An Interactive Tool to Develop Literature Reviews

Helmut Simonis[1], Cemalettin Öztürk[2]

[1]Insight SFI Centre for Data Analytics, Department for Computer Science and Information Technology, University College Cork, Cork, Ireland
[2]Munster Technological University, Bishopstown, Cork, Ireland

Abstract

In this short paper, we present an interactive tool to help create literature surveys of scientific literature. The tool uses existing web services to find papers related to an initial seed library and uses a user-defined ontology to consider which papers found are relevant to the survey. Examples show that comprehensive literature surveys can be developed in a short time period, while leaving the user in control of the overall process. The Java sources for the tool are available online.

Keywords

Literature Survey, Bibliography, Bibliometric, LaTeX

1. Motivation

Building a comprehensive literature review of a given field [1] is part of any new project we have to undertake, but this can be a time-consuming and tedious process. There are currently two main approaches to building a survey.

- The first is *additive*, we start with an initial set of papers, and manually find relevant related papers which we add. This time-consuming process will often overlook newer papers of which we are not aware and is quite likely to miss out on important contributions to the field.
- The second approach is *subtractive*, we start with a search of a literature database based on a set of keywords that define the area of interest. Very often, such a search returns a huge list of potential publications, we then have to refine the keywords and/or manually filter out unwanted publications to create our survey. Some publication types, like PhD theses, may be ignored completely if the database does not index them.

In contrast, we provide a tool that automates large parts of the first approach, while also using the contents of online databases to systematically find all papers that are either referred to by our current set of works or that are citing one of these works. We also use a domain-specific ontology to automatically estimate the relevance of any candidate publication, and reject non-relevant works.

1.1. Assumptions

The success of our approach depends on a number of key assumptions. The first assumption is that all significant papers in a research area will be connected to other papers in the area, forming a single connected component in the knowledge graph of all research papers. The second assumption is that all relevant works will have a DOI (digital object identifier, https://www.doi.org/), and will be indexed in at least one of the databases searched. The third core assumption is that we can define an area-specific ontology to find and classify all the works that we are looking for. While the tools automate much of the search and classification process, the final decision about which papers to include rests with the user, our tool is an interactive decision support tool.

submitted

2. Process

The tool can be used interactively, or as a batch process, the underlying process is shown in Figure 1.

Beginning with a LaTeX bibliography file that contains our seed papers, the system attempts to identify related literature by querying online services via REST requests using DOI keys. This finds all references and citing works of our working set. By extracting title, keywords, and abstracts from the meta-data we can determine the relevant concepts that apply to a work, and use that information to select which papers are relevant. These will be added to our working set, we continue this process until we reach a fix-point, and all remaining connected works are considered to be not relevant. Given a set of papers, we then manually download any full-text pdf files of the publications, and use text extraction with *pdfgrep* (https://pdfgrep.org/index.html) to perform a more comprehensive classification of the papers. We also use that information to produce various analysis reports.

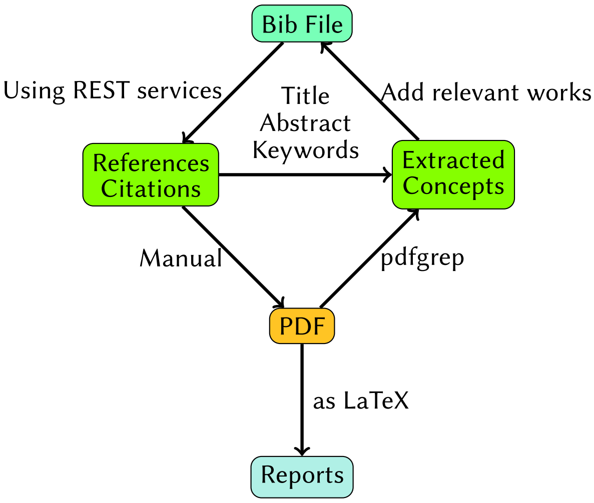


Figure 1: Process

Our main data feed is OpenCitations [2], which provides both references and citing works for a given DOI key. We also use the platforms Crossref (https://www.crossref.org/documentation/) and Scopus (https://www.elsevier.com/products/scopus) with their public REST APIs, which provide more detailed metadata for each work.

3. Case Studies

So far, we have used the tool to create four surveys in different computer science areas, a summary overview is given in Table 1. Each survey starts with a seed library, for example, we used a DBLP (https://dblp.org/) search for "Constraint Programming and Scheduling" as a starting point. We use a manually defined, survey-specific ontology to find relevant papers and to find sub-categories within the survey. Existing ontologies for Computer Science [3] are not fine-grained enough for the very specialized subject areas we are interested in. Note that the number of works included in the survey is only a small fraction of all works considered. That number far exceeds the number of publications we could check manually.

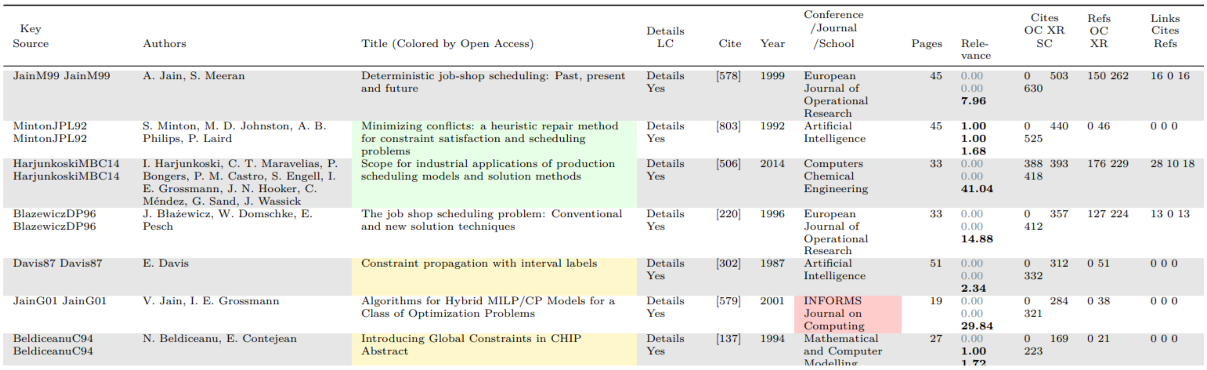
3.1. Survey Example

The current results of the "CP & Scheduling" survey can be found online at https://hsimonis.github.io/pthg24/, which also includes the sources of the application. The results are presented in different forms, based on type of publication, year of publication, the citation numbers for each work, and the number of internal connections inside the survey.

Table 1
Survey Case Studies

	Work	Seed	Ontology	Works	Works
Topic	With	Library	Terms	Found	Considered
CP & Scheduling	H. Simonis C. Öztürk	DBLP 223	407	1,263	28,313
Medical & Drones	G. Tacadao B. O'Sullivan L. Quesada H. Simonis	WoS 106	119	495	14,681
AI & Counter-Terrorism	H. Simonis B. O'Sullivan	Book 32	183	1,603	75,792
Uncertain & CP	J. Lopez H. Simonis	Manual 6	13	106	3,328

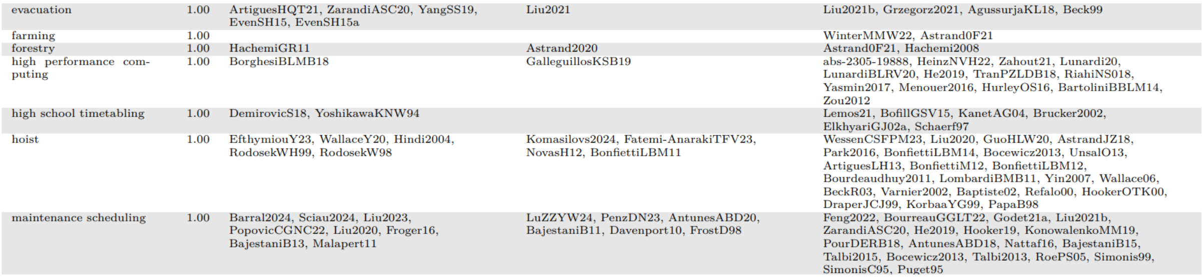
Figure 2 shows a list view of the most cited papers in the survey. Each line shows key properties of a work, with hyperlinks leading to more detailed information, or the paper itself. The citation numbers for each paper vary with the data feed, we use the maximum value of any feed for ranking. Note that the number of citations can be very different from the number of connections in the survey. The paper MintonJPL92[4] in the list for example has a large number of citations, but it is not referred to by other papers in the survey. This indicates that while this is an important paper for AI in general, its influence on CP Based Scheduling is quite limited. You can also see that the estimated relevance of that paper is quite low, just above the cut-off limit.



Key Source	Authors	Title (Colored by Open Access)	Details LC	Cite	Year	Conference /Journal /School	Pages	Rele- vance	Cites OC XR SC	Refs OC XR	Links Cites Refs
JainM99 JainM99	A. Jain, S. Meeran	Deterministic job-shop scheduling: Past, present and future	Details Yes	[578]	1999	European Journal of Operational Research	45	0.00 0.00 **7.96**	0 503 630	150 262	16 0 16
MintonJPL92 MintonJPL92	S. Minton, M. D. Johnston, A. B. Philips, P. Laird	Minimizing conflicts: a heuristic repair method for constraint satisfaction and scheduling problems	Details Yes	[803]	1992	Artificial Intelligence	45	**1.00** **1.00** **1.68**	0 440 525	0 46	0 0 0
HarjunkoskiMBC14 HarjunkoskiMBC14	I. Harjunkoski, C. T. Maravelias, P. Bongers, P. M. Castro, S. Engell, I. E. Grossmann, J. N. Hooker, C. Méndez, G. Sand, J. Wassick	Scope for industrial applications of production scheduling models and solution methods	Details Yes	[506]	2014	Computers Chemical Engineering	33	0.00 0.00 **41.04**	388 393 418	176 229	28 10 18
BlazewiczDP96 BlazewiczDP96	J. Błażewicz, W. Domschke, E. Pesch	The job shop scheduling problem: Conventional and new solution techniques	Details Yes	[220]	1996	European Journal of Operational Research	33	0.00 0.00 **14.88**	0 357 412	127 224	13 0 13
Davis87 Davis87	E. Davis	Constraint propagation with interval labels	Details Yes	[302]	1987	Artificial Intelligence	51	0.00 0.00 **2.34**	0 312 332	0 51	0 0 0
JainG01 JainG01	V. Jain, I. E. Grossmann	Algorithms for Hybrid MILP/CP Models for a Class of Optimization Problems	Details Yes	[579]	2001	INFORMS Journal on Computing	19	0.00 0.00 **29.84**	0 284 321	0 38	0 0 0
BeldiceanuC94 BeldiceanuC94	N. Beldiceanu, E. Contejean	Introducing Global Constraints in CHIP Abstract	Details Yes	[137]	1994	Mathematical and Computer Modelling	27	0.00 **1.00** **1.72**	0 169 223	0 21	0 0 0

Figure 2: Most Cited Works in Survey

Another type of analysis is based on the ontology, Figure 3 shows a few of the application area concepts that we have defined, together with papers that match each concept either a lot, somewhat or only in a limited way. We see for example that *farming* is not a popular concept in the survey, while *maintenance scheduling* is.



evacuation	1.00	ArtiguesHQT21, ZarandiASC20, YangSS19, EvenSH15, EvenSH15a	Liu2021	Liu2021b, Grzegorz2021, AgussurjaKL18, Beck99
farming	1.00			WinterMMW22, Astrand0F21
forestry	1.00	HachemiGR11	Astrand2020	Astrand0F21, Hachemi2008
high performance computing	1.00	BorghesiBLMB18	GalleguillosKSB19	abs-2305-19888, HeinzNVH22, Zahout21, Lunardi20, LunardiBLRV20, He2019, TranPZLDB18, RiahiNS018, Yasmin2017, Menouer2016, HurleyOS16, BartoliniBBLM14, Zou2012
high school timetabling	1.00	DemirovicS18, YoshikawaKNW94		Lemos21, BofillGSV15, KanetAG04, Brucker2002, ElkhyariGJ02a, Schaerf97
hoist	1.00	EfthymiouY23, WallaceY20, Hindi2004, RodosekWH99, RodosekW98	Komasilovs2024, Fatemi-AnarakiTFV23, NovasH12, BonfiettiLBM11	WessenCSFPM23, Liu2020, GuoHLW20, AstrandJZ18, Park2016, BonfiettiLBM14, Bocewicz2013, UnsalO13, ArtiguesLH13, BonfiettiM12, BonfiettiLBM12, Bourdeaudhuy2011, LombardiBMB11, Yin2007, Wallace06, BeckR03, Varnier2002, Baptiste02, Refalo00, HookerOTK00, DraperJCJ99, KorbaaYG99, PapaB98
maintenance scheduling	1.00	Barral2024, Sciau2024, Liu2023, PopovicCGNC22, Liu2020, Froger16, BajestaniB13, Malapert11	LuZZYW24, PenzDN23, AntunesABD20, BajestaniB11, Davenport10, FrostD98	Feng2022, BourreauGGLT22, Godet21a, Liu2021b, ZarandiASC20, He2019, Hooker19, KonowalenkoMM19, PourDERB18, AntunesABD18, Nattaf16, BajestaniB15, Talbi2015, Bocewicz2013, Talbi2013, RoePS05, Simonis99, SimonisC95, Puget95

Figure 3: Works Discussing Specific Concepts

The analysis also presents an summary of each work, with key extracted information, a justification for the considered relevance score, and a list of the most similar works based on the extracted ontology features.

3.2. Bibliometric Analysis

We now present some of the bibliometric analysis [5] results for the "CP & Scheduling" survey.

Figure 4 shows the institutions with the largest number of works included in the survey. The leads are the University of Toronto and the University of Bologna, but a significant number of other institutions were producing multiple works in this area. Note that this information is based on the affiliation meta-data provided in one of the data feeds, and considers the affiliation at the time of publication only.

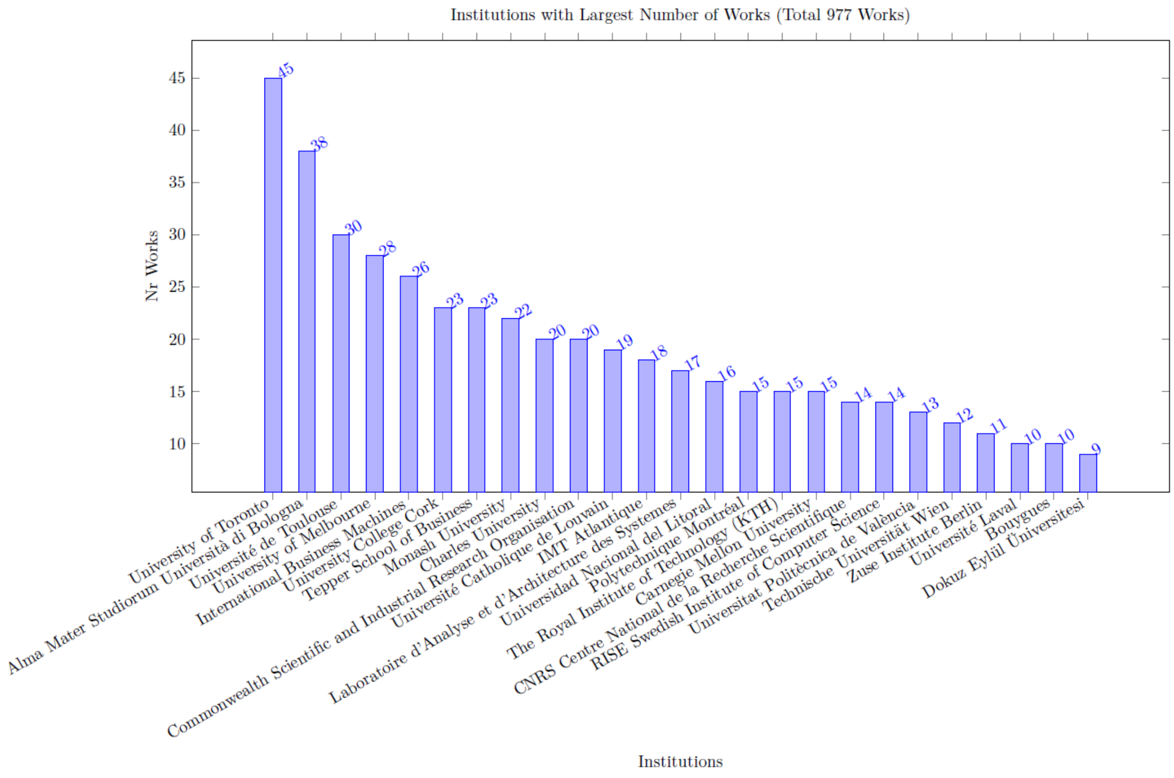


Figure 4: Publications by Institution

The next Figure (Figure 5) shows the number of publications by year, starting with early work in the 1980s. We distinguish conference papers, journal articles and PhD theses. The number of theses is probably still underrepresented, as they often miss DOI keys, and are therefore not found through database searches. It is interesting to see that since 2018 the number of journal articles outstrips the conference papers, whose numbers by year have been quite stable over time. While this may indicate a growing interest in CP based scheduling in the OR field, it can also be explained to a large part by the appearance of new actors in the scientific publishing field. Our full analysis report contains a more detailed discussion of this question.

Figure 6 shows a co-author graph produced from the survey data. Nodes represent authors, with the colors indicating the number of publications in the survey. Links connect authors who were coauthors for some publications, with color again indicating the number of collaborations. The graph shows that over time the CP scheduling community has worked together in many different combinations, with one large connected component of authors dominating the field, while there are only a few connected components of authors that are more isolated. This is not always the case, one of the other surveys shows that the co-author graph consists of many small connected components.

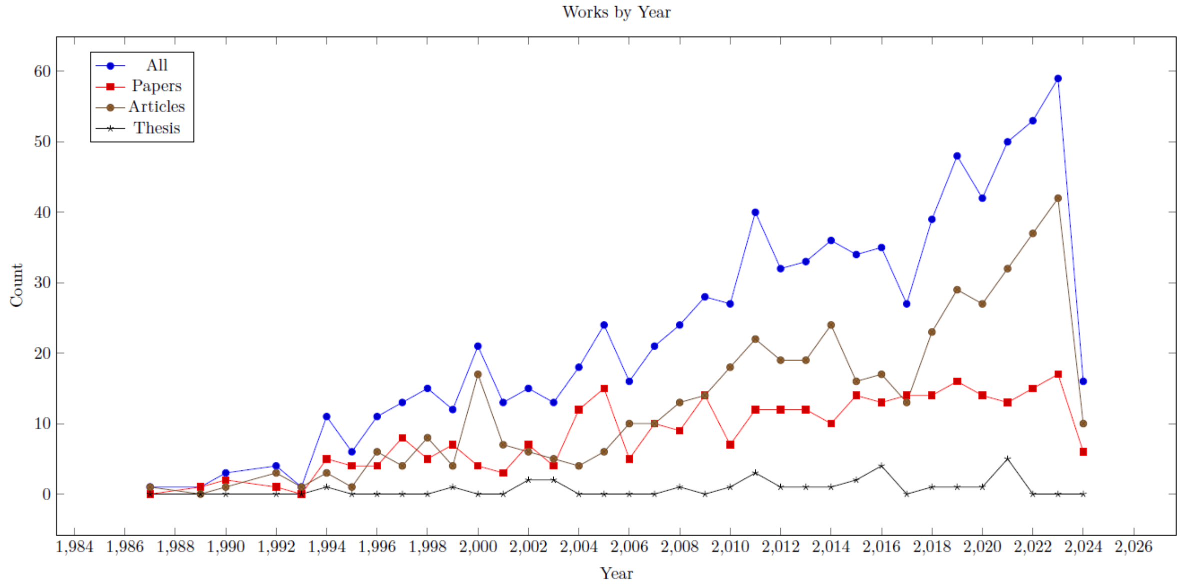


Figure 5: Publications by Year

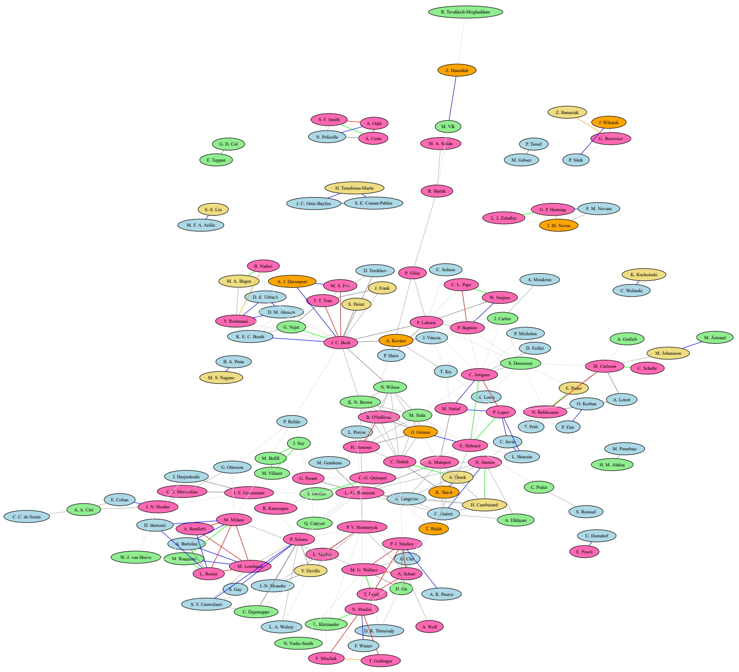


Figure 6: Coauthor Graph

Finally, Figure 7 shows the distribution of the cosine similarity measure based on the ontology features that we extract from the full text version of the works. We see that there are a few papers that are considered to be very similar to each other, these typically are papers by the same authors in different settings, for example a conference paper followed by an extended journal version of that paper. The similarity values are used in the survey to identify the closest works related to some specific paper, which might be of interest to readers of selected papers.

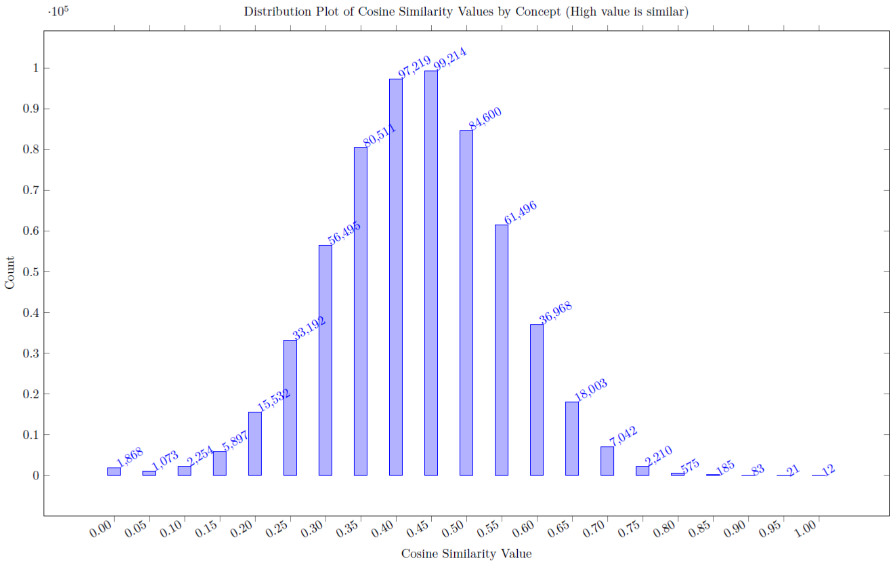


Figure 7: Similarity Distribution

4. Limitations and Future Work

While the tool already has shown its usefulness in generating surveys of specific Computer Science areas, we have identified several limitations that we may resolve in future versions.

The primary limitation arises from the fact that none of the database providers offer comprehensive data, which restricts the scope of the analyses we can conduct. For example, the abstract of a work is not always available in the meta-data, limiting our relevance analysis without access to the full text. As the database queries are using the DOI keys to identify works, they are not very helpful in analyzing works that do not have a DOI. This is a problem in particular for PhD theses, where universities have not assigned a DOI, and for older (before 2005) AI conferences, which were published by AI societies. For the "CP & Scheduling" survey we did perform a manual search of the proceedings by hand to find all relevant early publications on the topic. Finally, the use of one specific ontology in English for an area may result in ignoring works published in other languages that will use other technical terms for the same concepts. The use of a LaTeX tool-chain and non-uniform encoding of characters in web searches causes issues for non-Latin texts. We did find that while retrieving the full-text versions remains a manual process, this was not a limiting factor, as we could present the required links and keys to the user in a way that allowed very rapid access (less than a minute per paper).

There are of course further analysis that can be performed on the collected papers, for example visualizing the connection between all papers for a specific sub-area, to see when novel concepts were first introduced. This could help with providing further analysis of the publications in an area for a deeper understanding of a sub-field.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 12/RC/2289-P2 at Insight the SFI Research Centre for Data

References

[1] M. Pautasso, Ten simple rules for writing a literature review, PLoS Comput Biol 9 (2013). doi:https://doi.org/10.1371/journal.pcbi.1003149.

[2] S. Peroni, D. Shotton, OpenCitations, an infrastructure organization for open scholarship, Quantitative Science Studies 1 (2020) 428–444. URL: https://doi.org/10.1162/qss_a_00023. doi:10.1162/qss_a_00023. arXiv:https://direct.mit.edu/qss/article-pdf/1/1/428/1760920/qss_a_00023.pdf.

[3] M. Beck, S. T. R. Rizvi, A. Dengel, S. Ahmed, From automatic keyword detection to ontology-based topic modeling, in: X. Bai, D. Karatzas, D. Lopresti (Eds.), Document Analysis Systems, Springer International Publishing, Cham, 2020, pp. 451–465.

[4] S. Minton, M. D. Johnston, A. B. Philips, P. Laird, Minimizing conflicts: a heuristic repair method for constraint satisfaction and scheduling problems, Artificial Intelligence 58 (1992) 161–205. URL: http://dx.doi.org/10.1016/0004-3702(92)90007-k. doi:10.1016/0004-3702(92)90007-k.

[5] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, W. M. Lim, How to conduct a bibliometric analysis: An overview and guidelines, Journal of Business Research 133 (2021) 285–296. URL: https://www.sciencedirect.com/science/article/pii/S0148296321003155. doi:https://doi.org/10.1016/j.jbusres.2021.04.070.