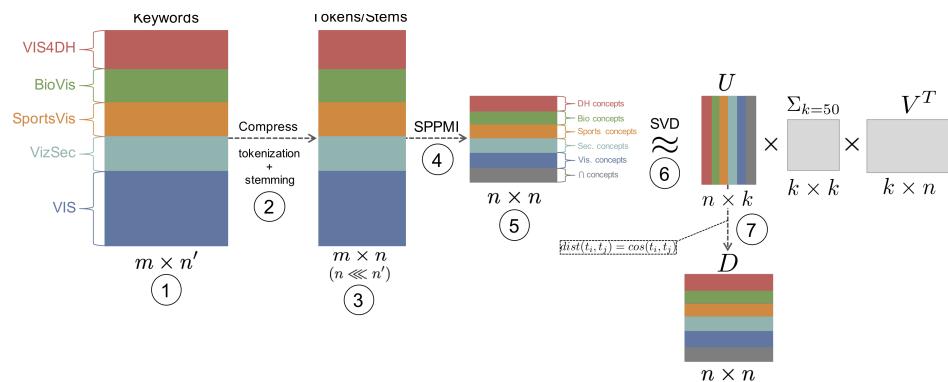
### 3.2. Visualization Literature

The visualization literature (VIS) is a set of 2259 visualization research papers presented at the IEEE set of conferences InfoVis, SciVis, VAST, and Vis between the years 1991 and 2019 that was compiled by other authors [10]. The collection includes a great variety of different algorithms, techniques, problems, and tasks typically related to visualization research in different and diverse application domains. This collection is used to connect the different communities of practice represented by the domain-specific literatures introduced in the previous sections.

## 4. Method

Our method aims to implement a standard visual text mining pipeline that is often seen in many VTA tools. Generally, these pipelines employ diverse well-known text mining algorithms whose results are presented to the user in an interactive graphical interface. To obtain deeper insight into text mining techniques commonly employed for text visualization, we refer the reader to the recent survey by Liu et al. [37], which offers a highly didactic introduction to the topic. Additionally, the work in [38] provides a general introduction to machine learning methods for text analysis.

As explained in previous sections, our implementation aims to detect *significant* inter-collection distributional similarities between exclusive terms appearing in each literature. To this end, we rely on a distance matrix  $S$  that is obtained from comparing dense vector representations of keyword tokens in a vector space model (VSM) using a cosine metric. The process to obtain the vectors from the keyword tokens generated in Section 4.1 is replicated here from our previous work in the field [8], which was inspired by other authors [28]. Figure 4 captures the steps we followed to build the embeddings, which are detailed in this section.



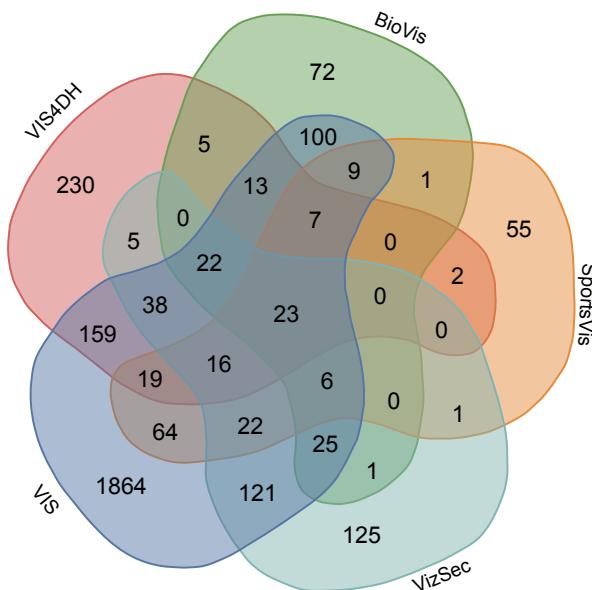
**Figure 4.** Diagram depicting the generation of distributional embeddings from keywords found in research papers. (1) Term-document matrix containing documents in the source (S-) and target (T-) literatures. (2) Keywords are tokenized and stemmed, (3) effectively reducing the number of columns in the term-document matrix. (4) An SPPMI matrix is built from annotating co-occurring keyword tokens in the corpus. (5) Rows in the resulting square matrix are sorted according to each token's provenance. Finally, the matrix is decomposed into the product of three matrices employing singular value decomposition (SVD). Vectors representing each keyword token are obtained from the left singular vectors of the factorization,  $U$  (6). Finally, we derive a distance matrix  $D$  employing a cosine metric (7) and that we use to extract inter-domain similarities.

#### 4.1. Data Processing

Before attempting to create vector representations from the keywords found in the collections presented in the previous section, we prepared the data in the same manner as in our previous studies: first, we built a document-term matrix with keywords extracted from each document (documents with less than two keywords were discarded), tokenized and translated into their American English forms. Tokens matching NLTK’s list of English stop words (e.g., “and” or “of”) were removed from the analysis, which yielded a total of 3005 different tokens. Next, each token was light-stemmed using the Porter algorithm. As author-assigned keywords are a very sparse feature, the stemming procedure had the positive effect of compressing the input vocabulary by linking related lexical forms together under the same root. The number of extracted unique tokens for each collection is shown in Table 1. Intersection sets between the five collections are displayed in Figure 5.

**Table 1.** Number of documents and keyword tokens per collection after processing. Among the four domain-specific collections, VIS4DH held the largest number of unique tokens (different tokens in a collection) and also had the highest ratio of exclusive (not found in any other collection) vs. unique tokens. In total, 659 different tokens could be found in two or more collections.

Dataset	# Documents	# Unique Tokens	Avg. Keyword Tokens per Doc.	# Exclusive Tokens
VIS4DH	221	539	$4.47 \pm 0.99$	230 (42.7%)
BioVis	69	284	$4.57 \pm 1.85$	72 (25.4%)
SportsVis	59	225	$4.73 \pm 1.55$	55 (24.4%)
VizSec	175	405	$4.63 \pm 1.75$	125 (30.9%)
VIS	2253	2508	$4.66 \pm 1.61$	1864 (74.3%)



**Figure 5.** Venn Diagram (generated with the online tool at <http://bioinformatics.psb.ugent.be/webtools/Venn/>) displaying intersections between the five datasets that were employed in our study. The specific keyword sets found on each intersection can be consulted in Table S1 of Supplemental Materials.

#### 4.2. Embedding Generation

The process starts by building a document-term matrix (Figure 4(1))  $D_{m \times n}$  that contains all documents and keywords in the five collections. This matrix was compressed (Figure 4(2)) by tokenizing multi-term keywords and stemming all 1-grams, as explained in f1s4.1(Figure 4(3)). Next, we built a pointwise mutual information (PMI) matrix (Figure 4(4)) that encoded the probability for a pair of tokens to be seen together in a document with respect to seeing those terms separately in the

whole corpus. In this approach, each document is treated as a bag-of-words in which the probabilities  $\text{infleq:pmi}$  can be empirically calculated from the corpus in the following manner; a keyword  $w$  appearing in a set of documents  $D$  with other keywords (its context  $c$ ) can be counted, giving a number  $\#(w, c) \cdot |D|$ . This number is divided by the product of the number of times that keyword appears in the whole corpus ( $\#(w)$ ) and the number of times all the other context keywords appear in the corpus ( $\#(c)$ ). As it is customary [29], we apply a smoothing factor  $\alpha$  to the distribution of each token's context  $P(c)$  (Equation (3)), obtaining  $\hat{P}_\alpha(c)$ , which aims to counteract PMI bias towards very infrequent events. During the experiments described in this paper, we employed  $\alpha = 0.95$ , which seems to work well for keyword similarity tasks and small-sized vocabularies like ours according to our past findings [5,8] (note that  $\alpha$  and the number of dimensions  $k$  are corpus-dependent factors).

$$\text{PMI}(w, c) = \log \frac{P(w, c)}{P(w)P(c)} = \log \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \quad (1)$$

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

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

Finally, and as the matrix SPMI can take infinite negative values when two tokens are never seen in the corpus since  $\log(0) = -\infty$ , we employ a positive version that takes 0 in such cases (Equation (4)).

$$\text{SPPMI}(w, c) = \max(\text{SPMI}(w, c), 0) \quad (4)$$

Then, we annotated the provenance for rows in the  $\text{SPPMI}$  matrix (Figure 4(5)) and recorded whether they appeared exclusively in one of the collection, or in several. The provenance was employed in a later step to detect similarities between elements in different collections (see the next section). The resulting square matrix  $\text{SPPMI}_{n \times n}$  is then factorized into the product of three matrices  $U_{n \times k} \times \Sigma_{k \times k} \times V_{k \times n}^T$  employing singular value decomposition (SVD) (Figure 4(6)), which is a popular algebraic method among NLP scholars that was first employed in the 1990s by the authors behind latent semantic analysis (LSA) [27]. The number of dimensions  $k$  was adjusted to 50, although we obtained similar results with values of  $k$  in the  $50 \pm 10$  range. The vector representations of keyword tokens are the left singular values  $U$  of the decomposition from which we built a distance matrix  $D$ , which we searched for significant inter-collection similarities as we explain the next section.

#### 4.3. Distance Matrix

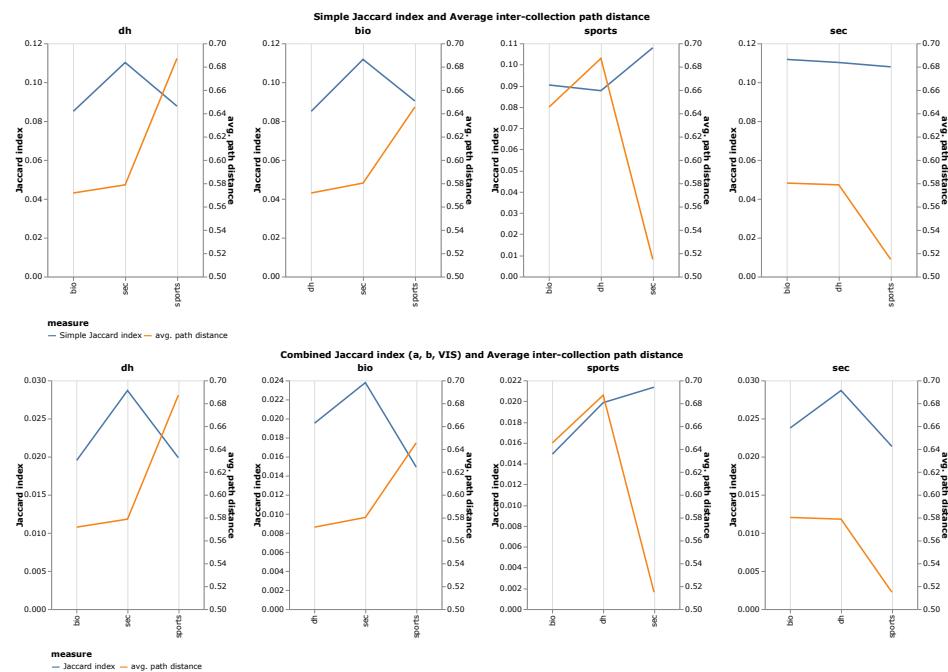
We calculated pairwise distances between keyword embeddings employing a cosine metric (Equation (5)). These distances were later captured in a distance matrix  $D$ . In turn, the matrix  $D$  was converted into a distance graph  $G$  that was later pruned by removing nodes that were not on the shortest paths connecting domain-specific terms in the different literatures.

$$\text{dist}(x, y) = \cos(x, y) = \frac{\sum_{i=1}^n x_i \cdot y_i}{\sqrt{\sum_{i=1}^n x_i^2 \cdot \sum_{i=1}^n y_i^2}} \quad (5)$$

Keyword vectors with pairwise cosine distance equal to or less than 0.01 were considered to be the distributionally identical for this task and were therefore combined into a single representation (509 in total). From this distance matrix, we generated a graph  $G$  that was explored to find interesting interdisciplinary connections. The exploration method we employed is discussed in the next section.

#### 4.4. Finding Interdisciplinary Connections

As discussed in previous sections, this study aimed to capture and visualize interesting interdisciplinary knowledge associations between the different domains represented in the collected sample data. To this end, we partitioned the graph  $G$  using Dijkstra's algorithm to discover least-cost paths connecting every domain-specific term in the four different domain-specific literatures to their closest exclusive tokens in every other literature. After running the algorithm, we obtained 563 shortest paths (note that  $P(a,b) = P(b,a)$ ) for which we annotated their distances and the collections their originating tokens belonged to. Average inter-collection distances are represented in the plot in Figure 6 as orange lines. From these depictions, some information can be decoded: for example, the collections pair presenting the highest average distance between their terms was formed by the VIS4DH and SportsVis datasets, whereas this last one showed the highest similarity of all with the VizSec dataset.

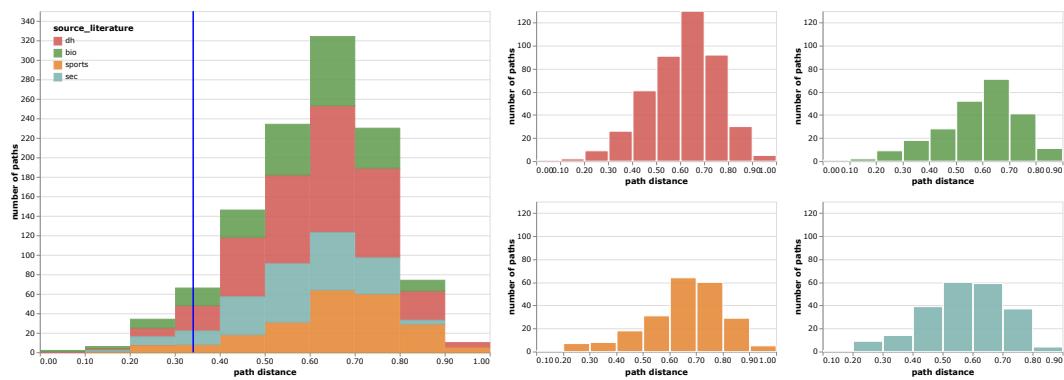


**Figure 6.** Diagrams showing possible interactions between intersection sizes between two collections (**top**) or two collections and the VIS collection (**bottom**) and average path distance between elements in those collections.

At this point, we wondered whether the size of each collection (and thus their number of overlapping tokens with other collections) had any influence on the distances obtained. To answer this question, we calculated three metrics: (1) the average path distance obtained for all tokens on each collection, (2) the Jaccard index between two sets of tokens (defined as  $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ ), and (3) the combined Jaccard index between the two sets of tokens and the VIS collection (which would help us clarify whether the number of overlapping items of both sets with VIS was also influencing the distances in any manner). These two variables were plotted in blue in the charts of f1f6.

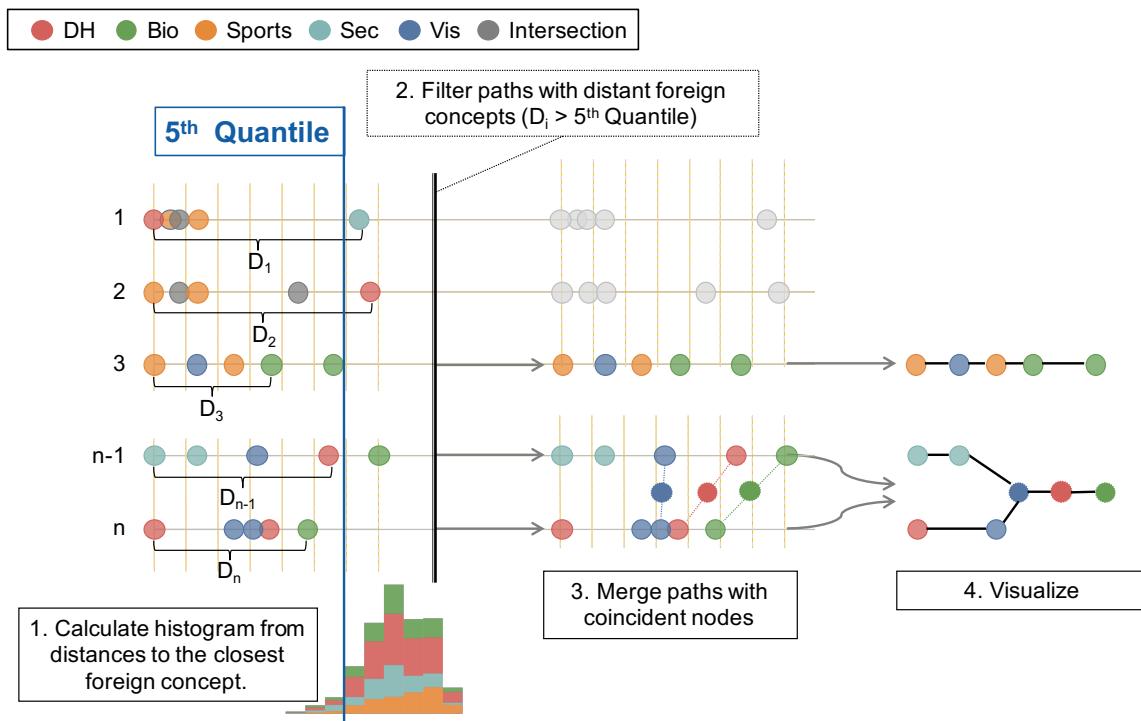
A Spearman correlation ( $p \gg 0.005$ ) test verified what can also be observed in the charts: we could not find any evidence that supported that the average proximity between collections was influenced by the number of overlapping tokens between the collections, and thus neither by the size of each collection, nor by the size of their intersections with the main VIS dataset. A plausible explanation for this fact may be that the similarity score is more influenced by how specific keywords on each collection associate with others in the rest of the dataset. Investigating the exact causes for this observation, however, is something that we considered exceeded the aims of this paper and was left for future studies.

We continued our study by analyzing inter-collection path distances in the histogram in Figure 7. By inspecting the charts, it can be seen that the distances of paths in the four collections approximately follow a Weibull distribution (KS test:  $D = 0.029$ ,  $p > 0.05$ ). In light of these results, we decided to use a cut-off value from the distribution head to filter longer, less interesting connections before moving to the visualization stage.



**Figure 7.** Left: histogram showing the distribution of distances for the found 563 shortest paths connecting domain-specific terms (notice that values on the Y-axis are doubled given that  $P(a, b) = P(b, a)$ ). The blue line shows the 5th percentile ( $x = 0.3408$ ), which was later used as a cut-off value for filtering out longer, and thus less interesting paths that were not visualized in the last stage of the study (see Figure 8). To the right, the same data are disaggregated into four charts, showing similar distance distributions for paths originating at the four collections.

To filter out paths and terms presenting long distances to other terms, we purposely selected a highly restrictive cut-off value (5th percentile,  $x = 0.3408$ ) to focus the visualization on representative connections only. Given that, we assessed that (1) the distribution of distances was similar in the four collections, and (2) the size of each collection did not influence these found distances; this cut-off would obtain a sample of high inter-collection similarities in which all the collections would be evenly represented according to their original sizes. The whole process to select these similarities is shown in Figure 8. After merging paths with coincident nodes, we continued to the next stage in which we employed our tool *GlassViz* to visualize the terms captured in the shortest paths and related documents.

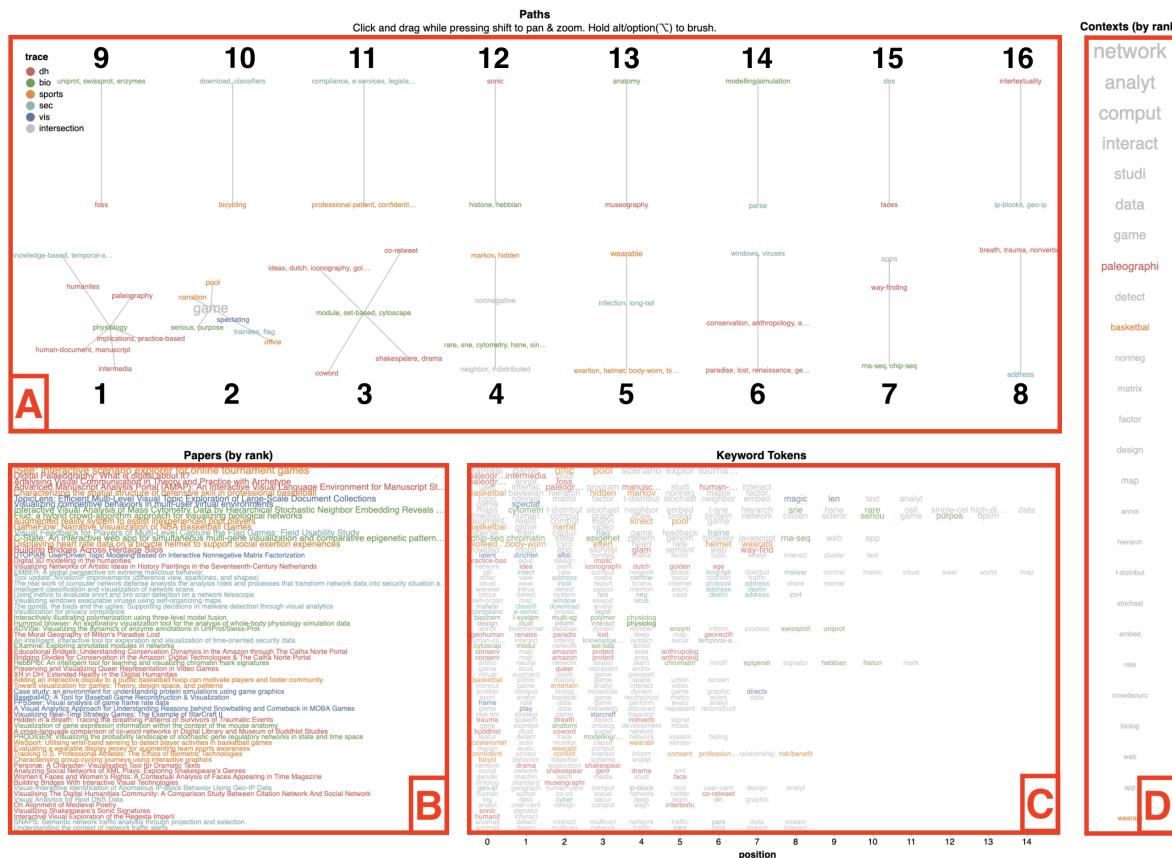


**Figure 8.** Chart depicting the process to compose inter-domain paths. The method relies on calculating a histogram of distances to find a cut-off value used to remove long paths from the final visualization.

## 5. Visualization

After filtering, we obtained 29 paths originating at the VIS4DH, BioVis, SportsVis, and VizSec collections, respectively. After merging paths with coincident tokens, we could identify 50 unique tokens distributed across 16 different components. Using the document-term matrix constructed in Section 4.2, these terms could be mapped to 64 different papers (21 VIS4DH, 10 BioVis, 11 SportsVis, 14 VizSec, and eight VIS). The 16 components were plotted in *GlassViz*'s main view using a node-link graph representation and a force-directed layout algorithm [39] in which the edges capture pairwise similarity (note that  $\text{sim}(x, y) = 1 - \text{dist}(x, y) = 1 - \cos(x, y)$ ) and nodes are tokens in the identified paths (label size is log-scaled to the absolute frequency of the token in the combination of the five corpora). The captured data were plotted in the *GlassViz* interface, which we modified to show paths and related documents detected by our method (shown in Figure 9).

The components were plotted on a designated area of View A, showing labels colored as per the categorical scheme employed throughout this paper, depending on the collection they can be found on (VIS4DH: red; BioVis: green; SportsVis: orange; VizSec: teal). Besides, whenever a token was found in multiple collections, it was colored in gray. Views B, C, and D are rank-frequency lists that show documents and keyword tokens that can be traced to the items selected in view A (in the image, the default selection, all, is applied). Specifically, view B lists documents according to the number of matching tokens with the current selection, meaning that documents that are more relevant to the user's current selection are shown at the top. View C displays keyword tokens for each document in view B in the same order as they were originally processed. Finally, view D provides a visual aggregation of the tokens in C, in which higher-ranked tokens are shown in larger font sizes and are placed closer to the top of the list.



**Figure 9.** Screen capture of *GlassViz* showing the 16 components (view A) extracted in the previous stage Section 4.4 and related documents (view B), along with their keyword tokens (view C) and co-occurring terms with nodes on each component (contexts, view D). We encourage readers to download the high-resolution images companion to this paper which can be zoomed and explored to get a better understanding of how *GlassViz* works.

As we have mentioned, the 16 components of Figure 9 represent different inter-domain associations that can be explored with the aid of *GlassViz*. Although in some cases, such as in components #2 or #3, the main underlying themes can be partially guessed by reading the labels in the graphs, *GlassViz* offers the user the possibility to zoom and brush each component separately to get specific details about a component in the other three auxiliary views. For example, the relation between “sonic” and “histone/hebbian” in component #12 is certainly not obvious and hard to interpret directly. However, if the user brushes this component, the relationship is immediately revealed in view D which, in this case, is “signatures” because the two terms can be mapped to papers in the BioVis and VIS4DH collections that mention it. A compilation of documents, keywords, and contexts for each of the 16 components, as shown in *GlassViz*, is provided in Table S2 of Supplemental Materials.

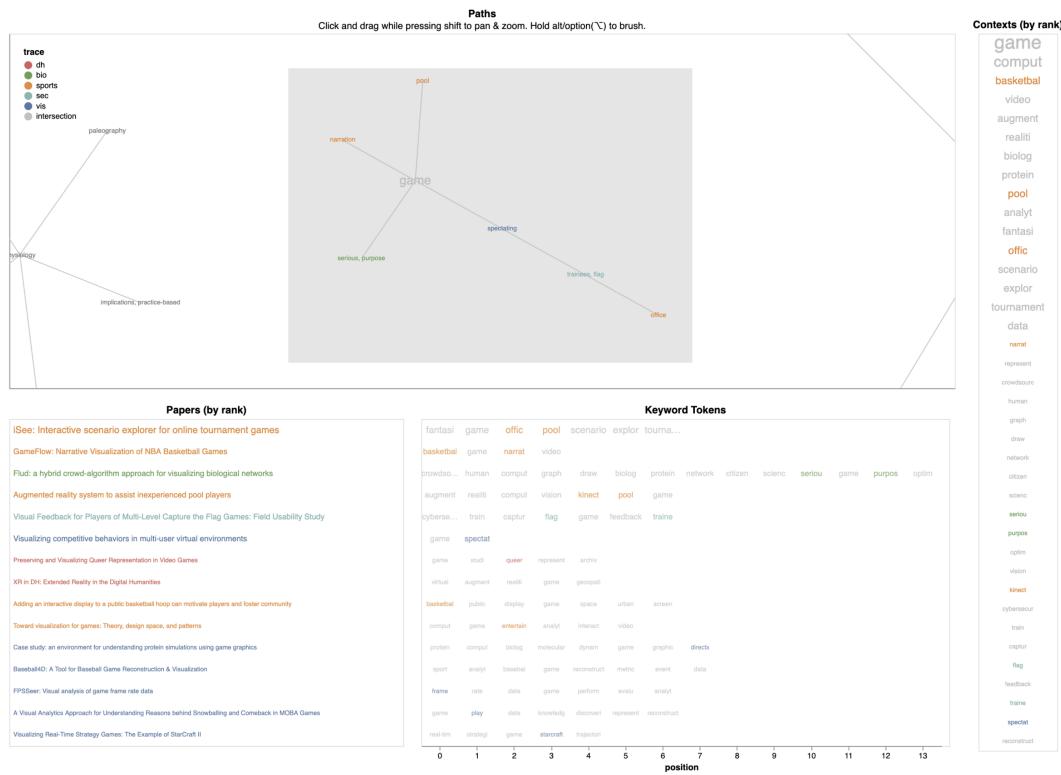
## 6. Use Cases

In this section, we exemplify the advantages of our method in two use cases in which we explore interesting associations of terms and documents. Concretely, we selected those cases in which more documents from distinct collections were captured, namely, components 2 and 4.

### *6.1. Case Study #1: Games and Virtual Reality*

Component #2 links together six different terms appearing in distinct collections (“office”, “pool”, and “narration” in SportsVis; “flag” and “trainee” in VizSec; purpose/serious in “BioVis”; and “spectating” in VIS) which are linked through the term “game/games/gaming” that appears in

all listed documents and it is central to the theme, as it can be observed by its position at the top of view D in Figure 10.

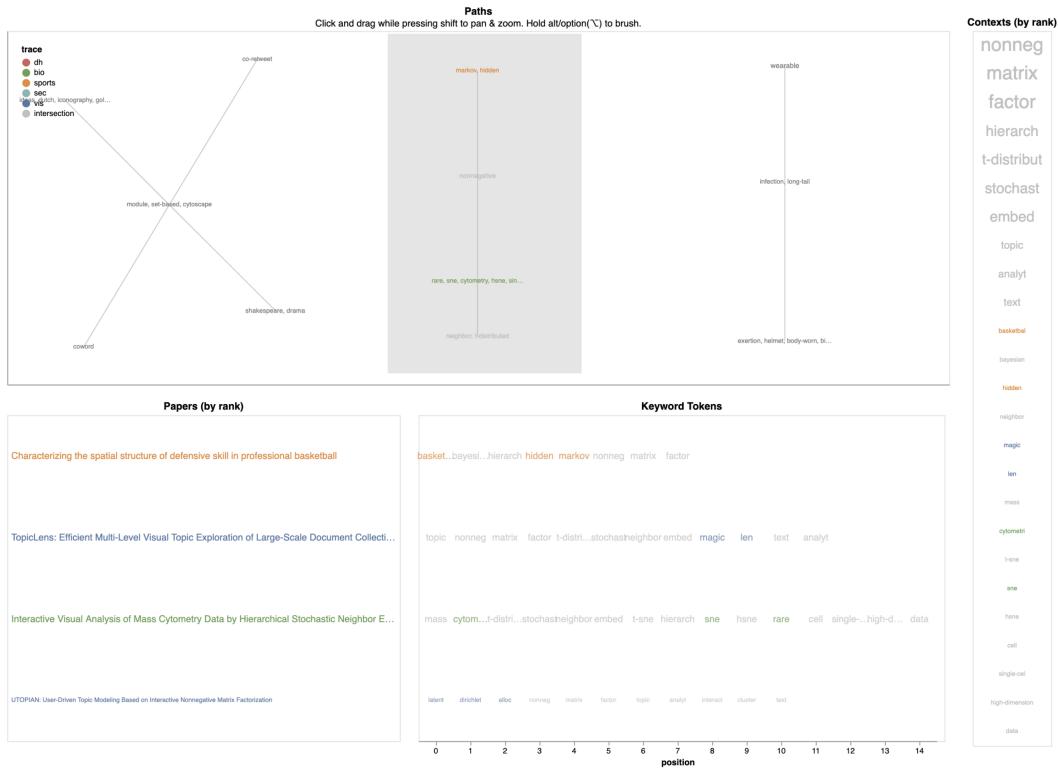


**Figure 10.** Close view of component #2 in *GlassViz*'s main view. We encourage readers to download the high-resolution images companion to this paper which can be zoomed and explored to get a better understanding of how *GlassViz* works.

Further reading of the tokens in view D helps identify other sub-themes that can be found among the documents, the first one being “comput.” for “computational biology” or “computer” for “computer vision” or “computer game”. In relation to this stem, there are four documents included in the list: the first two refer to games applied in the context of “computational biology” (documents 1.3 and 1.11, from collections BioViS and VIS, respectively), whereas the other two documents are associated to computers, as in “computer vision,” and “computer game.” (documents 1.4 and 1.10, both found in the SportsVis collection). Finally, another important sub-theme can also be identified, formed by the tokens “augment” and “realiti”. If we inspect documents containing these terms, we find document 1.4 again, and also 1.8, this last one pertaining to the VIS4DH collection. The papers describe two research experiences with augmented reality that are highly related as found by our model: the first one was built to assist novices in the game of pool, whereas the second one describes different pedagogical experiences in the humanities domain, forming a potentially interesting pair for knowledge transfer.

## 6.2. Case Study #2: Topic Models and Interaction Techniques

The next example is derived from examining component #4, and comprises fewer elements (only four in this case) which form a smaller and more concise theme than in the previous case. Again, the general composition of the theme is revealed in view D (Figure 11), whose top items refer to different topic modeling (e.g., “non-negative matrix factorization” and “latent dirichlet allocation”) and dimensionality reduction techniques (“t-SNE” and “h-SNE”) often employed in visualization approaches conceived to support classification tasks.



**Figure 11.** Close view of component #4 in *GlassViz*'s main view. We encourage readers to download the high-resolution images companion to this paper which can be zoomed and explored to get a better understanding of how *GlassViz* works.

The inspection of the documents captured under this component shows four documents pertaining to three different collections (SportsVis, document 4.1; BioVis, document 4.3; and VIS, documents 4.2 and 4.4) that are centered around the aforementioned main concepts. Using the graph representation of the component, more information can be decoded: for example, it can be seen that the domain-specific keywords shown in green (“rare”, “sne”, and “cytometry”) are placed closer to the terms “neighbor” and “t-distributed” in the lower part of the chart. This effect can be understood by manually inspecting the documents and keywords in views B and C: there, it can be seen that the authors of paper 4.3 employ a variation of t-sne, h-sne, to visualize mass cytometry data. Relatedly, the authors of paper 4.2 propose an interaction technique based on t-sne and the magic lens metaphor [40] to inspect topic models in textual data. These authors also employ the topic modeling technique “non-negative matrix factorization” to conduct their study, which is in turn used by the authors in publications 4.4 and 4.1 and serves as the connecting theme of the component.

## 7. Future Work

In previous sections, we have presented a proposal for automatically detecting shared interests between different communities of practice in PDVR. The results presented in Sections 4.1, 5, and 6 show that keywords carry great implicit knowledge by the authors that deserves being studied and analyzed in full. Although we are aware that the study of the language of keywords has many beneficial implications in science, in this paper we have seen how it can be used to determine important points of confluence between a priori unrelated groups of researchers, which represents an advancement towards addressing the critical problem of knowledge fragmentation in modern science. Not only knowing the number, but also understanding the manner in which the fundamental semantic components of keywords are combined, may open up novel ways to obtain holistic panoramas of science that may help overcome some of science’s current difficulties. In this contribution, we proposed a model and a VTA tool to capture and explore rare conceptual associations between research areas that would be hard to

find for a human actor. Although we consider the work presented in this paper to be still ongoing, and despite the development of *GlassViz* still is in its early stages, in light of the results, we are positive about the results and aim to keep improving the system in future research to cover more datasets and support more complex use cases. In this regard, we aim to improve *GlassViz*'s interactivity, which at the moment is rather limited. For example, it is currently not possible to obtain information on how the different components in view A are related to each other, and also how they relate to other parts in the different collection. To address this issue, we are currently conducting experiments on our data employing a novel dimensionality reduction technique, uniform manifold approximation and projection (UMAP) [41], that is showing very promising results. This would allow us to obtain a joint projection of documents and keywords, which should be preferred to our linked views approach due to its reduced cognitive load. In addition, a combination of UMAP with hierarchical density-based clustering (HDBSCAN) [42] could offer an automatic way to cluster connected components into larger thematic areas, an addition that would yield great opportunities for implementing direct manipulation interaction techniques [43]. This would in turn allow us to receive fine-grained information from the user to, for example, denormalize certain terms that were linked together by the stemming algorithm and that the user may want to split. These cases are usually hard to detect by automatic means, as they much depend on the user's own aims of the exploration. By adopting direct manipulation principles, the user could drive the execution of the algorithm at each step to obtain personalized results seamlessly.

## 8. Summary

In this paper, we have presented a study on keywords to identify thematic similarities and potential methodology transfers between different areas of PDVR. Our approach was supported by the collection and composition of four different datasets that represented keyword associations made by authors of research papers in diverse interdisciplinary visualization research areas. In addition, we proved that there is evidence to support the hypothesis that text-based, automatic methods to accomplish the aim of connecting communities of practice within a body of research may be proposed. Beyond that, we believe our approach could be further extended to other areas experiencing the same fragmentation. In this regard, we desire that our work serves to inspire future researchers to build more complex VTA tools that address this issue of modern science.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2076-3417/10/20/7248/s1>, Table S1: Keyword intersections, Table S2: Components, Dataset D1: publications datasets with keywords.

**Author Contributions:** Conceptualization, A.B.-S.; formal analysis, A.B.-S.; investigation, A.B.-S.; writing—original draft preparation, A.B.-S.; writing—review and editing, A.B.-S. and R.T.S.; supervision, R.T.S.; project administration, R.T.S.; funding acquisition, R.T.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work has received funding within the CHIST-ERA programme under the following national grant agreement PCIN-2017-064 (MINECO Spain).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript.

PDVR	Problem-Driven Visualization Research
LBD	Literature-Based Discovery
MT	Methodology Transfer
MTM	Methodology Transfer Model
VTA	Visual Text Analytics
DH	Digital Humanities

## References

1. Simon, S.; Mittelstädt, S.; Keim, D.A.; Sedlmair, M. Bridging the Gap of Domain and Visualization Experts with a Liaison. In *Proceedings of the Eurographics Conference on Visualization (EuroVis)*; The Eurographics Association: Cagliari, Italy, 2015; Volume 2015.
2. Brehmer, M.; Munzner, T. A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 2376–2385, doi:10.1109/TVCG.2013.124.
3. Miller, M.; Schäfer, H.; Kraus, M.; Leman, M.; Keim, D.A.; El-Assady, M. Framing Visual Musicology through Methodology Transfer. In Proceeding of the 4th Workshop on Visualization for the Digital Humanities (VIS4DH), Vancouver, BC, Canada, 20 October 2019.
4. Burkhard, R. Learning from Architects: The Difference between Knowledge Visualization and Information Visualization. In Proceedings of the Eighth International Conference on Information Visualisation, London, UK, 16 July 2004; pp. 519–524, doi:10.1109/IV.2004.1320194.
5. Benito-Santos, A.; Therón, R. GlassViz: Visualizing Automatically-Extracted Entry Points for Exploring Scientific Corpora in Problem-Driven Visualization Research. 2020 IEEE Visualization Conference (VIS), 2020, p. To appear in IEEE VIS 2020 Conference Proceedings. Available online: <https://arxiv.org/abs/2009.02094> (accessed on 16 October 2020).
6. Satyanarayan, A.; Moritz, D.; Wongsuphasawat, K.; Heer, J. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 341–350, doi:10.1109/TVCG.2016.2599030.
7. VanderPlas, J.; Granger, B.; Heer, J.; Moritz, D.; Wongsuphasawat, K.; Lees, E.; Timofeev, I.; Welsh, B.; Sievert, S. Altair: Interactive Statistical Visualizations for Python. *J. Open Source Softw.* **2018**, *3*, 1057, doi:10.21105/joss.01057.
8. Benito-Santos, A.; Therón Sánchez, R. Cross-Domain Visual Exploration of Academic Corpora via the Latent Meaning of User-Authored Keywords. *IEEE Access* **2019**, *7*, 98144–98160.
9. Benito-Santos, A.; Therón Sánchez, R. A Data-Driven Introduction to Authors, Readings and Techniques in Visualization for the Digital Humanities. *IEEE Comput. Graph. Appl.* **2020**, *40*, 45–57.
10. Isenberg, P.; Heimerl, F.; Koch, S.; Isenberg, T.; Xu, P.; Stolper, C.D.; Sedlmair, M.; Chen, J.; Möller, T.; Stasko, J. Vispubdata.Org: A Metadata Collection About IEEE Visualization (VIS) Publications. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 2199–2206, doi:10.1109/TVCG.2016.2615308.
11. Thomas, J.J.; Cook, K.A. A Visual Analytics Agenda. *IEEE Comput. Graph. Appl.* **2006**, *26*, 10–13, doi:10.1109/MCG.2006.5.
12. Keim, D.A.; Mansmann, F.; Schneidewind, J.; Thomas, J.; Ziegler, H. Visual Analytics: Scope and Challenges. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*; Simoff, S.J., Böhnen, M.H., Mazeika, A., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; pp. 76–90, doi:10.1007/978-3-540-71080-6\_6.
13. Berger, M.; McDonough, K.; Seversky, L.M. Cite2vec: Citation-Driven Document Exploration via Word Embeddings. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 691–700, doi:10.1109/TVCG.2016.2598667.
14. Fried, D.; Kobourov, S.G. Maps of Computer Science. In Proceedings of the 2014 IEEE Pacific Visualization Symposium, Yokohama, Japan, 4–7 March 2014 ; pp. 113–120, doi:10.1109/PacificVis.2014.47.
15. Shahaf, D.; Guestrin, C.; Horvitz, E. Metro Maps of Science. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2012; pp. 1122–1130, doi:10.1145/2339530.2339706.
16. Török, Z.G.; Török, Á. Cognitive Data Visualization—A New Field with a Long History. In *Cognitive Infocommunications, Theory and Applications*; Topics in Intelligent Engineering and Informatics; Klempous, R.; Nikodem, J.; Baranyi, P.Z., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 49–77, doi:10.1007/978-3-319-95996-2\_3.
17. Shakhnov, V.; Zinchenko, L.; Makarchuk, V.; Verstov, V. Visual Analytics Support for the SOI VLSI Layout Design for Multiple Patterning Technology. In Proceedings of the 2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Györ, Hungary, 19–21 October 2015; pp. 67–70, doi:10.1109/CogInfoCom.2015.7390566.
18. Soós, S.; Vida, Z. Topic Overlay Maps and the Cognitive Structure of Policy-Related SSH. In Proceedings of the 2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom), Vietri sul Mare, Italy, 5–7 November 2014; pp. 413–418, doi:10.1109/CogInfoCom.2014.7020490.

19. Chen, C. Visualising Semantic Spaces and Author Co-Citation Networks in Digital Libraries. *Inf. Process. Manag.* **1999**, *35*, 401–420, doi:10.1016/S0306-4573(98)00068-5.
20. Chen, C.; Kulpis, J.; Paul, R.J. Visualizing Latent Domain Knowledge. *IEEE Trans. Syst. Man Cybern. Part Appl. Rev.* **2001**, *31*, 518–529, doi:10.1109/5326.983935.
21. Chen, T.T. The Development and Empirical Study of a Literature Review Aiding System. *Scientometrics* **2012**, *92*, 105–116, doi:10.1007/s11192-012-0728-3.
22. Godwin, A. Visualizing Systematic Literature Reviews to Identify New Areas of Research. In Proceedings of the 2016 IEEE Frontiers in Education Conference (FIE), Erie, PA, USA, 12–15 October 2016; pp. 1–8, doi:10.1109/FIE.2016.7757690.
23. Thilakaratne, M.; Falkner, K.; Atapattu, T. A Systematic Review on Literature-Based Discovery. *Acm Comput. Surv. (CSUR)* **2019**, *5*, e235.
24. Swanson, D.R. Fish Oil, Raynaud’s Syndrome, and Undiscovered Public Knowledge. *Perspect. Biol. Med.* **1986**, *30*, 7–18, doi:10.1353/pbm.1986.0087.
25. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems 26*; Burges, C.J.C.; Bottou, L.; Welling, M.; Ghahramani, Z.; Weinberger, K.Q., Eds.; Curran Associates, Inc.: NY 12571, USA, 2013; pp. 3111–3119.
26. Thilakaratne, M.; Falkner, K.; Atapattu, T. Automatic Detection of Cross-Disciplinary Knowledge Associations. In Proceedings of ACL Student Research Workshop; Association for Computational Linguistics: Stroudsburg, Pennsylvania, USA, 2018; pp. 45–51.
27. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407, doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.
28. Levy, O.; Goldberg, Y. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: NY 12571, USA, 2014; pp. 2177–2185.
29. Levy, O.; Goldberg, Y.; Dagan, I. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 211–225.
30. Heimerl, F.; Han, Q.; Koch, S.; Ertl, T. CiteRivers: Visual Analytics of Citation Patterns. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 190–199, doi:10.1109/TVCG.2015.2467621.
31. Günther, F.; Dudschig, C.; Kaup, B. Latent Semantic Analysis Cosines As a Cognitive Similarity Measure: Evidence from Priming Studies. *Q. J. Exp. Psychol.* **2016**, *69*, 626–653, doi:10.1080/17470218.2015.1038280.
32. Eppler, M.J. Visuelle Kommunikation—Der Einsatz von graphischen Metaphern zur Optimierung des Wissenstransfers. In *Wissenskommunikation in Organisationen: Methoden · Instrumente · Theorien*; Reinhardt, R., Eppler, M.J., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 13–31, doi:10.1007/978-3-642-17130-7\_2.
33. Burkhard, R.A. Strategy Visualization: A New Research Focus in Knowledge Visualization and a Case Study. *Proc. Know* **2005**, *5*, 1–8.
34. Elouni, J.; Ltifi, H.; Ayed, M.B. Knowledge Visualization Model for Intelligent Dynamic Decision-Making. Hybrid Intelligent Systems; Abraham, A.; Han, S.Y.; Al-Sharhan, S.A.; Liu, H., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2016; pp. 223–235, doi:10.1007/978-3-319-27221-4\_19.
35. Fadiran, O.A.; van Biljon, J.; Schoeman, M.A. How Can Visualisation Principles Be Used to Support Knowledge Transfer in Teaching and Learning? In Proceedings of the 2018 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 8–9 March 2018; pp. 1–6, doi:10.1109/ICTAS.2018.8368739.
36. Perin, C.; Vuillemot, R.; Stolper, C.D.; Stasko, J.T.; Wood, J.; Carpendale, S. State of the Art of Sports Data Visualization. *Comput. Graph. Forum* **2018**, *37*, 663–686, doi:10.1111/cgf.13447.
37. Liu, S.; Wang, X.; Collins, C.; Dou, W.; Ouyang, F.; El-Assady, M.; Jiang, L.; Keim, D. Bridging Text Visualization and Mining: A Task-Driven Survey. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 2482–2504, doi:10.1109/TVCG.2018.2834341.
38. Aggarwal, C.C. *Machine Learning for Text*; Springer International Publishing: Cham, Switzerland, 2018, doi:10.1007/978-3-319-73531-3.
39. Fruchterman, T.M.J.; Reingold, E.M. Graph Drawing by Force-Directed Placement. *Software: Pract. Exp.* **1991**, *21*, 1129–1164, doi:10.1002/spe.4380211102.

40. Bier, E.A.; Stone, M.C.; Pier, K.; Buxton, W.; DeRose, T.D. Toolglass and Magic Lenses: The See-through Interface. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*; Association for Computing Machinery: New York, NY, USA, 1993; pp. 73–80, doi:10.1145/166117.166126.
41. McInnes, L.; Healy, J.; Saul, N.; Grossberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861.
42. McInnes, L.; Healy, J.; Astels, S. HdbSCAN: Hierarchical Density Based Clustering. *J. Open Source Softw.* **2017**, *2*, 205, doi:10.21105/joss.00205.
43. El-Assady, M.; Kehlbeck, R.; Collins, C.; Keim, D.; Deussen, O. Semantic Concept Spaces: Guided Topic Model Refinement Using Word-Embedding Projections. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 1001–1011, doi:10.1109/TVCG.2019.2934654.

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

# Bibliography

*No man is an island entire of itself; every man  
is a piece of the continent, a part of the main.*

John Donne - Excerpt from Meditation XVII

- [BEC<sup>+</sup>18] A. J. Bradley, M. El-Assady, K. Coles, E. Alexander, M. Chen, C. Collins, S. Jänicke, and D. J. Wrisley. Visualization and the Digital Humanities: Moving Toward Stronger Collaborations. *IEEE Computer Graphics and Applications*, 38(6):26–38, November 2018.
- [BMS17] M. Berger, K. McDonough, and L. M. Seversky. Cite2vec: Citation-Driven Document Exploration via Word Embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):691–700, January 2017.
- [BO06] L J Bracken and E A Oughton. ‘What do you mean?’ The importance of language in developing interdisciplinary research. *Transactions of the Institute of British Geographers*, 31(3):371–382, July 2006.
- [BSTS16] Alejandro Benito Santos and Roberto Therón Sánchez. *Visualización de Datos En Humanidades Digitales*. Tesis de Master, Universidad de Salamanca, España, 2016.
- [BTL<sup>+</sup>17] Alejandro Benito, Roberto Therón, Antonio Losada, Eveline Wandl-Vogt, and Amelie Dorn. Exploring Lemma Interconnections in Historical Dictionaries. In *Proc. 2nd Workshop on Visualization for the Digital Humanities (VIS4DH)*, 2017.
- [BTL<sup>+</sup>18] Alejandro Benito-Santos, Roberto Theron, Antonio Losada, Jaime E. Sampaio, and Carlos Lago-Peñas. Data-Driven Visual Performance

- Analysis in Soccer: An Exploratory Prototype. *Frontiers in Psychology*, 9, 2018.
- [Bur04] R.A. Burkhard. Learning from architects: The difference between knowledge visualization and information visualization. In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, pages 519–524, July 2004.
- [Bur05] Remo Aslak Burkhard. Strategy visualization: A new research focus in knowledge visualization and a case study. In *Proceedings of I-Know*, volume 5, pages 1–8, 2005.
- [Che97] Chaomei Chen. Tracking latent domain structures: An integration of pathfinder and Latent Semantic Analysis. *AI & SOCIETY*, 11(1):48–62, March 1997.
- [Che99] Chaomei Chen. Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management*, 35(3):401–420, May 1999.
- [CKP01] C. Chen, J. Kuljis, and R. J. Paul. Visualizing latent domain knowledge. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 31(4):518–529, November 2001.
- [Cox87] D Cox. Computer art/design curricula in universities: Beyond the traditional approach. *Teaching computer graphics: An interdisciplinary approach*, pages 207–233, 1987.
- [DSG<sup>+</sup>12] Cody Dunne, Ben Shneiderman, Robert Gove, Judith Klavans, and Bonnie Dorr. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12):2351–2369, 2012.
- [DWCR11] W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 231–240, October 2011.
- [ELA16] Jihed Elouni, Hela Ltifi, and Mounir Ben Ayed. Knowledge Visualization Model for Intelligent Dynamic Decision-Making. In Ajith Abraham,

- Sang Yong Han, Salah A. Al-Sharhan, and Hongbo Liu, editors, *Hybrid Intelligent Systems, Advances in Intelligent Systems and Computing*, pages 223–235, Cham, 2016. Springer International Publishing.
- [Epp04] Martin J. Eppler. Visuelle Kommunikation — Der Einsatz von graphischen Metaphern zur Optimierung des Wissenstransfers. In Rüdiger Reinhardt and Martin J. Eppler, editors, *Wissenskommunikation in Organisationen: Methoden · Instrumente · Theorien*, pages 13–31. Springer, Berlin, Heidelberg, 2004.
- [FvS18] Olakumbi A. Fadiran, Judy van Biljon, and Marthie A. Schoeman. How can visualisation principles be used to support knowledge transfer in teaching and learning? In *2018 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6, March 2018.
- [GL18] H. Guo and D. H. Laidlaw. Topic-based Exploration and Embedded Visualizations for Research Idea Generation. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2018.
- [HHKE16] F. Heimerl, Q. Han, S. Koch, and T. Ertl. CiteRivers: Visual Analytics of Citation Patterns. *IEEE Transactions on Visualization & Computer Graphics*, 22(1):190–199, January 2016.
- [HJQ<sup>+</sup>16] F. Heimerl, M. John, Qi Han, S. Koch, and T. Ertl. DocuCompass: Effective exploration of document landscapes. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 11–20, October 2016.
- [HM17] Sam Henry and Bridget T. McInnes. Literature Based Discovery: Models, methods, and trends. *Journal of Biomedical Informatics*, 74:20–32, October 2017.
- [HPLC19] Jiangen He, Qing Ping, Wen Lou, and Chaomei Chen. PaperPoles: Facilitating adaptive visual exploration of scientific publications by citation links. *Journal of the Association for Information Science and Technology*, 70(8):843–857, 2019.
- [IHK<sup>+</sup>17] 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*

- Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, September 2017.
- [IIS<sup>+</sup>14a] Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. Toward a deeper understanding of visualization through keyword analysis. Technical Report RR-8580, INRIA, France, August 2014.
  - [IIS<sup>+</sup>14b] Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. Visualization According To Research Paper Keywords. Technical report, INRIA, France, 2014.
  - [IIS<sup>+</sup>17] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller. Visualization as Seen through its Research Paper Keywords. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):771–780, January 2017.
  - [KAF<sup>+</sup>08] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual Analytics: Definition, Process, and Challenges. In Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, editors, *Information Visualization: Human-Centered Issues and Perspectives*, Lecture Notes in Computer Science, pages 154–175. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
  - [KK89] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 1989.
  - [KKP<sup>+</sup>17] M. Kim, K. Kang, D. Park, J. Choo, and N. Elmquist. TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):151–160, January 2017.
  - [KM13] R. M. Kirby and M. Meyer. Visualization Collaborations: What Works and Why. *IEEE Computer Graphics and Applications*, 33(6):82–88, Nov.-Dec. 2013.
  - [KMS<sup>+</sup>08] Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual Analytics: Scope and Challenges. In Simeon J. Simoff, Michael H. Böhlen, and Arturas Mazeika, editors, *Visual Data Mining: Theory, Techniques and Tools for Visual Analyt-*

- ics*, Lecture Notes in Computer Science, pages 76–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- [LG14] Omer Levy and Yoav Goldberg. Neural Word Embedding As Implicit Matrix Factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, pages 2177–2185, Cambridge, MA, USA, 2014. MIT Press.
- [LGD15] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3(0):211–225, May 2015.
- [LGS<sup>+</sup>14] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, December 2014.
- [LTB16] Antonio G Losada, Roberto Therón, and Alejandro Benito. Bkviz: A Basketball Visual Analysis Tool. *IEEE Computer Graphics and Applications*, 36(6):58–68, 2016.
- [McC87] Bruce Howard McCormick. Visualization in scientific computing. *Computer graphics*, 21(6), 1987.
- [MHSG18] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.
- [MSC<sup>+</sup>13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.
- [MSK<sup>+</sup>19] Matthias Miller, Hanna Schäfer, Matthias Kraus, Marc Leman, Daniel A. Keim, and Mennatallah El-Assady. Framing Visual Musicology through Methodology Transfer. *Proceedings of the Workshop on Visualization for the Digital Humanities (VIS4DH) at IEEE VIS 2019*, October 2019.
- [PC05] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analy-

- sis. In *Proceedings of International Conference on Intelligence Analysis*, pages 2–4, 2005.
- [PC18] Qing Ping and Chaomei Chen. LitStoryTeller+: An Interactive System for Multi-level Scientific Paper Visual Storytelling with a Supportive Text Mining Toolbox. *Scientometrics*, 116(3):1887–1944, September 2018.
- [Ped17] Thomas Lin Pedersen. Hierarchical sets: Analyzing pangenome structure through scalable set visualizations. *Bioinformatics*, 33(11):1604–1612, June 2017.
- [PEM16] Antoine Ponsard, Francisco Escalona, and Tamara Munzner. PaperQuest: A Visualization Tool to Support Literature Review. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA ’16, pages 2264–2271, San Jose, California, USA, May 2016. Association for Computing Machinery.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.
- [SMKS15] Svenja Simon, Sebastian Mittelstädt, Daniel A. Keim, and Michael Sedlmair. Bridging the gap of domain and visualization experts with a Liaison. In *Accepted at the Eurographics Conference on Visualization (EuroVis 2015, Short Paper)*, volume 2015, Cagliari, Italy, 2015. The Eurographics Association.
- [SMM12] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, December 2012.
- [Swa86] Don R. Swanson. Fish Oil, Raynaud’s Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.
- [Swa88] Don R. Swanson. Migraine and Magnesium: Eleven Neglected Connections. *Perspectives in Biology and Medicine*, 31(4):526–557, 1988.

- [TC06] J. J. Thomas and K. A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, January 2006.
- [TFA18] Menasha Thilakaratne, Katrina Falkner, and Thushari Atapattu. Automatic Detection of Cross-Disciplinary Knowledge Associations. In *Proceedings of ACL 2018, Student Research Workshop*, pages 45–51, July 2018.
- [TFA19] Menasha Thilakaratne, Katrina Falkner, and Thushari Atapattu. A Systematic Review on Literature-based Discovery. *ACM Computing Surveys (CSUR)*, December 2019.
- [WLQ<sup>+</sup>16] Yun Wang, Dongyu Liu, Huamin Qu, Qiong Luo, and Xiaojuan Ma. A Guided Tour of Literature Review: Facilitating Academic Paper Reading with Narrative Visualization. In *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction*, VINCI ’16, pages 17–24, New York, NY, USA, 2016. ACM.

*Education isn't something you can finish.*

*Isaac Asimov*

