





Who Is Who in Literature-Based Discovery: Preliminary Analysis

Andrej Kastrin^(✉)  and Dimitar Hristovski 

Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
{andrej.kastrin,dimitar.hristovski}@mf.uni-lj.si

Abstract. Literature-based discovery (LBD) has undergone an evolution from being an emerging area to a mature research field. Hence it is necessary to summarize the literature and scrutinize general bibliographic characteristics and publication trends. This paper presents very basic scientometric review of LBD in the period 1986–2020. We identified a total of 237 publications on LBD in the Web of Science database. The Journal of Biomedical Informatics published the greatest amount of papers on LBD. The United States plays a leading role in LBD research. Thomas C. Rindflesch is the most productive co-author in the field of LBD. Drawing on these first insights, we aim to better understand the historical progress of LBD in the last 35 years and to be able to improve the publishing practices to contribute to the field in the future.

Keywords: Literature-based discovery · Bibliometric study · Research performance

1 Introduction

Literature-based discovery (LBD) is a text mining approach for automatically generating research hypotheses [11]. LBD is a complex, continually evolving and collaborative research field. To the best of our knowledge, five traditional literature reviews were recently written to elucidate the extent of current knowledge in the LBD research domain. In the same year, Sebastian et al. [10] and Henry et al. [7], published extensive review papers on LBD. The first group of authors provides an in-depth discussion on a broad palette of existing LBD approaches and offers performance evaluations on some recent emerging LBD methodologies. Later authors likewise introduced historical and modern LBD approaches and provided an overview of evaluation methodologies and current trends. Both papers provided the unifying framework for the LBD paradigm, its methodologies, and tools. In 2019 three new review papers appear. Gopalakrishnan et al. [6] provide a more comprehensive analysis of the LBD field; their paper may serves

Supported by the Slovenian Research Agency (Grant No. Z5-9352 and J5-1780).

© Springer Nature Switzerland AG 2020
W. Lu and K. Q. Zhu (Eds.): PAKDD 2020 Workshops, LNAI 12237, pp. 51–59, 2020.
https://doi.org/10.1007/978-3-030-60470-7_6

as a methodological introduction behind particular tools and techniques. Thilakaratne et al. [13] analyzed methodologies used in the LBD using a novel classification scheme and provide a timeline with key milestones in LBD research. In their second paper, Thilakaratne et al. [14] present a large-scale systematic review of the LBD workflow by manually analysing 176 LBD papers. Although these reviews successfully provide insight into the field of LBD through dissecting the research evidence and appropriate classification of research themes, they have not used more sophisticated tools, such as bibliometric and scientometric analysis. Recently, Chen et al. [4] performed first scientometric analysis in the LBD field. They use LBD domain as a proxy to illustrate an intuitive method to compare multiple search strategies in order to identify the most representative body of scientific publications. Consequently, the in-depth analysis in the LBD field is urgently needed, to provide newcomers, researchers, and clinicians with a state-of-the-art scientometric overview of the area.

2 Methods

2.1 Collection of Bibliographic Data

We used Web of Science (WoS) (Clarivate Analytics, Philadelphia, PA, USA) as the data sources for retrieving publications and related metadata in the LBD research domain. In this preliminary analysis, our objective was to include as complete set of publications on LBD as possible without much manual intervention. The search strategy for WoS was defined as: `TS=(((‘literature-based discovery’)) OR (‘undiscovered public knowledge’))`. The time span was from 1986 until 2020. We applied no language, geographic, or any other constraints on the database retrieval procedure. We are aware of at least two limitations of the simple search strategy described above. One limitation is that many conference papers are not indexed in WoS, and therefore, were not included in our analysis. However, we do know there is a considerable number of important LBD papers published in conferences. For example, our group has written at least four well-cited conference papers that are currently not included. The other limitation is that in quite a few cases, the authors have been creative with the titles and abstracts of their LBD papers, and had avoided mentioning the well-established phrases such as *literature-based discovery*. We will address both limitations in our future work by developing a more complex search strategy (or a set of strategies), and by doing various manual interventions.

2.2 Data Analysis

We prepare and summarise statistics on most prolific authors, countries, and journals. We obtain journal metrics including impact factors of the top 10 journals from Journal Citation Reports (Clarivate Analytics, Philadelphia, PA, USA) on January 30, 2020. The main part of the analysis and visualizations were performed in R using the `bibliometrix` package [2].

3 Results

A last search of the databases was performed on January 30, 2020. In further analysis we included a total of 237 bibliographic records. Publications cover a time period of 35 years (1986–2020) beginning with Swanson’s first paper on the LBD [12].

The majority of records were original articles ($n = 139$), followed by conference papers ($n = 58$), review papers ($n = 17$), book chapters ($n = 8$), and other material. As of January 30, 2020, the complete set of publications had been cited $n = 5400$ times.

3.1 Publication Evolution over the Years

In the time period 1986–2020, $n = 237$ publications were published about LBD and indexed in WoS. The maximum number of papers ($n = 22$) was published in 2017. It is noteworthy that the term *Literature-Based Discovery* was included in Medical Subject Headings (MeSH) vocabulary in 2013, indicating its high bibliographic importance. Figure 1 depicted the changing pattern of publications (actual and cumulative frequencies) in our data set from 1986 until 2020 for WoS. The reader can observe that the number of publications on LBD increased slowly from 1986 to 2000, but since then it has been increasing significantly. This fact indicates that the field of LBD has acquired significant attention in the last two decades.

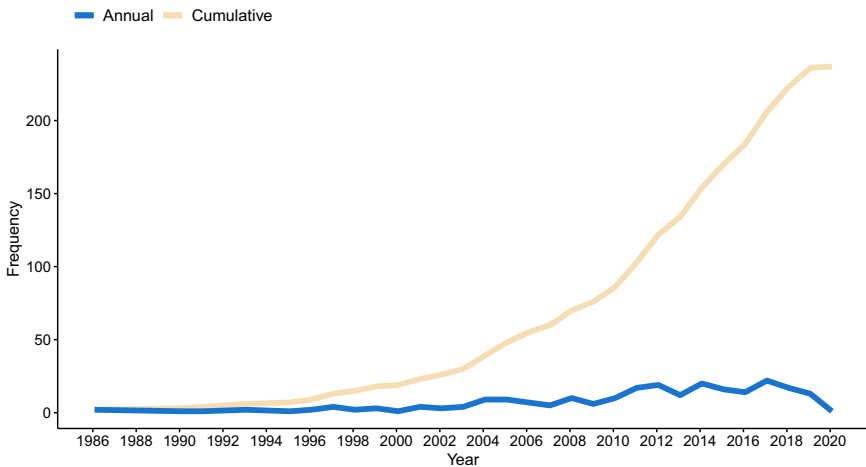


Fig. 1. Number of LBD publications in WoS collection during the period 1986–2020

3.2 Most Productive Authors

Our analysis identifies 530 distinct authors. The majority of the authors write in collaboration with colleagues ($n = 497$). On average we detected 2.24 authors per document and 0.45 documents per author. The authors with the highest number of publications and citations have a tendency to be scientists who drive the research field and have the casting vote for its development. The 10 top authors with the most publications are presented in Table 1. Thomas C. Rindfleisch clearly holds the first position with 20 publications, although he is the first author in only one paper on LBD. In addition to raw number of publication, we also present fractionalized number of publications. The author fractionalized number of publications (fNP) is the sum of a unit's publications after assigning each publication the value 1 and dividing the assigned value with the number of authors. Low values in relation to the number of publications indicate a high level of co-authors. For instance, Kostoff achieves high fNP value, because he authored the greatest number of solo publications.

In Fig. 2 we present a co-authorship network of authors in the LBD domain as derived from the WoS database. Although our group has been collaborating in LBD with Thomas C. Rindfleisch since the early 2000s, it came to us as a surprise that he is the author with most LBD publications. He is mostly known for the development of SemRep, a natural language processing system that extracts semantic predication from biomedical text [9]. However, Fig. 2 illustrates well that over the years he has collaborated with several other groups, and in the last decade he had his own group publishing in LBD. In network analysis term, he has the highest betweenness centrality and he is the major hub in the co-author network. In the current analysis, we count authorship regardless of the author's position. But most of the time, in most publications, the first author is the one doing most of the work and usually being responsible for the major novelty of the publication. Therefore, as further work, we will create an additional table with the first authors only.

3.3 Most Productive Countries

A total of 34 countries contributed to the selected data set of LBD literature. First, it is worth noting that the LBD production is unevenly distributed across countries. United States commit exactly half of the body of the literature to the treasury of knowledge on LBD ($n = 117, 50\%$). This indicates that the US is leading in LBD research. Interestingly, Slovenia, a small country in the heart of Europe, is the second-most productive country with 18 publications (7.6%). Surprisingly, India has no researcher who published about LBD as the first author (Table 2).

3.4 Most Relevant Journals

When analyzing research productivity, it is essential to study the journals in which papers are published. LBD is so specific research field that it has no

Table 1. Top 10 authors based on the total number of publications

Rank	Author	NP	Author	fNP
1	Rindflesch TC	20	Kostoff RN	10.28
2	Kostoff RN	16	Smalheiser NR	7.20
3	Hristovski D	13	Swanson DR	6.37
4	Smalheiser NR	12	Rindflesch TC	4.82
5	Song M	12	Hristovski D	3.62
6	Swanson DR	10	Song M	3.29
7	Kastrin A	9	Kastrin A	2.67
8	Cohen T	7	Preiss J	2.33
9	Persidis A	7	Cohen T	2.20
10	Lee D	6	Ahmed A	2.00

Note: NP = Number of Publications, fNP = fractionalized Number of Publications

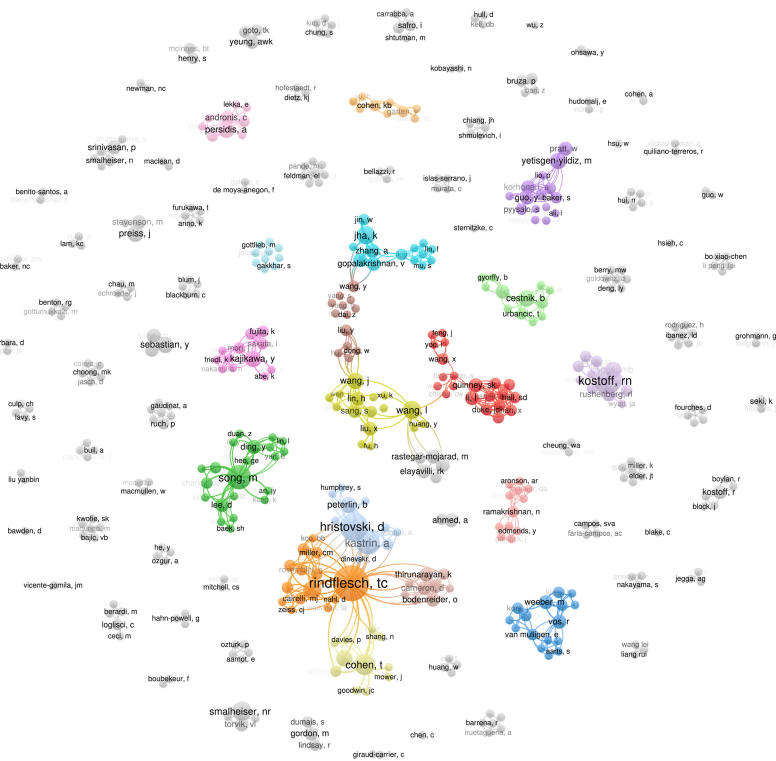


Fig. 2. Co-authorship network of authors on LBD themes

Table 2. Top 10 productive countries for LBD research

Rank	Country	NP	Prop	SCP	MCP	MCP _r
1	USA	117	0.50	98	19	0.162
2	Slovenia	18	0.08	7	11	0.611
3	China	17	0.07	15	2	0.118
4	Korea	13	0.06	8	5	0.385
5	United Kingdom	12	0.05	8	4	0.333
6	Japan	11	0.05	9	2	0.182
7	Australia	5	0.02	3	2	0.400
8	Netherlands	5	0.02	3	2	0.400
9	Spain	5	0.02	5	0	0.000
10	Canada	4	0.02	3	1	0.250

Note: NP = Number of Publications, Prop = Proportion of Publications, SCP = Single Country Publications, MCP = Multiple Country Publications

specialized journals. Instead, the LBD research is published mainly in journals related to (biomedical) informatics and bioinformatics. Table 3 summarizes the details about the top 10 journals. Not surprisingly, with respect to the number of publications, *Journal of Biomedical Informatics* had published 14 papers on LBD research, followed by *Technological Forecasting and Social Change* with 11 published articles. *Briefings in Bioinformatics*, which has the highest impact factor in our list, has published only 6 papers on LBD. Out of the 10 journals, the majority are published in the United States. Journals from the top 10 list publish LBD papers from the beginning of the 2000s, with the exception of the *Journal of the American Society for Information Science* which was active between the years 1987 and 1999.

Table 3. Journals with most LBD publications

Rank	Source	NP	IF
1	J. Biomed. Inform.	14	2.950
2	Technol. Forecast. Soc. Chang.	11	3.815
3	J. Am. Soc. Inf. Sci. Technol.	10	2.452
4	BMC Bioinformatics	7	2.511
5	Bioinformatics	6	4.531
6	Brief. Bioinform.	6	9.101
7	J. Am. Med. Inf. Assoc.	6	4.292
8	PLoS One	6	2.776
9	Scientometrics	6	2.770
10	J. Doc.	5	1.573

Note: NP = Number of Publications, IF = Impact Factor

3.5 Publication Hallmarks

By employing the processed bibliometric data, we can identify the most important hallmarks of LBD research. The top 10 most cited papers are listed in Table 4, including their first author, year of publication, journal, the total number of citations and number of citations per year. Data are ranked by the number of citations. Swanson is the author of five listed publications including his seminal paper on fish oil and Raynaud’s disease which is the first on the list [12]. The second most cited paper is a review article published by Cohen et al. [5] in which they discuss various text mining approaches including automated hypothesis generation from literature. Swanson’s paper is categorically the first hallmark of LBD research. However, it is important to note that Cohen’s paper has more than two-times more citations per year. This is probably due to the high impact factor of the journal in which the paper was published and because of its interestingness for the broader domain of researchers. These ten publications covered the theoretical research as well as practical applications of LBD. However, all these papers were published before 2005, although important scientific achievements in LBD were published also later on.

Table 4. The top 10 papers in the LBD domain based on the number of citations

Rank	Paper	TC	TCY
1	Swanson DR, 1986, <i>Perspect Biol Med</i>	402	11.82
2	Cohen AM, 2005, <i>Brief Bioinform</i>	363	24.20
3	Dumais ST, 2004, <i>Annu Rev Inform Sci</i>	290	18.12
4	Swanson DR, 1997, <i>Artif Intell</i>	230	10.00
5	Swanson DR, 1986, <i>Libr Quart</i>	169	4.97
6	Srinivasan P, 2004, <i>J Am Soc Inf Sci Tec</i>	155	9.69
7	Kostoff RN, 2004, <i>Technol Forecast Soc</i>	150	9.38
8	Hristovski D, 2005, <i>Int J Med Inform</i>	140	9.33
9	Zweigenbaum P, 2007, <i>Brief Bioinform</i>	131	10.08
10	Weeber M, 2001, <i>J Am Soc Inf Sci Tec</i>	126	6.63

Note: TC = Total Citations, TCY = Total Citations per Year

4 Discussion

Through very basic scientometric analysis, this study aimed to reveal worldwide scientific productivity and research trends in LBD over the last three decades (1986–2020). To the best of our knowledge, this paper, although in its preliminary version, is the first scientometric analysis in the field of LBD.

Understanding the past and current body of publications is the sine qua non for the advancement of LBD research in the future. In the last decade, a

plethora of studies has been published examining knowledge structure and evolution through the bibliographic lens of particular scientific fields. The lack of a similar study in the LBD area makes it difficult if not impossible to compare LBD with other research fields. However, LBD is inherently lean to biomedicine and to medical informatics in particular. There are two reasons for this fact. First, historically, LBD originates from the medical applications. Second, practically, MEDLINE distribution is freely available to researchers that is not the case with Scopus or WoS.

A conspicuous change in the number of papers published per year suggests that a major turning point is occurring in the field. We found that the number of publications increased over the last 20 years, particularly since 2000. The development of the LBD field is associated with great progress in computer science and natural language processing in particular. The total number of citations accumulate over the years and consequently, the recent papers do not have enough time to acquire more citations. However, the growth of publications and citations in recent years indicates a promising future of LBD.

Scientific productivity is strongly correlated with international collaboration among researchers, countries, and institutions [8]. Studies investigating the scientific impact of cross-institution groups confirmed that their papers have a higher citation rate in comparison to papers produced by a single research group. Papers with international co-authorship have an even higher impact [15]. Most of the research produced in the field of LBD is generated in the cliques of researchers. Even though the collaboration and internationalization among researchers have certain downsides, it provides great benefits. Abramo et al. [1] demonstrated an increasing trend in collaboration among institutions that could be attributed to different policies stimulating research collaboration (e.g., the EU Framework Programme for Research and Innovation). We are aware of at least one successful EU FP7 funded project from the domain of LBD named BISON (2008–2011) that investigates novel methods for discovering new, domain bridging connections and patterns from heterogeneous data sources [3].

For further work, we intend to greatly expand the analysis to the Scopus, Pubmed, Google Scholar, and Dimensions databases. To build an universum of relevant publications, we will employ a strategy that combines regular query search with cascading citation expansion approach as proposed recently by Chen et al. [4]. Preliminary work reveals that such expansion improves the results a lot. Next, we already work on the science mapping of the LBD domain. When completed, this study will elucidate a past, present, and future image of LBD in great detail.

References

1. Abramo, G., D'Angelo, C.A., Solazzi, M.: The relationship between scientists' research performance and the degree of internationalization of their research. *Scientometrics* **86**(3), 629–643 (2011)
2. Aria, M., Cuccurullo, C.: Bibliometrix: an R-tool for comprehensive science mapping analysis. *J. Inf.* **11**(4), 959–975 (2017)

3. Berthold, M.R. (ed.): *Bisociative Knowledge Discovery*. LNCS (LNAI), vol. 7250. Springer, Heidelberg (2012). <https://doi.org/10.1007/978-3-642-31830-6>
4. Chen, C., Song, M.: Visualizing a field of research: a methodology of systematic scientometric reviews. *PLoS One* **14**(10), e0223994 (2019)
5. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. *Brief. Bioinform.* **6**(1), 57–71 (2005)
6. Gopalakrishnan, V., Jha, K., Jin, W., Zhang, A.: A survey on literature based discovery approaches in biomedical domain. *J. Biomed. Inform.* **93**, 103141 (2019)
7. Henry, S., McInnes, B.: Literature based discovery: models, methods, and trends. *J. Biomed. Inform.* **74**, 20–32 (2017)
8. Lee, S., Bozeman, B.: The impact of research collaboration on scientific productivity. *Soc. Stud. Sci.* **35**(5), 673–702 (2005)
9. Rindfleisch, T.C., Fiszman, M.: The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J. Biomed. Inform.* **36**(6), 462–477 (2003)
10. Sebastian, Y., Siew, E.G., Orimaye, S.: Emerging approaches in literature-based discovery: techniques and performance review. *Knowled. Eng. Rev.* **32**, E12 (2017)
11. Smalheiser, N.: Rediscovering Don Swanson: the past, present and future of literature-based discovery. *J. Data Inf. Sci.* **2**(4), 43–64 (2017)
12. Swanson, D.: Fish oil, Raynaud’s syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **30**(1), 7–18 (1986)
13. Thilakaratne, M., Falkner, K., Atapattu, T.: A systematic review on literature-based discovery: general overview, methodology, & statistical analysis. *ACM Comput. Surv. (CSUR)* **52**(6), 1–34 (2019)
14. Thilakaratne, M., Falkner, K., Atapattu, T.: A systematic review on literature-based discovery workflow. *PeerJ Comput. Sci.* **5**, e235 (2019)
15. Thonon, F., et al.: Measuring the outcome of biomedical research: a systematic literature review. *PLoS One* **10**(4), e0122239 (2015)