Discovering seminal works with marker papers

Robin Haunschild and Werner Marx

Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart, Germany
{r.haunschild@fkf.mpg.de, w.marx@fkf.mpg.de}

Abstract. Bibliometric information retrieval in databases can employ different strategies. Commonly, queries are performed by searching in title, abstract and/or author keywords (author vocabulary). More advanced queries employ database keywords to search in a controlled vocabulary. Queries based on search terms can be augmented with their citing papers if a research field cannot be curtailed by the search query alone. Here, we present another strategy to discover the most important papers of a research field. A marker paper is used to reveal the most important works for the relevant community. All papers co-cited with the marker paper are analyzed using Reference Publication Year Spectroscopy (RPYS). For demonstration of the marker paper approach, density functional theory (DFT) is used as a research field. Comparisons between a prior RPYS on a publication set compiled using a keyword-based search in a controlled vocabulary and a co-citation RPYS (RPYS-CO) show very similar results. Similarities and differences are discussed.

Keywords: Bibliometrics, RPYS, RPYS-CO, marker paper, seminal papers, historical roots, DFT

1 Introduction

Information retrieval in databases can be performed using different routes. Commonly, searches are performed via search terms (author vocabulary) in the full-text or in certain sections of a paper (e. g., title, abstract, and/or author keywords). Some databases also offer controlled vocabulary (i. e., keywords assigned by the database producer) to be searched. Searches in author vocabulary often require a strategy which is called "interactive query formulation" and was extensively discussed by Wacholder [1]. This strategy was applied for example in Haunschild, Bornmann and Marx [2] and Wang, Pan, Ke, Wang and Wei [3] to analyze the literature about climate change. A search in controlled vocabulary often needs less search terms and less complicated queries. For example, Haunschild, Barth and Marx [4] used a rather concise search query in the controlled vocabulary of CAplus^{SM} to analyze the literature about density functional theory (DFT), a widely used method in the field of computational chemistry.

Besides keyword searches, the citing papers of one specific key-paper (or a few key papers) can be used to retrieve fundamental literature, see e. g., Marx, Haunschild

and Bornmann [5]. This enables bibliometricians to cover publication sets which are hard to narrow down using keyword searches only.

Here, we apply a methodology using a single marker paper (or a few marker papers) for retrieving the set of most influential publications of a topic. This methodology (RPYS-CO) is based on the co-citation network of publications [6]. We will compare the results from our RPYS-CO analysis with the previous RPYS analysis by Haunschild, Barth and Marx [4] which is based on a keyword search in a controlled vocabulary. Previously, the methodology has been applied to the history of the greenhouse effect [7]. The references within the citing papers of the marker paper are used in a RPYS (Reference Publication Year Spectroscopy) analysis. The publication set to be analyzed contains all papers which have been co-cited with the marker paper. In case of a few marker papers, the papers of the publication set are co-cited with at least one of the marker papers.

RPYS is a bibliometric method for locating seminal papers and the historical roots in publication sets covering specific research topics or fields [8]. The method analyzes the cited references of the papers of the relevant publication set. The references most frequently cited are analyzed in graphical and tabular forms. This provides a more objective answer to the question about seminal papers and historical roots (based on the "wisdom of the crowd"). Individual scientists in the field can answer this question only subjectively. However, many scientists with knowledge in the studied field deliver a broader view which is the basis for the interpretation of the RPYS results.

2 Methods

2.1 Dataset used

This analysis is based on the Web of Science (WoS, Clarivate Analytics) custom data of our in-house database derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), and Arts and Humanities Citation Index (AHCI) produced by Clarivate Analytics (Philadelphia, USA). Our in-house database contains the WoS publications since the publication year 1980.

A good marker paper should be of high relevance of the field under study. As a marker paper, we selected the publication by Becke [9] in which he proposed a very popular density functional approximation for the exchange energy which was for example used together with the LYP correlation functional [10] and in the very popular B3LYP functional [11]. Therefore, Becke [9] (also known as "Becke88") seems to be a very promising candidate for a marker paper. We exported all papers (n= 34,437) from our in-house database which cited this marker paper.

2.2 Software

We used the CRExplorer (see: http://crexplorer.net) to perform the RPYS analysis. The program can be downloaded for free and a comprehensive handbook explaining all functions is also available. With the program meta-knowledge [12] and the web tool RPYS i/o [13] two other resources have been developed in recent years for doing

cited references analyses, too. However, CRExplorer has a much broader functionality than both other resources.

2.3 Methodology

We used the CRExplorer script language to process the 668,007 unique reference variants (n=1,992,244 cited references, CRs). The script in **Listing 1** was used to perform the RPYS analysis. The command `importFile` is used to import all WoS papers citing Becke [9] which were published between 1980 and 2017. The range of reference publication years (RPYs) is restricted to 1950-1990 in order to analyze the same time frame as reported in Haunschild, Barth and Marx [4]. Clustering and merging equivalent CR variants is done via the commands `cluster` and `merge`. All CRs which were referenced less than 100 times are removed via the `removeCR` command. The value of 100 should be adjusted to the size of the studied data set in terms of cited references. Finally, the command exportFile is used to write the results (CR file and spectrogram file) in CSV format to files. The R package BibPlots (see: https://cran.r-project.org/web/packages/BibPlots/index.html and https://tinyurl.com/y97bb54z) is used to plot the spectrograms.

```
importFile(file: "citing_papers.wos.txt", type: "WOS",
RPY: [1950, 1990, false], PY: [1980, 2017, false], maxCR:
0)
cluster(threshold: 0.75, volume: true, page: true, DOI:
false)
merge()
removeCR( N_CR: [0, 99])
exportFile(file: "full_rpys_CR.csv", type: "CSV_CR")
exportFile(file: "full_rpys_GRAPH.csv", type:
"CSV_GRAPH")
```

Listing 1: CRExplorer script to perform RPYS on the WoS papers citing Becke [9]

3 Results

3.1 RPYS-CO with a suitable marker paper

A suitable marker paper should fulfill at least two requirements: (i) it should be cited fairly well considering the topic under study, and (ii) it should reasonably represent the studied topic. The paper by Becke [9] is highly cited. Furthermore, Becke [9] presents a very popular functional approximation for the exchange energy. Every researcher using this approximation should cite this paper. Therefore, this paper presents a very good candidate for a marker paper. Other very good candidates would be, e. g., Hohenberg and Kohn [14], Kohn and Sham [15], Lee, Yang and Parr [10], Perdew [16], Perdew, Burke and Ernzerhof [17], and Perdew, Ernzerhof and Burke

[18]. The proper choice of suitable marker papers requires at least some knowledge of the topic under study.

Fig. 1 shows the number of cited reference (NCR) curves for the RPYS-CO in this study and the RPYS from Haunschild, Barth and Marx [4] for the time frame 1950-1990. The NCR curves show differences and similarities. The peaks are positioned in or around the same RPYs (1951, 1955, 1964/65, 1970, 1972/73/74, 1976/77, 1980, 1985/86, and 1988) but the peak heights differ. The peak papers from the RPYS analysis were discussed in Haunschild, Barth and Marx [4]. **Fig. 2** shows the spectrogram of the RPYS-CO analysis using Becke [9] as a marker paper. The peak papers of the RPYS-CO analysis are listed in **Table 1**.

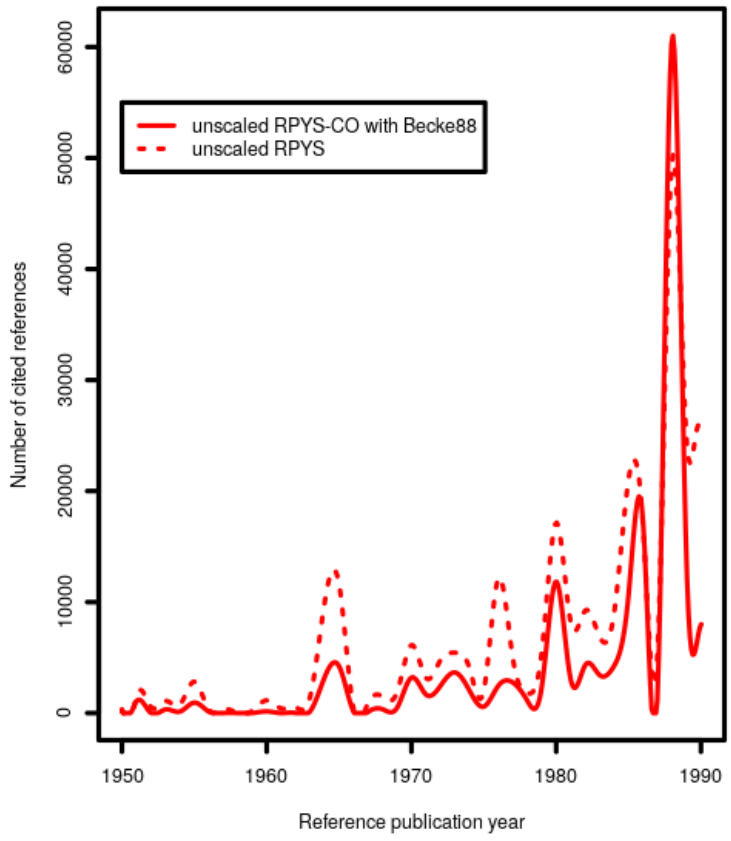

Fig. 1. Comparison of NCR curves from the RPYS analysis using DFT papers from a keyword search in controlled vocabulary of the CAS thesaurus for the time frame 1950-1990 from Haunschild, Barth and Marx [4] with the RPYS-CO analysis in this study using Becke [9] as a marker paper

The CRs 11, 12, 13, 15, and 16 appear in the RPYS-CO but were not mentioned in the RPYS analysis of Haunschild, Barth and Marx [4]. These five CRs of course occurred in the RPYS analysis, too, but did not seem to be as significant as in the RPYS-CO analysis performed in this study. The other 14 CRs of the RPYS-CO also appeared in the RPYS of Haunschild, Barth and Marx [4]. Some CRs even have very similar NCR values, e. g., CR1 with NCR = 793 in the RPYS-CO and NCR = 737 in the RPYS of Haunschild, Barth and Marx [4]. The largest absolute deviation between the results of RPYS and RPYS-CO are found for the marker paper CR18 with NCR = 33,850 in the RPYS-CO and NCR = 14,150 in the RPYS. The peak in the RPY 1976/77 in this RPYS-CO is broader than in the RPYS of Haunschild, Barth and Marx [4]. The different focus can be seen by the comparison of the NCR values of CR10: NCR = 407 in RPYS-CO and NCR = 6506 in RPYS. Monkhorst and Pack proposed a new method to generate special points in the Brillouin zone which enables more efficient integrations of periodic functions. This method had much more impact in the overall DFT community than in the publication set of our RPYS-CO.

In CR11, Ziegler and Rauk proposed a methodology for calculating bonding energies and bond distances using the Hartree-Fock-Slater method. Optimized basis sets for 3*d* orbitals were presented by Hay in CR12. Hirshfeld proposed a molecular partial charge analysis in CR 13. Hay presented very frequently used ab-initio effective core potentials for molecular calculations in CRs 15 and 16. These CRs had more impact in the publication set of our RPYS-CO than in the RPYS analysis based on keywords as presented by Haunschild, Barth and Marx [4].

In fact, we captured the most important seminal papers in **Table 1** as we can see from ordering the CRs by the NCR value. All 10 most frequently occurring CRs appear in **Table 1** except two of them (Dunning [19] with NCR = 2658 and Parr and Yang [20] with NCR = 2263). Dunning [19] proposed very popular atom-centered basis sets. Parr and Yang [20] is a very popular textbook about DFT. Both CRs were published in 1989. We see that 1989 is on the lower end of the downward slope of the 1988 peak. It is a matter of choice of the scope of the analysis if such RPYs should also be investigated. However, inspection of the most frequently occurring CRs is always recommended. The scope of our study is on the RPYS-CO method rather than on the seminal papers of DFT itself. Studies which have a specific topic as a focus, should investigate the RPYS results more deeply than performed here. For example, the CRExplorer also offers advanced indicators to discover papers with significant impact over many citing years [21].

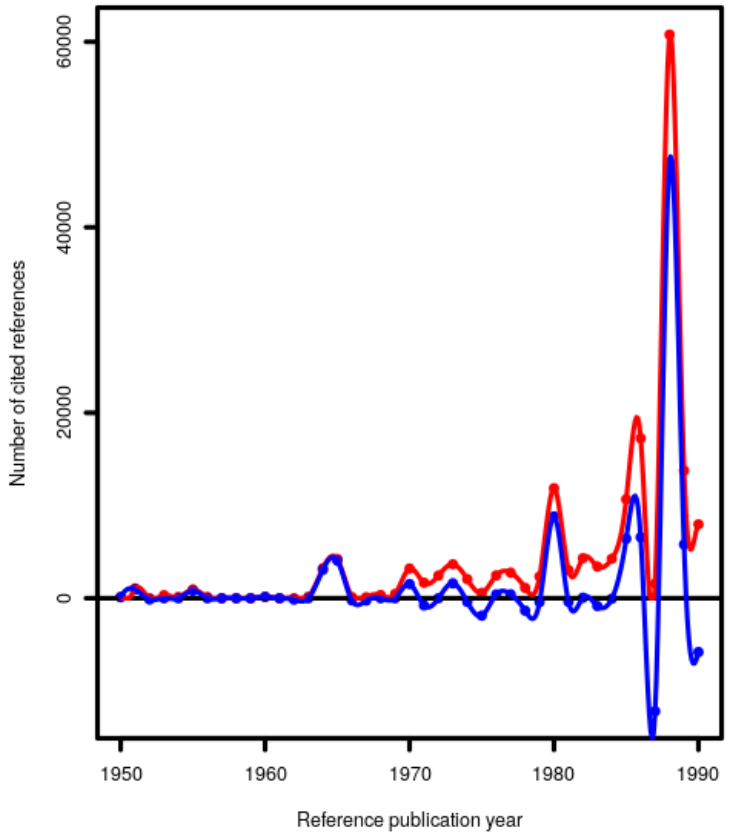

Fig. 2. RPYS-CO analysis using papers co-cited with Becke [9] for the time frame 1950-1990. The red curve and dots show the NCR values. The blue curve and dots show the five-year median deviation. Both curves are used to locate peaks.

Table 1. Peak papers of the RPYS-CO using papers co-cited with Becke [9] for the time frame 1950-1990

No	RPY	CR	NCR
CR1	1951	Slater JC, 1951, Physical Review, V81, P385	793
CR2	1951	Roothaan CCJ, 1951, Reviews of Modern Physics, V23, P69	267
CR3	1955	Mulliken RS, 1955, Journal of Chemical Physics, V23, P1833	642
CR4	1964	Hohenberg P, 1964, Physical Review B, V136, Pb864	2,713
CR5	1965	Kohn W, 1965, Physical Review, V140, P1133	3,688
CR6	1970	Boys SF, 1970, Molecular Physics, V19, P553	1,584

CR7	1972	Hehre WJ, 1972, Journal of Chemical Physics, V56, P2257	1,815
CR8	1973	Harihara PC, 1973, Theoretica Chimica Acta, V28, P213	1,957
CR9	1973	Baerends EJ, 1973, Chemical Physics, V2, P41	1,446
CR10	1976	Monkhorst HJ, 1976, Physical Review B, V13, P5188	407
CR11	1977	Ziegler T, 1977, Theoretica Chimica Acta, V46, P1	645
CR12	1977	Hay PJ, 1977, Journal of Chemical Physics, V66, P4377	428
CR13	1977	Hirshfeld FL, 1977, Theoretica Chimica Acta, V44, P129	398
CR14	1980	Vosko SH, 1980, Canadian Journal of Physics, V58, P1200	6,962
CR15	1985	Hay PJ, 1985, Journal of Chemical Physics, V82, P299	2,340
CR16	1985	Hay PJ, 1985, Journal of Chemical Physics, V82, P270	1,710
CR17	1986	Perdew JP, 1986, Physical Review B, V33, P8822	10,308
CR18	1988	Becke AD, 1988, Physical Review A, V38, P3098	33,850
CR19	1988	Lee CT, 1988, Physical Review B, V37, P785	21,887

3.2 RPYS-CO without a suitable marker paper

In order to choose a suitable marker paper, one needs at least some insight into the topic under study. Furthermore, a preliminary query using search terms is helpful for determining the usual citation rate of the topic. In this section, we demonstrate, by applying the RPYS-CO methodology iteratively, the procedure starting with a rather poor marker paper. We choose to start with Sun, Haunschild, Xiao, Bulik, Scuseria and Perdew [22]. This paper has been cited 69 times (date of search 05 March, 2019). For the size of a topic like DFT, even a rather poor marker paper should not be cited much less. This paper is a rather special paper which presents density functional approximations which have not yet been widely applied.

Listing 1 (without the command "`removeCR`" and the command "`RPY: [1950, 1990, false]`" replaced as "`RPY: [1950, 2017, false]`" in order to also capture newer papers in the initial step) is used for the initial RPYS-CO using Sun, Haunschild, Xiao, Bulik, Scuseria and Perdew [22] as a marker paper. In the first step, we only look at the ten most frequently occurring CRs ordered by NCR as shown in **Table 2**.

Table 2. Ten most frequently occurring CRs of the RPYS-CO using papers co-cited with Sun, Haunschild, Xiao, Bulik, Scuseria and Perdew [22] for the time frame 1950-1990

No	RPY	CR	NCR
CR20	2013	Sun JW, 2013, Journal of Chemical Physics, V138	51
CR21	1996	Perdew JP, 1996, Physical Review Letters, V77, P3865	45
CR22	2003	Tao JM, 2003, Physical Review Letters, V91	37
CR23	1965	Kohn W, 1965, Physical Review, V140, P1133	36

CR24	2006	Zhao Y, 2006, Journal of Chemical Physics, V125	31
CR25	2009	Perdew JP, 2009, Physical Review Letters, V103	29
CR26	2012	Sun JW, 2012, J Chem Phys, V137	27
CR27	1988	Becke AD, 1988, Physical Review A, V38, P3098	26
CR28	2008	Zhao Y, 2008, Theoretica Chimica Acta, V120, P215	25
CR29	2008	Perdew JP, 2008, Physical Review Letters, V100	25

We see that CR21, CR23, and CR27 were mentioned in the previous section as possible suitable marker papers. Furthermore, CR21 has a rather similar NCR value as our rather poor marker paper (CR20). This is already an indication that our choice of the initial marker paper might not have been very good. Therefore, we use CR21 as a new marker paper in the next step of the iterative RPYS-CO, this time using again **Listing 1**. The resulting NCR curve is compared with the one from the RPYS by Haunschild, Barth and Marx [4] based on a keyword search in controlled vocabulary in **Fig. 3**. Both NCR curves show peaks at the same locations although the heights of the peaks differ substantially. The RPYS-CO spectrogram using CR21 as a marker paper is shown in **Fig. 4**. The corresponding peak papers are listed in **Table 3**. Nine out of 14 CRs in **Table 3** also appeared as peak papers in the RPYS-CO analysis using Becke [9] as a marker paper. The other five CRs also appeared in the other RPYS analyzes although not as pronounced peak papers. CR30 studied elastic behavior of a crystalline aggregate. CR31 discusses relations between the elastic and plastic properties of pure polycrystalline metals. Both CRs are important for several applications of DFT to solid state physics. CR37 presents studies of electrochemical photolysis of water at a semiconductor electrode. The latter three CRs are experimental studies which were extensively referenced in DFT papers. The results in CR40 were used to construct correlation functionals. In CR41, a very popular employed ansatz for molecular dynamics in DFT is proposed. The slight differences in the two RPYS-CO analyses presented here show the different foci which can be carried over from different maker papers into the RPYS-CO results. At least when studying large topics, it might be advisable to perform multiple iterative RPYS-CO analyses in practice and combine the results.

Table 3. Peak papers of the RPYS-CO using papers co-cited with CR21 for the time frame 1950-1990

No	RPY	CR	NCR
CR30	1952	Hill R, 1952, Proceedings of the Physical Society of London Section A, V65, P349	1185
CR31	1954	Pugh SF, 1954, Philosophical Magazine, V45, P823	1294
CR32	1955	Mulliken RS, 1955, Journal of Chemical Physics, V23, P1833	833
CR33	1964	Hohenberg P, 1964, Physical Review B, V136, PB864	7509
CR34	1965	Kohn W, 1965, Physical Review, V140, P1133	8946

CR35	1970	Boys SF, 1970, Molecular Physics, V19, P553	1138
CR36	1972	Hehre WJ, 1972, Journal of Chemical Physics, V56, P2257	628
CR37	1972	Fujishima A, 1972, Nature, V238, P37	605
CR38	1976	Monkhorst HJ, 1976, Physical Review B, V13, P5188	13558
CR39	1980	Vosko SH, 1980, Canadian Journal of Physics, V58, P1200	2180
CR40	1980	Ceperley DM, 1980, Physical Review Letters, V45, P566	1980
CR41	1985	Car R, 1985, Physical Review Letters, V55, P2471	1242
CR42	1988	Lee CT, 1988, Physical Review B, V37, P785	4981
CR43	1988	Becke AD, 1988, Physical Review A, V38, P3098	4048

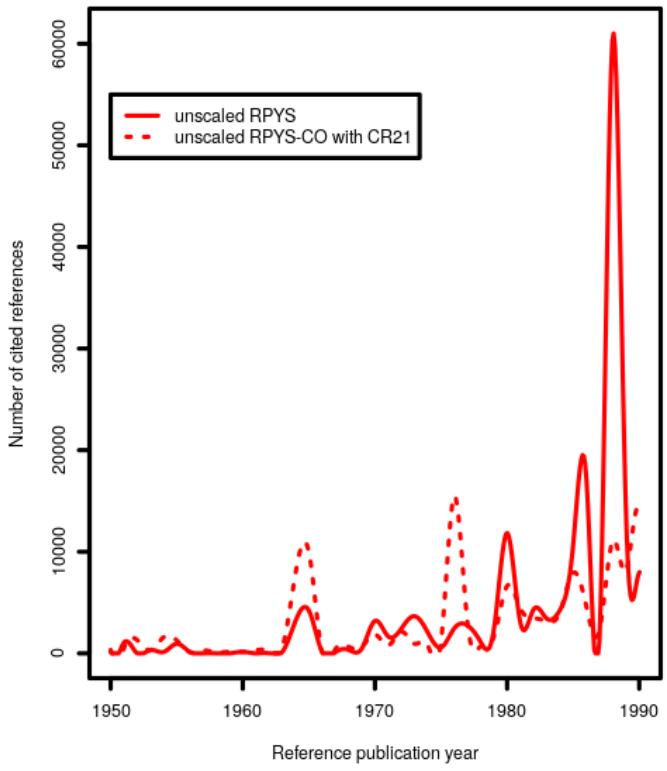

Fig. 3. Comparison of NCR curves from the RPYS analysis using DFT papers from a keyword search in controlled vocabulary of the CAS thesaurus for the time frame 1950-1990 from Haunschild, Barth and Marx [4] with the RPYS-CO analysis in this study using CR21 as a marker paper

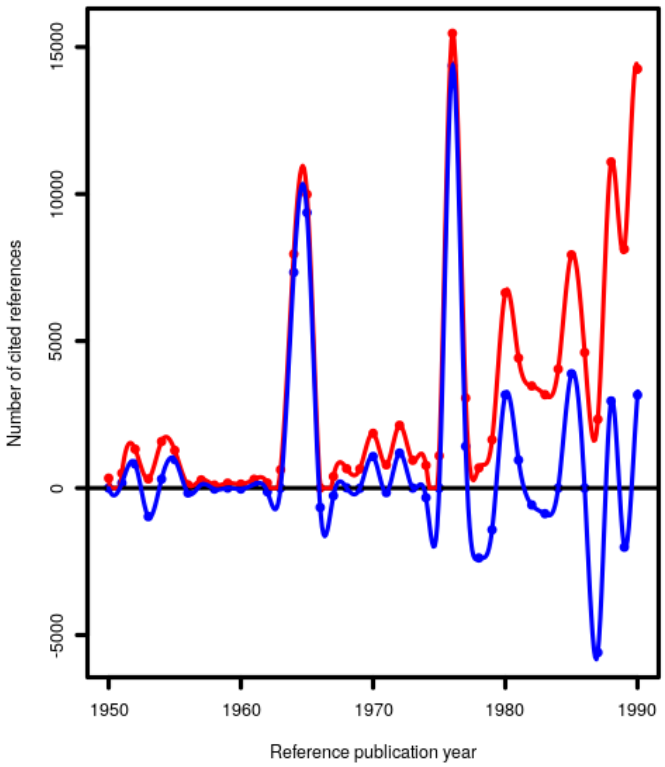

Fig. 4. RPYS-CO analysis using papers co-cited with CR21 for the time frame 1950-1990. The red curve and dots show the NCR values. The blue curve and dots show the five-year median deviation. Both curves are used to locate peaks.

4 Discussion and Conclusions

Overall, the results of the RPYS-CO presented here and the RPYS of Haunschild, Barth and Marx [4] are very similar although the methodology and the employed database are quite different. Haunschild, Barth and Marx [4] started from a keyword search in index terms of the CAplus database (controlled vocabulary of the database provider) while the RPYS-CO performed in this study is based on papers co-cited with one marker paper in the WoS database. Despite the different approaches, quite similar results were obtained.

The approach of using a marker paper for finding other seminal papers in research fields might become an interesting tool for scientists to explore their research fields in addition to a keyword-based literature search. If a good marker paper is not known a priori, the RPYS-CO methodology can be applied iteratively.

The RPYS-CO analysis has several advantages over build-in functionalities of several databases: (i) not only source records of the database can be found but also seminal papers which appear only in the cited references. (ii) The CRExplorer provides additional analysis features, such as filtering for papers which had a significant impact over many citing years by using the advanced indicators. (iii) The RPYS-CO methodology is not restricted to any database. In principle, the RPYS-CO methodology can be applied to datasets from any database which has cited references included.

The focus on the cited references, however, has a disadvantage: Search results have to be processed outside the database or reimported into the database. Such a reimport is usually not complete as non-source records appear in the results of an RPYS analysis.

CitNetExplorer (see http://www.citnetexplorer.nl/), a tool based on Eugene Garfield's work on algorithmic historiography and the corresponding program HistCite (the program is no longer in active development or officially supported) show the time evolution of a given research topic via the citation network of major papers, which have been selected before using other methods. In contrast to CitNetExplorer, using CRExplorer and applying the RPYS-CO method aims to detect the publications most important for the relevant community during the evolution of a given research topic. An alternative method for retrieving relevant literature based on co-citations is the Related Records Search function offered by the WoS. However, this method retrieves a publication set without any weighting with regard to the citation impact within the relevant community.

Future work should employ other databases and look for similar marker papers in DFT. Also, the method should be applied to other research topics.

5 References

1. Wacholder, N.: Interactive Query Formulation. Annu Rev Inform Sci Technol 45, 157-196 (2011)
2. Haunschild, R., Bornmann, L., Marx, W.: Climate Change Research in View of Bibliometrics. PloS one 11, 19 (2016)
3. Wang, B., Pan, S.Y., Ke, R.Y., Wang, K., Wei, Y.M.: An overview of climate change vulnerability: a bibliometric analysis based on Web of Science database. Nat. Hazards 74, 1649-1666 (2014)
4. Haunschild, R., Barth, A., Marx, W.: Evolution of DFT studies in view of a scientometric perspective. J. Cheminformatics 8, 12 (2016)
5. Marx, W., Haunschild, R., Bornmann, L.: Global Warming and Tea Production-The Bibliometric View on a Newly Emerging Research Topic. Climate 5, 14 (2017)
6. Small, H.: Cocitation in Scientific Literature - New Measure of Relationship Between 2 Documents. J. Am. Soc. Inf. Sci. 24, 265-269 (1973)
7. Marx, W., Haunschild, R., Thor, A., Bornmann, L.: Which early works are cited most frequently in climate change research literature? A bibliometric approach

based on Reference Publication Year Spectroscopy. Scientometrics 110, 335-353 (2017)
8. Marx, W., Bornmann, L., Barth, A., Leydesdorff, L.: Detecting the Historical Roots of Research Fields by Reference Publication Year Spectroscopy (RPYS). Journal of the Association for Information Science and Technology 65, 751-764 (2014)
9. Becke, A.D.: Density-functional exchange-energy approximation with correct asymptotic-behavior. Physical Review A 38, 3098-3100 (1988)
10. Lee, C.T., Yang, W.T., Parr, R.G.: Development of the Colle-Salvetti correlation-energy formula into a functional of the electron-density. Phys. Rev. B 37, 785-789 (1988)
11. Stephens, P.J., Devlin, F.J., Chabalowski, C.F., Frisch, M.J.: Ab-Initio Calculation of Vibrational Absorption and Circular-Dichroism Spectra using Density-Functional Force-Fields. J. Phys. Chem. 98, 11623-11627 (1994)
12. McLevey, J., McIlroy-Young, R.: Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science. Journal of Informetrics 11, 176-197 (2017)
13. Comins, J.A., Leydesdorff, L.: RPYS i/o: software demonstration of a web-based tool for the historiography and visualization of citation classics, sleeping beauties and research fronts. Scientometrics 107, 1509-1517 (2016)
14. Hohenberg, P., Kohn, W.: Inhomogeneous electron gas. Phys. Rev. B 136, B864-+ (1964)
15. Kohn, W., Sham, L.J.: Self-consistent equations including exchange and correlation effects. Physical Review 140, 1133-& (1965)
16. Perdew, J.P.: Density-Functional Approximation for the Correlation-Energy of the Inhomogeneous Electron-Gas. Phys. Rev. B 33, 8822-8824 (1986)
17. Perdew, J.P., Burke, K., Ernzerhof, M.: Generalized gradient approximation made simple. Phys. Rev. Lett. 77, 3865-3868 (1996)
18. Perdew, J.P., Ernzerhof, M., Burke, K.: Rationale for mixing exact exchange with density functional approximations. J. Chem. Phys. 105, 9982-9985 (1996)
19. Dunning, T.H.: Gaussian-Basis Sets for use in Correlated Molecular Calculations .1. The Atoms Boron Through Neon and Hydrogen. J. Chem. Phys. 90, 1007-1023 (1989)
20. Parr, R.G., Yang, W.: Density-Functional Theory of Atoms and Molecules. Oxford University Press, USA (1989)
21. Thor, A., Bornmann, L., Marx, W., Mutz, R.: Identifying single influential publications in a research field: New analysis opportunities of the CRExplorer. Scientometrics 116, 591–608 (2018)
22. Sun, J.W., Haunschild, R., Xiao, B., Bulik, I.W., Scuseria, G.E., Perdew, J.P.: Semilocal and hybrid meta-generalized gradient approximations based on the understanding of the kinetic-energy-density dependence. J. Chem. Phys. 138, 044113 (2013)