Research online Publishing
www.ronpub.com

Open Access

Open Journal of Information Systems (OJIS)
Volume 4, Issue 1, 2017

http://www.ronpub.com/ojis
ISSN 2198-9281

Purposeful Searching for Citations of Scholarly Publications

Fabian Rosenthal, Sven Groppe

Institute of Information Systems, University of Lübeck, Ratzeburger Allee 160, D-23562 Lübeck, Germany,
fabian.rosenthal@xennis.org, groppe@ifis.uni-luebeck.de

Abstract

Citation data contains the citations among scholarly publications. The data can be used to find relevant sources during research, identify emerging trends and research areas, compute metrics for comparing authors or journals, or for thematic clustering. Manual administration of citation data is limited due to the large number of publications. In this work, we hence lay the foundations for the automatic search for scientific citations. The unique characteristics are a purposeful search of citations for a specified set of publications (of e.g., an author or an institute). Therefore, search strategies will be developed and evaluated in this work in order to reduce the costs for the analysis of documents without citations to the given set of publications. In our experiments, for authors with more than 100 publications about 75 % of the citations were found. The purposeful strategy examined thereby only 1.5 % of the 120 million publications of the used data set.

Type of Paper and Keywords

Regular research paper: *scholarly publications, citations, citation data, citation search, citation matching*

1 Introduction

Bjor et al. [1] estimate that 1.35 millions English-language publication were published in the year 2006 and van Noorden [55] estimate 1.8 millions in the year 2011. Publishers and also authors publish the papers with an increasing speed in the web. But for the access to these publications a user account or a payment is often required. According to the estimation of Khabsa et al. [28] from 2013 overall 114 million English-language, scientific documents[1] are available in the web and 27 millions of these are freely accessible.

Moreover, bibliographic databases and search engines like Google Scholar [22], CiteSeerX [34, 56] and DBLP [33] were established in the web. Bibliographic databases comprise references to scientific papers and their metadata like title, author, year of publication and abstract. Citations ratios are used to determine the impact of authors or journals and make impacts comparable. The basic idea of the therefore used metrics is, that an often cited paper is a significant work because it is often referenced. Consequently, it is particular for young scientists relevant to be often quoted.

One of the most widespread author metrics is the h-index [25], introduced 2005 by Hirsch. A scientist has the index h, if h of his publications have each at least h citations and his other publications have h or less citations. These and other metrics (like the *mf*-index [44]) to measure the performance and output of scientists emphasizes even more the importance that scientists must be aware of citations to their papers.

Since papers refer to other publications or were

[1] Scientific documents are accordingly to [28]: Journal and conference papers, dissertations, master theses, academic books, technical reports and working paper. Patents are excluded.

cited in other publications, relations between these papers emerge. Furthermore, relations between authors, journals, conferences and fields of studies become visible when papers were grouped by these attributes. Citation relations are used to determine and to compare the influence of authors or journals.

Citation databases like Web of Science [52], Google Scholar or CiteSeerX are accessible in the web. However, the citation data is often not freely accessible. For the Web of Science, a membership is required and Google Scholar shows citations, but does not provide an API. CiteSeerX provides its publication but not the citations over an API. Furthermore, most databases are primarily bibliographic databases with a focus on collection and search for metadata of publications (without citations).

Hence it seems to be necessary for single researchers or whole institutes to collect their citations themselves. However, building a citation network of almost all publications is too cost-intensive in terms of time and processing power. In order to reduce these costs it is necessary to reduce the overall number of publications to be considered for analysis and extraction of references. Thus, we propose in this paper various search strategies for finding a big proportion of the overall citations quickly. We evaluate our search strategies in a comprehensive experimental evaluation.

2 BASICS

We describe popular bibliographic databases in Section 2.1, tools for extracting header information and references from scientific documents in Section 2.2 and further related work in Section 2.3.

2.1 Bibliographic Databases

This section discusses a selection of established bibliographic databases, which are relevant for the computer science field of study. We especially investigate if and to what extent the data of bibliographic databases can be used. According to [29] above half of the databases have no (CCSB, Google Scholar, Mendeley, ResearchGate) or no free (BASE, MAG) access, because a registration is necessary. The data of other databases like Arnetminer, arXiv, CiteSeerX and DBLP are freely accessible. Most databases provide an OAI-PMH interface, some additionally or exclusively an own *Application Programming Interface (API)* or download of the data.

OAI Protocol for Metadata Harvesting: The *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)* (also called OAI-protocol) serves for collecting and redistribution of publication metadata between different services [31]. The protocol has been developed by the *Open Archives Initiative (OAI)* for the purpose of facilitating the access to electronic publications, which are stored at servers of universities and institutions [48]. It defines the interface between so called data providers, which provide metadata of stored publications, and service providers, which access this data for further processing [31]. The protocol is widely used. The OAI officially registered over 3,000 data providers [49], which include many universities and large bibliographic databases like CiteSeerX, arXiv and BASE. Any data provider has the choice to register at OAI. But this is not a requirement to offer an OAI-PMH interface, such that there are more data providers than those registered at OAI. The protocol specifies the Dublin Core format as default metadata format, but each data provider can additionally support other formats [31]. OAI-PMH uses HTTP `GET`- and `POST`-queries.

CiteSeerX: CiteSeerX is a bibliographic database with several millions of entries in the area of computer science [50, 56]. At the same time CiteSeerX is also a search engine, which automatically searches for publications, extracts information from and stores data about these publications [34, 56]. Among other things, one module of CiteSeerX is the so called CiteSeerExtractor, which we describe in Section 2.2.3. The database offers its data through an OAI-PMH interface and per download upon request under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License [51].

arXiv: The bibliographic database arXiv covers many fields of study like computer science, physiques, mathematics and many more [11]. The database offers an OAI-PMH interface and additionally its data can be retrieved via an own XML-based API [9]. The OAI-PMH interface supports own formats (`arXiv` and `arXivRaw`) [10] besides Dublin Core.

Comparable to OAI-PMH the XML API [7, 8] can be also accessed via HTTP `GET`- and `POST`-queries, and returns metadata in an own XML format. In contrast to OAI-PMH the APIs offer a free search, but disallow a continuous retrieval of all data. Furthermore, the free search provides neither an option for a specific filtering of the results nor a refinement of the search.

DBLP: DBLP is a bibliographic database for the area of computer science and includes more than 3,7 millions publications [15, 16]. The data can be accessed via download of a large XML file [14] or via an XML

based API [33]. The format of the XML file as well as the result of the API orientate themselves according to BibTex. According to Ley [33] this so called DBLP-XML format can be regarded as BibTeX in XML form with small modifications. Each publication owns an unique identifier called `key`. BibTeX declares only one field for all authors, but in DBLP-XML each author is stored in an own `author`-field. Further information about the format is provided in [33].

Microsoft Academic Graph: Microsoft Adademic Search (MAS) was a bibliographic database and search engine for scientific publications comparable to Google Scholar. In contrast to Google Scholar, MAS offered an API [39] for the access to its data after authorization via a personal key [38]. The database is not updated any more and the API is not available any more. The successor Microsoft Academic and its associated Academic Knowledge API [36] will probably offer an API with fees [37].

The *Microsoft Adademic Graph (MAG)* [40] is a part of the so called Microsoft Academic Service [45] and the bibliographic database Microsoft Academic is also a part of the service. The graph contains publications, authors and citations. It is based on data of MAS and the Microsoft's search engine Bing and is the underlying database of Microsoft Academic [36]. The MAG was updated once and could be downloaded in form of *Comma Separated Values (CSV)*-files on a web page of Microsoft Research [40]. However, the data currently cannot be downloaded any more. The graph is very comprehensive, does not restrict itself to a certain field of study and contains many citations. Beside about 120 millions publications and authors, and 312 millions citations among publications, the graph contains also conferences, journals, institutes as well as keywords and URLs of publications.

2.2 Extraction Tools

The databases contain metadata like the title and the authors and sometimes additional links to the documents. A larger number of references between publications is provided by none of the observed databases. The only exception is the Microsoft Academic Graph, but which was recently removed from its webpage. Accordingly the publications have to be extracted manually to find these references. Thereby the metadata and links provided by the databases can be used to find the documents.

A scientific document is composed of different parts. The header of the document contains information about the publication itself like the title, names of the authors and an abstract. The text body represents the proper text of the work. Beyond text content elements like figures, tables, equations and code are part of the body. The bibliography section at the end of the document lists the cited publications. A reference is composed for instance of the title, authors, name of the journal, publication year and page references. Additional parts at the end of the document beside the bibliography can be acknowledgments or an appendix with further explanations.

To find citations primarily the bibliography is relevant since it lists all used sources. However, the text body and especially the header are relevant, too. The metadata from the header can be used to identify a document. Moreover the metadata can be used to obtain more information about the publication since the provided data from the databases are often incomplete or flawed. The text body is relevant when the context of the citation should be extracted.

The popular tools for the extraction of scientific documents are using linear chain condition random fields (CRFs), a statistical method to create probabilistic models for segmentation and labeling of linear sequential data [30]. CRFs can be trained with different learning approaches. They are used for citation extraction since the labeling of a character sequence depends on the neighborhood [30] and and thus the context is taken into account.

There are several libraries available, which implement Linear Chain CRFs. The C++-library CRF++[2] and the C-library Wapiti[3] are widely-used in the area of citation extraction. Among other tools, ParsCit and FreeCite use CRF++ [12, 2], whereas Grobit applies Wapiti [23]. We introduce ParsCit in Section 2.2.1 and Grobid in Section 2.2.4 in more detail.

2.2.1 ParsCit

ParsCit [12] is an open source CRF-parser for the extraction of metadata from publications. The input of the library (coded in Perl) is the document as text. The library first analyses the logical structure of the document. Afterwards the library identifies header metadata like title and author, text components like sections and section titles, and references [27]. The core of the library is a trained CRF model, which labels references. According to Councill et al. [12] the library internally uses CRF++ as CRF-library. Via a heuristic model[4] the sections containing the references are searched for in the overall logical

[2] `https://taku910.github.io/crfpp/`
[3] `https://wapiti.limsi.fr`
[4] We refer the interested reader to [12] for more details: First strings like *References* and *Bibliography* are searched for and afterwards further criteria are checked.

structure. Furthermore, the contexts of the considered references are determined within the overall document via the given reference markers.

The focus of ParsCit is the extraction of references, but also metadata of the document header is retrieved. The CiteSeerExtractor and its successor PDFMEF (see Section 2.2.3) extract via other tools first of all the text of the analyzed publications, which are stored as PDF documents. The text serves then as input for ParsCit in order to extract the references. Both tools use other tools than ParsCit for the extraction of the header metadata [34, 57].

2.2.2 SVMHeaderParse

SVMHeaderParse [24] extracts metadata in the head of a publication by classifying the rows of the head by a *Support Vector Machine (SVM)*. Analogous to ParsCit, SVMHeaderParse applies regular expressions for determining header metadata like titles and authors [56]. Afterwards, the SVM classifies each row of the head in one or several of 15 classes (author, title, abstract, et cetera[5]) [24]. The classification is iteratively improved by considering e.g. class labels of neighbored rows.

SVMHeaderParse is applied in CiteSeerExtractor for the extraction of the header metadata. The successor PDFMEF uses Grobid for this purpose, because Grobid delivers better results in experiments [57].

2.2.3 CiteSeerExtractor und PDFMEF

CiteSeerExtractor [34, 56] is a web service for automatically extracting metadata from scientific publications. The extractor has been designed for the CiteSeerX database and the code is publicly freely accessible[6]. The extractor includes a server, which offers a *Representational State Transfer (REST)* API for the extraction of metadata from documents. The document formats *Portable Document Format (PDF)*, *PostScript (PS)* and TXT test files are supported. The extraction itself consists of four components:

- *Text extractor*: The server temporarily stores an uploaded document, from which the text is extracted and stored as text file. For this purpose PDFBox[7] is used for PDF documents and ps2txt [56] for PS files.
- *Citations extractor*: Afterwards the CRF extractor ParsCit (see Section 2.2.1) is applied for the purpose of extracting references.
- *Header extractor*: The Support Vector Machine SVMHeaderParse (see Section 2.2.2) is used for the extraction of the metadata from the head of the document.
- *Text body extractor*: The text body is determined by removing the references. The text body can then be used for text analyzes or for free text search.

Basically, the CiteSeerExtractor offers an API and implements wrappers for the integration of existing tools, which do the underlying main work of the extraction. The result of these tools are afterwards combined in order to mine more information (or information with a higher quality) from the considered document.

PDF Multi-Entity Knowledge Extraction Framework (PDFMEF)[8] [57] is the successor of CiteSeerExtractor. The idea of this tool is to extract more information from a document than already done by CiteSeerExtractor. Hence, PDFMEF extracts additionally figures, tables and algorithms from documents. The citation extraction is still done by ParsCit, but PDFMEF uses Grobid (see Section 2.2.4) instead of SVMHeaderParse for the extraction of the header metadata. Figures and tables are extracted by using PDFFigures[9] [6] and algorithms via [53]. PDFMEF implements, analogous to CiteSeerExtractor, wrappers for the different tools and offers an API. In contrast to CiteSeerExtractor the API is not a REST API, but a Python API, which processes one document on average within 1.3 seconds on a server with 16 cores [57].

2.2.4 Grobid

GeneRation Of BIbliograhic Data (Grobid) [35] is an independent library for the extraction of metadata from scientific publications. The code of the tool is open source[10]. Grobid extracts header metadata, the text body as well as references from text and PDF documents. The Java-library Grobid classifies like ParsCit via Linear Chain CRFs. For this purpose, it uses the Wapiti CRF-library. In addition to a Java API, Grobid also offers a REST API and can be used as a web service. According to [23], Grobid processes 4,000 PDFs in 10 minutes and 3,000 references in 18 seconds on a modern MacBook Pro.

The output format of the extractor is TEI [47], which is an XML format of the *Text Encoding Initiative (TEI)* for publications.

[5] See [24] for all classes
[6] https://github.com/SeerLabs/CiteSeerExtractor
[7] https://pdfbox.apache.org/
[8] https://github.com/SeerLabs/new-csx-extractor
[9] https://github.com/allenai/pdffigures
[10] https://github.com/kermitt2/grobid

2.3 Further Related Work

This section introduces further related work (not discussed so far) relevant to the topic of this paper. Section 2.3.1 summarizes the contributions to search and extraction of publications. Section 2.3.2 provides an overview over citation networks, which are an essential application for citation data and which use citation data to determine new knowledge. Finally Section 2.3.3 introduces contributions of further applications of citations and the search for citations without restricting itself to scientific publications.

2.3.1 Search for and Extraction from Publications

Lafferty et al. [30] introduce Conditional Random Fields for labeling sequences of strings. This method is applied by widely-used citation extractors like ParsCit [12] and Grobid [35]. More recent publications like Clark et al. [6] deal with extracting further components like tables and figures from documents.

In the area of searching for publication CiteSeerX [32, 34, 56] should be mentioned together with its extractor CiteSeerExtractor. Wu et al. [58] describe how CiteSeerX automatically retrieves publications via a web crawler. CiteSeerX primarily uses a white-list strategy [59], i.e., a list of given websites are investigated. Based on Caragea et al. [3] a filter is used in order to detect whether or not a document is a scientific publication. The filter checks, whether or not the structure of the considered document follows the one of a scientific publication with e.g. an abstract somewhere in the beginning and a list of references somewhere in the end of document. Wu et al. [57] introduces PDFMEF, which internally uses ParsCit and Grobid, and provides the extraction of tables and figures, which allows to retrieve more and preciser information from a document [6].

2.3.2 Citation Networks

A publication is often cited by several other papers and cites itself also other contributions that also contain citations of publications. Hence, citations determine a directed graph, where the publications may become the nodes in the graph and citations directed edges from the citing to the cited publication. The nodes may also represent authors, journals, conferences or fields of study by grouping the publications accordingly. Citation networks are applications of citation data in order to determine new knowledge like the connectivity of journals or different communities in fields of study.

Connectivity of journals: Nerur et al. [41] investigate the citation data of 27 journals in order to determine how many publications of a journal A cite and are cited by how many publications of another journal B. Based on this citation network the authors analyze, how much the considered journals are connected with each other. On the one hand the authors detect thematic groups, i.e., not surprisingly, journals with related topics are more connected with each other. Additionally the authors recognize a separation of European and North American journals.

Communities in fields of study: Newman et al. research on the search for and evaluation of community structures in networks [42] and provide algorithms [42, 43] for this purpose. These cluster algorithms analyze a network by iteratively removing edges, such that the network is divided into different communities.

Kajikawa et al. [26] apply these cluster algorithms on publications related to the topic sustainability science in order to achieve an overview over the topics of this research area. Their analysis groups about 10,000 publications in 15 fields like agriculture, fisheries and tourism. They additionally determined the states for each of these fields, which published most publications for this field. Community structures are also computed and analyzed for a lot of other research areas by considering citation networks, e.g., in the area of physique [5] or learning analytics [13].

2.3.3 Other Applications of Citation Search

Besides scientific publications there are other applications, for which citations are relevant. Many publications deal with patents. Two examples include Chang et al. [4] and Érdi et al. [21], which analyze citations of patents in order to predict new, relevant technologies. The basic idea is that if a patent A cites another patent B, then A builds upon the knowledge of B. Based on this relationship, the patents can be grouped. It can be investigated in which groups many current patents are. Chang et al. and Érdi et al. use the data set of the United States Patent and Trademark Office [54] for the application of their methods.

Another, less investigated field deals with the citations of news articles in the Internet. According to Spitz et al. [46] there are structural similarities to scientific citation networks, such that citation networks of news articles can be analyzed with the existing methods. They collected about 59,000 articles of relevant German online news sources like SPIEGEL, ZEIT and Tagesschau. They analyzed this data set by looking at the references among these articles in order to investigate the distribution of information in mass media.

3 SEARCH STRATEGIES

The processing of a large number of publications is one of the most big challenges. According to the estimations of Khabsa et al. [28] and van Noorden [55], 114 millions English publications are accessible in the Internet and 1.35 millions contributions are published in the year 2011. The MAG data set of 2015 contains 121 millions publications. Also the relative small data set of DBLP consists of several millions publications and the statistics shows a non-linear grows for the years 1995 to 2016 [15].

Most data sets only provide metadata, in which data about citations are missing and a manual extraction of the citations for all publications of a bibliographic data set is costly. In order to reduce these costs, our proposed search strategies are designed for purposeful searching for citations of a given set publication (like the one of an author or a venue). The strategies propose publications in which citations to this given set of publications are likely to occur.

Let $\mathcal{P}$ be the set of publications for which citations should be searched for. Strategies typically first analyze certain properties of $\mathcal{P}$, such that publications are ranked for the automatic extraction of references based on this analysis. Section 3.1 introduces these search strategies. Another approach is to cluster publications, authors, journals or conferences according to their considered topics and then to search within these groups. We introduce this kind of strategies in Section 3.2.

3.1 Metadata Strategies

So called metadata strategies use directly the metadata of papers. One the one hand the proposed strategies examine the metadata in the publication set $\mathcal{P}$ to determine, for example, all authors in the set. On the other hand the strategies consider the metadata of other publications, for example to search for all papers of an author. This already corresponds to the first strategy:

- **Author strategy**: Inspection of all papers of the authors who write the publications in $\mathcal{P}$. The goal of the strategy is to find self-citations. For the case $\mathcal{P}$ represents all publications of one author, the strategy consequently checks all co-authors and can be processed recursively: When citation are found by this strategy the authors of these publications can be checked to find further citations.
- **Journal strategy**: Inspection of the publications of all journals in which the papers $\mathcal{P}$ are published. The idea is hereby that scientists of the same study field publish in the same journals or refer to the papers of these journals in their own work.
- **Conference strategy**: Analogous to the journal strategy, but with conferences.

3.1.1 Optimizations

Improvements of the basic strategies proposed in Section 3.1 restrict the search space such that a) less publications need to be analyzed (see the first optimization in the following enumeration), b) reorder the papers such that citations may be found more early (see second optimization), which allows to process these results more early in succeeding processing steps (like publishing them on a website), and c) consider the already found citing publications to search for further citations (see the third optimization):

1. **Minimum year of publication**: Based on the publication year, the oldest paper in $\mathcal{P}$ can be determined. All publications before this year need not be considered. Alternatively, a year can also be predefined by the user from which on papers should be searched for. This can be for instance the year in which an author publishes her/his first paper or a journal released its first edition.
2. **Sorting by frequency**: The search strategies determine for instance the authors, journals or conferences in $\mathcal{P}$. The entries can be sorted in descending order by frequency of occurrence in $\mathcal{P}$. The idea for the author strategy is that a commonly occurring author has probably more often cited papers in $\mathcal{P}$. This idea also applies for the other strategies. If the majority of the publications were released in a particular journal, it is probably worth to examine it first.
   (a) Applying a threshold: Only entries are considered that occur at least s times in $\mathcal{P}$. Thereby s is a predetermined threshold. Entries below this threshold can be either discarded or examined later. This can be more meaningful then to examine a large number of papers due to a rarely occurring property in $\mathcal{P}$.
3. **Citing papers**: Let us assume that through a strategy a set $\mathcal{P}_{Z1}$ of papers is already found that cite the publications in $\mathcal{P}$. One or multiple strategies can be re-applied on the citing papers $\mathcal{P}_{Z1}$. This is applicable up to a predefined depth n. The idea here is to check the citing papers since they possibly cite the original papers of $\mathcal{P}$ again.

3.2 Cluster Strategies

We call all those search strategies *cluster-strategies*, which group publications, authors or venues (i.e.,

journals, conferences and workshops) by related topics. The grouping can be based on already found citations or on metadata of publications like authors and keywords. For example, related journals can be detected by many common authors or by many citations among articles of the considered journals. However, we want to first look at the grouping of publications by topics by considering keywords in more detail.

3.2.1 Grouping Publications by Topics and Fields of Study

Many publications contain or are associated with keywords (given by their authors) in order to simplify the search for these publications. Keywords belong to - like the title or the publication year - the metadata of a publication. Most OAI-provider provide keywords by using the `subject`-tag of the Dublin Core format. MAG also contains keywords, but DBLP data does not. It is also possible to extract keywords from the documents by using extraction tools like Grobid.

For the purpose of grouping publications, keywords must be assigned to one or more fields of study. Only in this way two publications having the keywords *machine learning* and *data processing* respectively can be grouped according to the computer science field of study. Determining assignments of keywords to fields of study is a separate big topic, which we do not discuss further in this paper. Only considering fields of study is also too course-grained, as one field of study already contains millions of publications. Instead complex ontologies must be built and considered for each field of study itself, which allows to restrict the number of considered publications further. For example, an ontology for computer science may contain *machine learning* as subclass of *artificial intelligence* being itself a subclass of *computer science*, and *data processing* may be a subclass of *data management*, which may again be a subclass of *computer science*.

MAG contains these data and hence can be used in order to assign the keywords of a publication to fields of study. Having these data, not only publications can be assigned to fields of study, but also authors, journals and conferences. One just has to look at the keywords (and hence the fields of studies) of their publications. The most often occurring field of study is the main one of the author, journal or conference. In this way we may also calculate a numerical indicator for expressing how much an author, journal or conference belongs to a specific field of study (e.g., by determining the percentage of publications belonging to the specific field of study).

Sinha et al. [45] claim for MAG that new publications are assigned to fields of study based on already assigned publications. For this purpose, they use an *in-house knowledge base*, but do not describe the assignment process in more detail.

Table 1: Example of a citation matrix for three entries. (The columns contain the number of references, e.g.: publications of A refer to 30 publications of B.)

References ↓	A	B	C
A	200	10	0
B	30	102	45
C	25	5	165

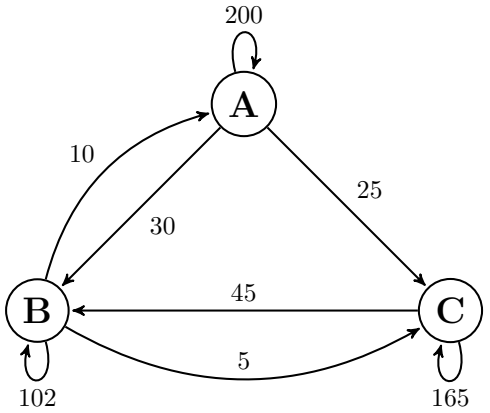


Figure 1: Directed graph of the citation matrix in Table 1. (The directed edge from A to B with value 30 can be read as: publications of A refer to 30 publications of B.)

Grouping by already found citations: Besides looking at the keywords, also already found citations can be used to cluster publications. In this way citation relatedness between authors, journals or conferences can be determined and further analyzed. The idea behind this grouping is the observation that publications typically cite those publications with a related topic. For example, Nerur et al. [41] analyze the citation relatedness between journals, where a grouping by topic has been recognized.

Calculating citation relatedness requires that citations are already known. The citations could be already determined by other strategies or can be retrieved from other databases like the one of MAG. Afterwards a *citation matrix* must be calculated. Table 1 contains an example of a citation matrix with three entries A, B and C. We assume A, B and C to be journals in the following paragraphs. Note that the entries of a citation matrix may be also authors or conferences. The first column of the citation matrix represents that publications

of A refer 200 times to publications of A, 30 times to publications of B, and 25 times to publications of C. Here, a citation is regarded as one reference between two publications. Hence, for example, one publication of A may refer to several publications of B, but also several publications of A may refer to the same publication of B, both increasing the value in the citation matrix. The diagonal in the citation matrix contains the *self-citations* (within the same journal, conference or of an author). The citation matrix represents a citation network. Figure 1 presents the citation matrix of Table 1 as directed graph.

To search for the citations of the journal B, it may be a good idea to look at the citation relatedness of the journals A and C to B. For this purpose, we consider the values of the second row of Table 1. We first sort the absolute values of this row in order to determine the journal having the most citations to journal B. As C 45 times and A 30 times refer to publications of B, we should first search for publications within C and afterwards within A. The idea is that publications of C already cited B quite often, such that these citations may cover related topics. Hence there is a high probability that other (not so far analyzed (maybe more recent) publications) of C also refer to B. In order to avoid the repeated analysis of publications of C, the search algorithm should store and check already analyzed publications and their references.

Only considering the absolute values in the citation matrix can be misleading: For example, if A contains 80 publications with 30 references to B and C 8.000 publications with 45 references to B, then B and A are more related to each other in comparison to B and C. The number of citation must be hence considered in relation to the number of publications for which references are already known. The citation relatedness CR_{BA} of B to A is hence defined as follows:

$$CR_{BA} = \frac{\text{Number of references of } A \text{ to } B}{\text{Number of analyzed publications of } A} \quad (1)$$

with $CR_{BA} := 0$ in the case of missing references between A and B. As larger the determined value is as more related is B with A. As one publication can refer to more than one source, the value can be bigger than one. The direction of references is considered in this definition, such that CR_{BA} and CR_{AB} are in general not the same.

Cluster strategies for searching for citations: We propose concrete search strategies for grouping in this paragraph. As before, let $\mathcal{P}$ be the set of publications for which we search citations. The strategies under the item 1 in the following enumeration deal directly with publications, whereas the strategies under the item 2 group authors, journals or conferences respectively:

1. **Publications cluster strategy**: Grouping publications by topics. The search considers publications with the same topics as the publications in $\mathcal{P}$. The keywords of publications can be directly or indirectly used for the assignment of publications according to topics:
   (a) Overlapping keywords: Publications are grouped according to their keywords. Two publications are related to each other, if a certain number of their keywords are equal. The advantage of this strategy is that the keywords do not need to be assigned to a field of study (or to their subclasses). However, the keywords of authors are typically very individual, such that they are often too general or too specific, such that the strategy does not achieve good results.
   (b) Fields of study based on keywords: Search within the publications of the same fields of study to which the publications of $\mathcal{P}$ belong. According to their keywords publications can be assigned to fields of study (as described above). The idea of this strategy is based on the observation that publications of the same field of study refer more often to each other than to publications outside of their field of study.
2. X **cluster strategy**: This strategy groups publications according to $X \in \{$authors, journals, conferences$\}$ and searches within the publications of the same group of the publications in $\mathcal{P}$. For this purpose, first the entries of X (i.e., depending on the considered X all authors, journals or conferences) must be grouped. The grouping can be based on metadata of the publications or already retrieved citations:
   (a) Fields of study of publications: The fields of study of publications can be used in order to group the entries of X. For this purpose, the most often occurring fields of study can be chosen or a numerical indicator representing the frequency of the single fields of study can be used (as already discussed before).
   (b) Citation relatedness: If there are already citations known, we can use them for grouping by calculating the citation relatedness between two entries of X. According to our definition in Equation 1,

a higher value of the citation relatedness reflects a higher relatedness.

(c) Common authors: Two entries of X are grouped whenever they have a certain number of authors in common. The idea of this strategy is based on the observation that two journals or conferences are related with each other if many of their authors publish in both venues, as authors often only research in one area or in related areas.

4 Experimental Evaluation

The main goal of the proposed search strategies for citations is to find citations as fast as possible for a given set of publications. We need evaluation criteria and a comparison measure in form of a fixed data set for the evaluation of the proposed search strategies. We first define these in Section 4.1 before we present the results of the evaluation of the proposed metadata and cluster strategies in Section 4.2 and in Section 4.3 based on a fixed data set (see Section 4.1.2).

4.1 Criteria and Data Set

We need definitions for the effectiveness and efficiency of the strategies, which we calculate based on a fixed data set. The used data set should already contain citations, such that we can compare the number of citations found with our proposed search strategies with the total number of citations for the considered publications (based on the used data set).

4.1.1 Evaluation Criteria

The goal of the system is to detect as many citations as possible by analyzing as few documents as possible. Hence, both, the effectiveness as well as the efficiency, are relevant in the overall comparison. A strategy is effective, if it finds as many citations as possible. We determine hence the effectiveness by dividing the number of citations found by the strategy by the total number of citations (as given in the used data set):

$$\text{Effectiveness} = \begin{cases} \frac{\mathcal{F}}{\mathcal{T}} & \mathcal{T} > 0 \\ 0 & \mathcal{T} = 0 \end{cases} \quad (2)$$

with $\mathcal{F}$ number of citations found by the strategy and $\mathcal{T}$ the total number of citations (as given in the used data set). The result is within the interval 0 to 1 and represents the percentage of found citations: An effectiveness of 0 means that no citations are found, whereas all citations are found by the strategy for an effectiveness of 1.

A very effective strategy, which takes very long, is not efficient. It is efficient, if it finds the citations in very short time. The concrete cost in terms of time is dependent on the configuration of the experimental environment like hardware, software and even the length of the analyzed documents. An independent measure is the number of analyzed publications. The idea is in principal that a strategy analyzing the double number of publications takes also about the double time. Hence the efficiency is determined by dividing the number of found publications by the number of analyzed publications:

$$\text{Efficiency} = \frac{\mathcal{F}}{\mathcal{A}} \quad (3)$$

with $\mathcal{F}$ number of citations found by the strategy and $\mathcal{A}$ number of analyzed publications. The value of the efficiency is equal to or greater than 0. An efficiency of 1 means that for each analyzed publication a new citation is found on average. Note that one analyzed publication may also contain more than one reference to the publications of the considered set $\mathcal{P}$ of publications. The efficiency is 0 whenever no citations are found. The efficiency converges to 0 whenever all citations are already found but still new publications are analyzed.

4.1.2 Used Data Set

We use the Microsoft Academic Graph (MAG) [45] as experimental data set and comparison measure. The graph contains about 120 millions publications and authors and 312 millions references among publications. Furthermore, the graph contains data about journals, conferences, institutes, keywords and URLs of publications. Hence it was[11] presumably the most comprehensive, freely available citation graph and fulfills the requirements for our evaluation.

Our evaluation uses a normalized version of the Microsoft Academic Graph, which we call *Normalized Microsoft Academic Graph (MAGN)*. Basically, in MAGN the hexadecimal identifiers have been replaced with integer values and entries with too long strings have been ignored. For example, the maximum string length of publication titles is 250 in MAGN, whereas the original MAG contains titles even above 700 characters, as also wrong data has been stored as titles in MAG. We use the graph of the 17th November 2015, which could be downloaded from the website of Microsoft Research [40]. We additionally use the fields of study hierarchy of the version of the 5th February 2016 for the evaluation of the fields of study strategy (see Section 3.2). MAGN contains about 311 millions references and about 120 millions publications and authors[12].

[11] Unluckily, the graph has been recently removed by Microsoft.

[12] MAGN includes 119,806,634 authors (85,567 less than MAG), 120,305,892 publications (581,941 less) and 310,715,601 references (1,558,658 less)

Table 2: Distribution of data in the test set of authors

	Number of Publications	Number of Citations
Minimum	10,00	0,00
Maximum	300,00	9.904,00
Average	31,08	584,30

We have randomly[13] chosen 100 authors with at least 10 publications from MAGN. We call these chosen authors the *test set* in the following sections. The MAGN contains many authors with only one or two publications. In order to avoid that the test set contains many authors with only few publications, we additionally required a minimum number 10 of publications for the authors in the test set. Table 2 shows that the authors in the test set wrote 31 publications on average with a minimum of 10 and a maximum of 300 publications. The authors obtained about 584 citation on average, whereas the maximum is above 9,900 with a minimum of 0 citations.

4.2 Metadata Strategies

These strategies use the metadata of publications for their citation search. We evaluate the basic strategies, their optimizations and the combination of these strategies in the next section.

4.2.1 Basic Metadata Strategies

The basic metadata strategies are the author, journal and conference strategy, which we evaluate at first. Figure 2 presents the chronological sequence of the strategies for two different authors, which are typical for the test set of authors. The x-axis represents the number of already analyzed publications and the y-axis the number of already found citations. Figure 2(a) presents the chronological sequence of found citations for an author with few publications (35) and few citations (75), whereas (b) presents an author with some publications (71) and many citations (1,405). Hence we assume that the author of (b) is more known than (a) in her/his community. Both chronological sequences show that the author strategy finds many citations in relation to the number of analyzed publications, which results in a high rise of the line of found citations. The line for the journal strategy does not rise so much, as many more publications are analyzed. However, the difference between (a) and (b) is that for the more widely-known author the number of found citations by the journal strategy outweighs the number of found publications by the author strategy. The conference strategy does not find any citations for both authors. One reason could be that in the used MAGN data set only about 800,000 of the 120 millions publications are associated with a conference, whereas 44 millions publications are associated with journals. Hence the conference strategy should be re-evaluated once a more complete data set about publications and citations is available.

Figure 3 presents the (a) effectiveness and (b) efficiency of the strategies. The author strategy provides a significantly higher efficiency, as citations are found very quickly. These citations are self-citations of the considered author and of her/his co-authors. The journal strategy analyzes many more publications. Nonetheless, the journal strategy is more effective on average (see Figure 3(a)), because it finds more citations.

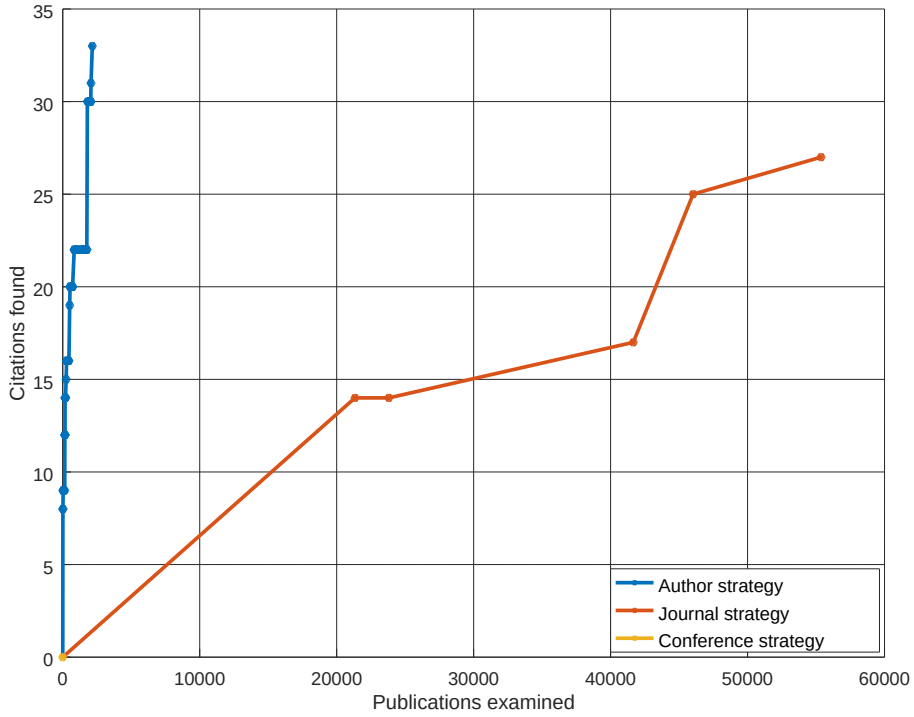
4.2.2 Optimization and Combination of the Basic Metadata Strategies

We propose in Section 3.1.1 general optimizations for our search strategies. Figure 4 presents the experimental results of two optimizations for the journal strategy in comparison to the basic strategy. The first optimization sorts the journals according to their frequency in the test set, such publications of the most often occurring journals are analyzed first. Figure 4 shows that the lines of the strategies are moved enormously to the y-axis. This rise is significantly higher in comparison to the original pure journal strategy. Hence many more citations are found in the beginning of the overall search. As only the order of the analysis is switched, the strategies find the same number of total citations. If this optimization is further combined with the restriction of the minimum publication year, then the effect is even stronger, because all publications before the minimum year of the publications in the test set are ignored for this strategy. The total number of found citations is slightly reduced, as MAG does not contain a publication year for every publication.

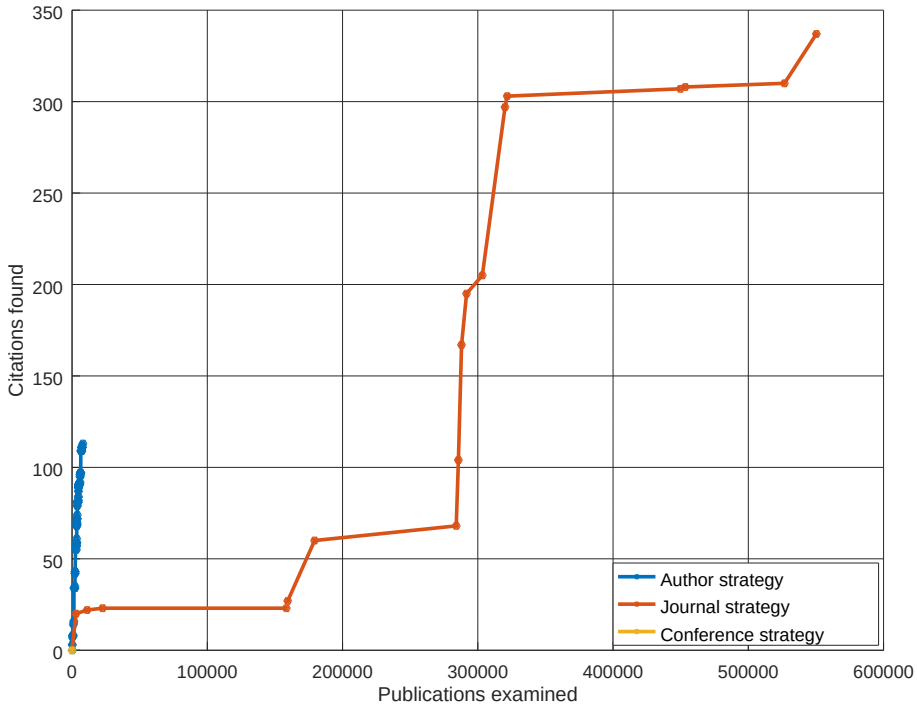
The effects of the optimizations of Figure 4 are representative for the other authors of the test set and also for the other strategies. As explained in Section 3.1.1, we can optimize in this way all metadata strategies (i.e., journal, conference and author strategies) in order to find citations faster. By applying a threshold, the efficiency can be increased, too.

In addition to applying the pure metadata strategies, we evaluate also the combination of these metadata strategies in the following paragraphs. If two strategies are combined, a second strategy will be applied after a first one, but the second strategy considers additionally the already found citations of the first strategy. For example, by combining the author with the journal

[13] We have used the pseudo-random Python function `random.randint(min, max)`.

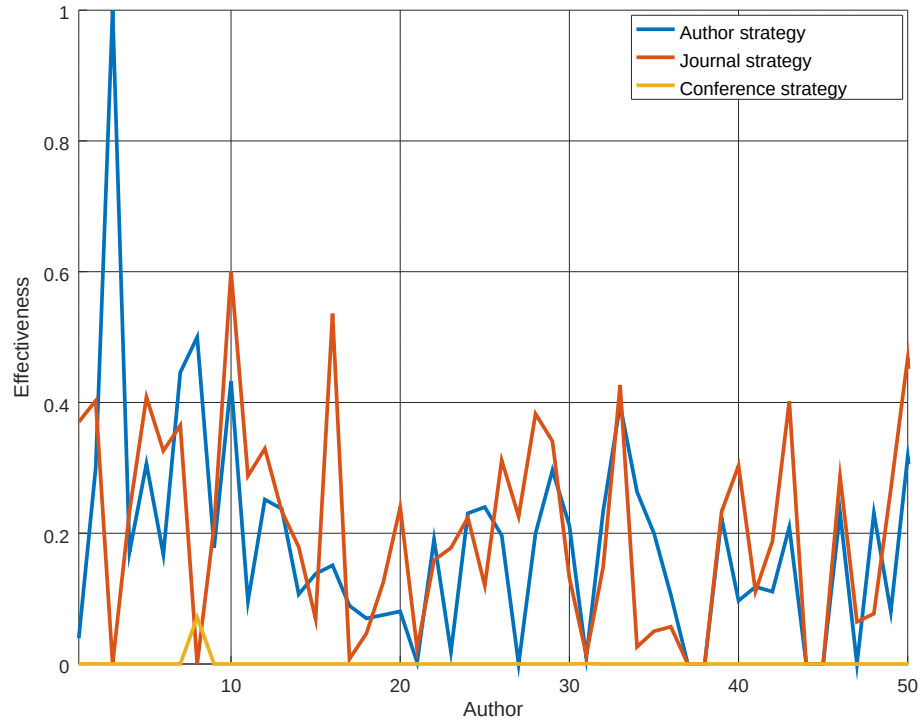


(a) Chronological sequence of an author with few publications (35) and citations (75)

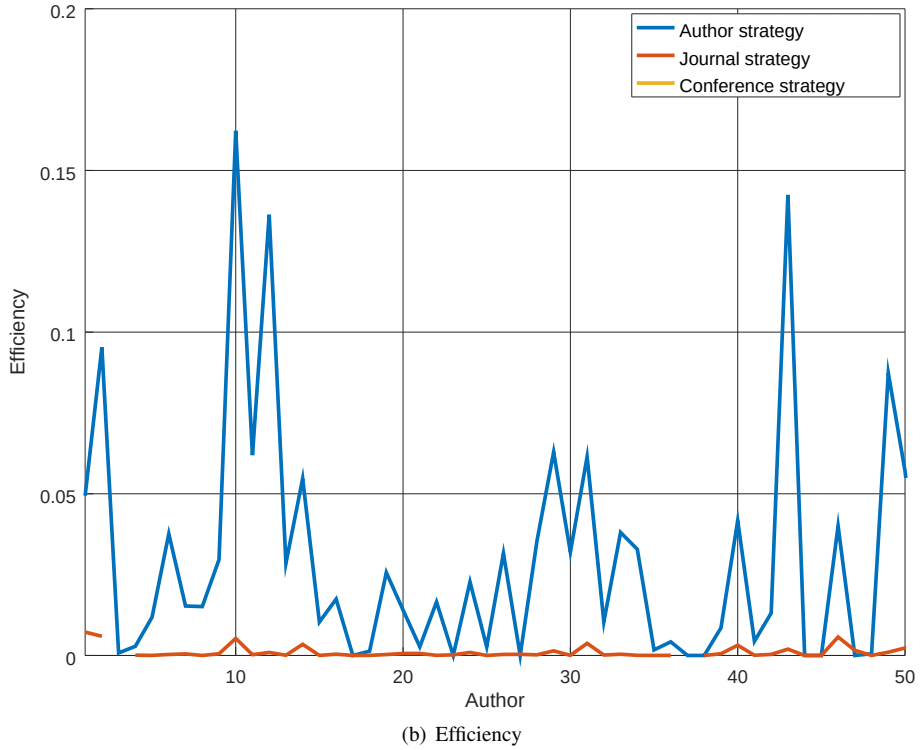


(b) Chronological sequence of an author with some publications (71) and many citations (1,405)

Figure 2: Two typical chronological sequences of the basic strategies of authors (a) with few publications and citations (b) and with some publications and many citations



(a) Effectiveness



(b) Efficiency

Figure 3: (a) Effectiveness and (b) efficiency of the basic metadata strategies.
(We present the first 50 authors of the test set of the authors. One experiment conforms to the analysis of the publications of one author.)

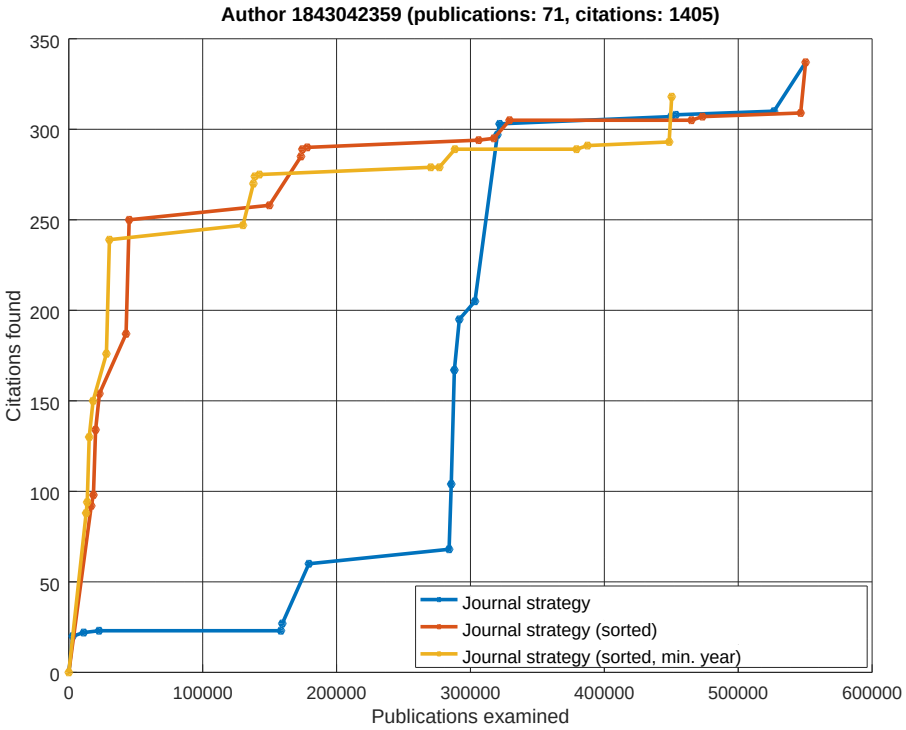


Figure 4: Comparison of the optimizations of the journal strategy.
(Sorted: Journals are descendant sorted according to their frequencies in the test set; min. year: Only publications are analyzed with a publication year equal to or greater than the minimum publication year in the test set.)

strategy, the author strategy first analyzes all publications of authors of the given publications for which citations are searched for. Afterwards, the journal strategy investigates all journals of the given publications and of the journals in which the preceding strategy already found citations. This applies also for a recursive application of the strategies, which we discuss in Section 3.1.1 as optimization. For example, if the author strategy is one time recursively applied, then the authors of the found citations are considered (without already analyzed authors).

We discussed already that the author strategy is the most efficient among the basic metadata strategies. Hence it is a good idea to first apply the author strategy. Furthermore, the author strategy should be applied recursively by searching citations among the authors of the citations in order to quickly find more citations. In this way succeeding strategies can be applied to a big set of of already found citations. We call the combination of the recursive author with the journal and conference strategies the *combination of the metadata strategies*. For these strategies, we already optimize by sorting according to the frequencies and considering the minimum publication year. We discuss the result of these combined strategies in Section 4.2.3.

Table 3: Effectiveness and Efficiency of the basic strategies and their combination

Strategy	Effectiveness	Efficiency
Combination	34.65 %	0.001116
Journal	17.12 %	0.001185
Author	15.24 %	0.033694
Conference	0.08 %	0.000311

4.2.3 Analysis of the Combination of Metadata Strategies

Table 3 contains the results about the effectiveness and efficiency of the metadata strategies and their combinations. The combined strategies have the highest effectiveness (ca. 35 %, i.e., on average 35 % of the total citations are found for the authors). The journal strategy has an effectiveness of 17 % and the author strategy 15 %. As already discussed, MAGN associates only very few publications with conferences, which explains the very low effectiveness of 0.08 % of the conference strategy. The author strategy owns the highest efficiency

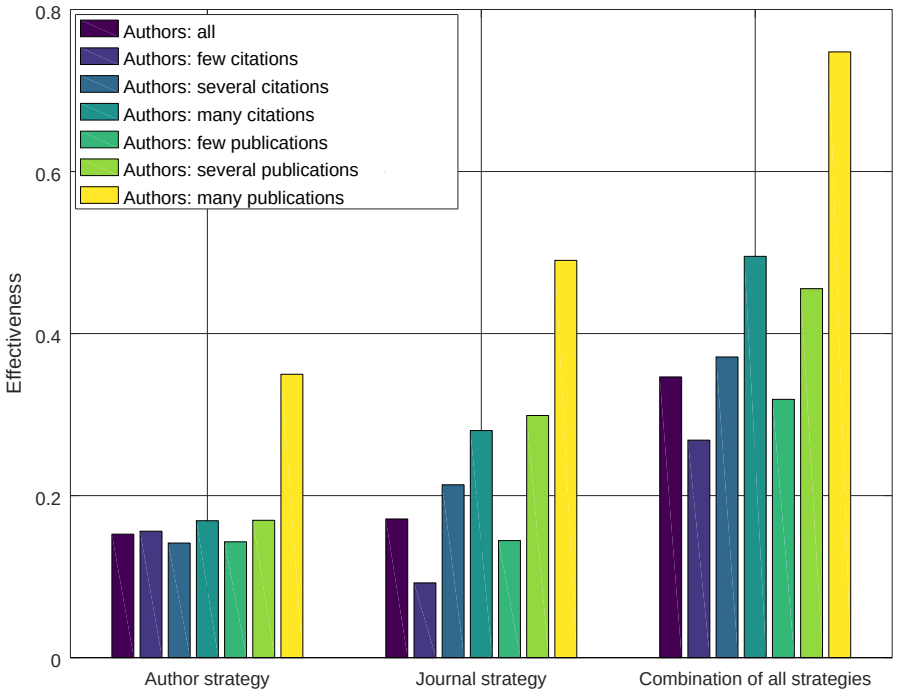


Figure 5: Effectiveness of the metadata strategies and their combinations for different classes of authors. (The first column presents the effectiveness of all authors in the test set. The next three ones represent the effectiveness for authors grouped according to their number of citations, whereas for the remaining columns the authors are grouped according to their number of publications.)

(0.03), because it analyzes only few publications and finds many self-citations.

As the characteristics of the considered authors (see Table 2) vary quite much with 0 to about 10,000 citations to 31.08 publications on average, we group the authors in several classes and analyze the results for each of these classes separately. We group the authors according to the number of their published papers and additionally to the number of their citations. Let z be the number of citations of an author. Then we group the authors in following classes:

- Authors without citations: $z = 0$
- Authors with few citations: $1 \leq z \leq 99$
- Authors with several citations: $100 \leq z \leq 999$
- Authors with many citations: $z \geq 1,000$

Furthermore, let p be the number of publications of an author. Then we additionally group the authors in following classes:

- Authors with few publications[14]: $10 \leq p \leq 49$
- Authors with several publications: $50 \leq p \leq 99$
- Authors with many publications: $p \geq 100$

Figure 5 presents the results of the author and journal strategies, and their combination according to the proposed classes of authors. As already discussed in Section 4.2.1, the MAGN contains too few associations of publications to conferences, such that we do not present the results for the conference strategy here. It is noticeable that the strategies have a significant higher effectiveness for authors with many publications. The combination reaches an effectiveness of nearly 75 %. Even the journal strategy obtains about 50 % effectiveness. The strategies are also better for authors with several to many citations (which typically have several to many publications). Hence the strategies achieve better results for well-known and active authors. Furthermore, we recognize that the author strategy has better results for unknown authors (based on their number of citations) in comparison to the journal

[14] All authors of the text set have at least 10 publications. We reason about this issue in Section 4.1.2.

Table 4: Effectiveness and efficiency of the publications cluster strategy (according to the field of study). (We restricted the maximum level of fields of study and consider only levels from 0 to 3. The minimum publication year and a sort according to the frequencies are considered with a restriction to maximal 10 fields of study as optimizations. Additionally, the last column contains the number of averagely analyzed publications per author.)

Max. Level	Effectiveness	Efficiency	Examined Publications
Level 0	21.23 %	0.000578	434,661.7
Level 1	21.11 %	0.000598	418,676.1
Level 2	20.87 %	0.000597	403,128.5
Level 3	19.35 %	0.000525	395,848.0

strategy, because their works are mainly cited by themselves or by their co-authors.

Figure 6 presents the efficiency in an analogous way to Figure 5. The author strategy has the highest efficiency for all classes of authors. The reason is that the author strategy analyzes only quite few publications and finds many self-citations. For the journal strategy and its combination, we achieve best results for authors with many citations and with many publications (for which we already obtained the highest effectiveness).

Summarizing the results, the author strategy achieved an effectiveness of 35 %, the journal strategy of about 50 % and the combination 75 %. For these results, about 26,000, 841,000 and 1,850,000 publications have been analyzed, which corresponds to 0.02 %, 0.70 % and 1.54 % of the publications in MAGN. Hence for the best result of 75 %, the purposeful search need to examine only 1.54 % of the 120 millions publications of MAGN in order to find three-fourths of all citations.

4.3 Cluster-Strategies

We introduced cluster strategies in Section 3.2 in order to group publications, journals, conferences or authors. In this way citations can be searched for in related groups, which are not restricted to a certain journal or conference. We analyze to what extent such kind of strategies are suitable for citation search in the following sections. We evaluate the effectiveness and efficiency of the publications cluster strategy in Section 4.3.1. Finally we explore the X cluster strategies in Section 4.3.2, if they are suitable for a purposeful citation search.

4.3.1 Publications Cluster Strategy

We group the publications according to their fields of study for the publications cluster strategy. We describe in Section 3.2 that MAG assigns a field of study to publications according to a hierarchy from level 0 to 3. As bigger the number, as more specific is the field of study. We analyzed 25 entries of the test set for this purpose.

There are many publications in level 0 and 1. Hence it is reasonable to optimize according to the minimum publication year and sort according to frequencies, such that less publications need to be considered. In order to further reduce the number of considered publications, we only consider 10 of the most often occurring fields of study. Experiments with more considered fields of study increases the effectiveness, but reduces the efficiency quite much. This is an expected result as more citations are found for more analyzed publications. The effectiveness decreases analogously for smaller numbers of maximal considered fields of study, but a higher efficiency is achieved.

The effectiveness of the publications cluster strategy (with the discussed optimizations and restrictions) reaches 21 % with an efficiency of 0.0006. In comparison to the metadata strategies in Table 3, the effectiveness of the publications cluster strategy lies between the combined metadata strategy and the journal strategy on the second place. However, the efficiencies of the combined metadata and journal strategies are about double times larger. The reason for the low efficiency of the publications cluster strategy is that for one author above 430,000 publications are analyzed on average. If we now consider the levels 2 or 3 of the MAG fields of study hierarchy (see Table 4), then - in contrast to our expectations - the effectiveness decreases slightly, as less publications are analyzed, but the efficiency nearly remains the same.

About 64%[15] of the MAG publications do not contain any keywords and hence also not a field of study. There are efforts to address this issue [45] by considering publications with a field of study for assigning a field of study to the remaining publications. We expect much better results for the publications cluster strategy for a data set with more publications assigned to fields of study.

4.3.2 X Cluster Strategies

X cluster strategies group authors, journals or conferences ($X \in \{$authors, journals, conferences$\}$). We analyze the different approaches for the journals given in Table 5 and show that the X cluster strategies can be used for searching for related publications.

We first analyze the citation relatedness between the considered journals. Table 6 contains the number of

[15] 78,3 millions publications

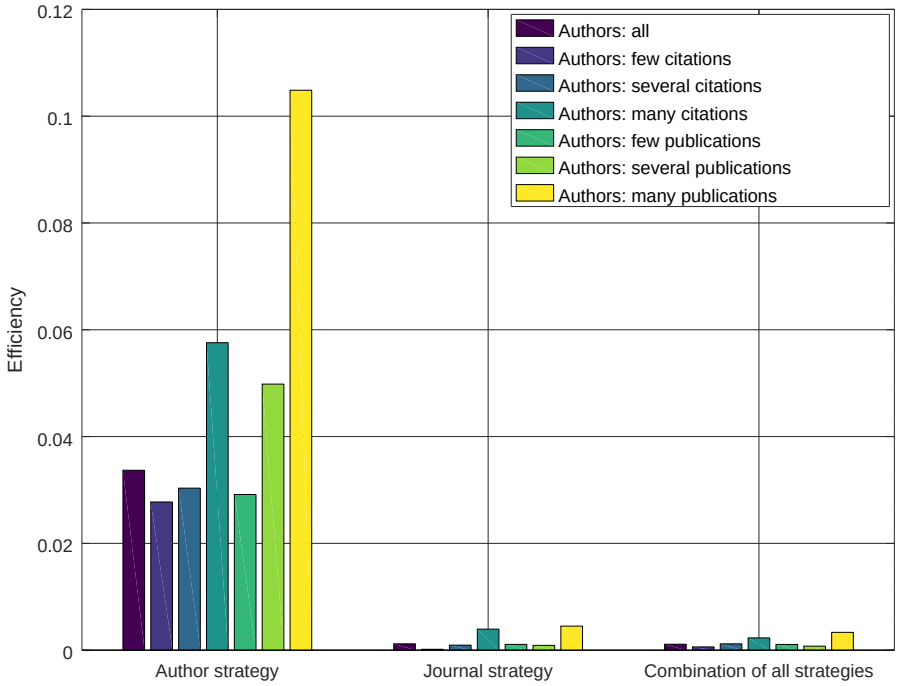


Figure 6: Efficiency of the metadata strategies and their combinations for different classes of authors. (Using the same classes as in Figure 5)

Table 5: Journals with their abbreviations, number of publications (Pub.) and the number of authors per publications. (The values of the last two columns are based on the data of MAGN.)

Journal	Abbr.	Pub.	$\frac{\text{Authors}}{\text{Pub.}}$
Expert Systems With Applications [17]	ESA	10.288	2,0
Knowledge Based Systems [20]	KBS	2.567	2,3
Int'l J. of Human-Computer Interaction [18]	HCI	885	2,4
Journal of Systems and Software [19]	JSS	3.900	2,1

Table 6: Citation relatedness between four journals. (Top: Number of absolute citations between journals. Bottom: Relative citation relatedness in percentage. Columns contain the number of references, e.g.: publications of ESA refer to 1,271 (i.e., 12.35 %) publications of KBS.)

references ↓	ESA	KBS	HCI	JSS
ESA	24,407	2,046	17	195
KBS	1,271	4,294	10	40
HCI	20	7	837	35
JSS	346	111	15	3,184
ESA	237.24 %	79.10 %	1.92 %	5.00 %
KBS	12.35 %	167.28 %	1.13 %	1.03 %
HCI	0.19 %	0.27 %	94.58 %	0.90 %
JSS	3.36 %	4.32 %	1.69 %	81.64 %

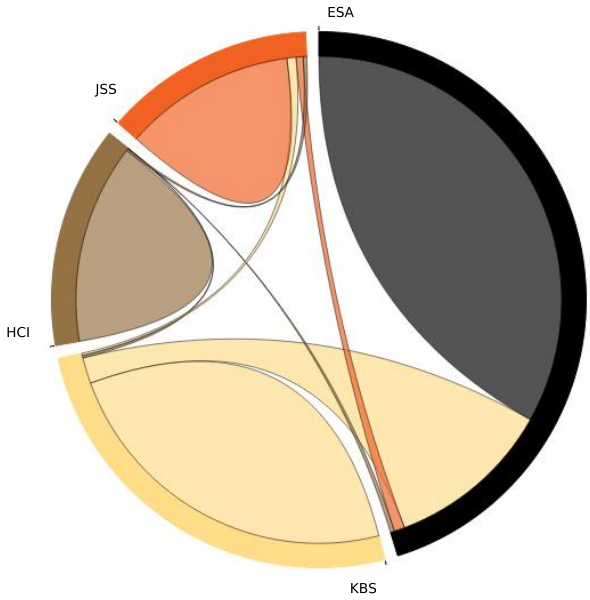


Figure 7: Graphical presentation of the relative citation relatedness between the considered journals. (There are references going from one journal to another, e.g.: ESA is much often cited from KBS than the other way around. Hence the connection becomes as thinner as closer the connection is to KBS.)

absolute citations among the journals. The references on the diagonal contain the references within the same journal. As expected the values on the diagonal are significantly higher, which corresponds to the journal strategy. Furthermore, we recognize a big number of citations between KBS and ESA (and the other way around). There are many references between ESA and JSS in both directions, too. We calculate the relative citation relatedness (see Equation 1 and bottom of Table 6) based on the absolute number of citations and the number of publications[16]. Sometimes the relatedness between two journals is very low, but all journals are related to each other. This is not surprising as all journals are computer science journals. Figure 7 presents the relative citation relatedness in a graphical way and illustrates well the differences in size. ESA is cited most often from itself, but also from KBS. A smaller portion is taken by JSS and HCI. ESA is more often cited than ESA cites other journals.

Table 7 contains the number of common authors. The common authors of each journal are obviously all authors of this journal (in the diagonal of the table). ESA and KBS have 1,032 authors in common, whereas KBS and HCI have only 16 authors in common. Table 8 presents the number of common fields of study, where we only consider the top 100 of the most occurring fields

[16] We assume that the publications in MAGN correspond to the number of extracted publications.

Table 7: Number of common authors of the journals. (The values in the diagonal are all authors of the corresponding journal.)

	ESA	KBS	HCI	JSS
ESA	20,948			
KBS	1,032	5,929		
HCI	34	16	2,112	
JSS	371	149	19	8,189

Table 8: Number of common fields of study of the journals by considering the 100 most occurring fields. (Hence the values on the diagonal correspond to the 100 most often occurring fields of study of the corresponding journal.)

	ESA	KBS	HCI	JSS
ESA	100			
KBS	57	100		
HCI	24	19	100	
JSS	22	16	19	100

of study. The number of common fields of study can be very large. Hence it may be more suitable to assign fields of study to journals instead of calculating the common fields of study. For this assignment, the more often occurring fields of study are typically considered. According to this idea (as discussed in Section 3.2) we also restrict the number of considered fields of study in this evaluation. All journals have at least 16 of their 100 most often occurring fields of study in common. The number of common fields of study is with 57 the largest for ESA and KBS.

Analysis of the results: We notice that all approaches (absolute/relative number of citations, common authors/fields of study) lead to similar results. Table 9 lists for each journal the most related journals (ordered descendant). For each approach we assume a higher relatedness as higher the calculated value (for the number of absolute/relative citations, of common authors or fields of study) is. The rankings for the absolute citations and the common authors are (with one exception) identical. If many common authors publish in two journals, then on the one hand these journals are related to each other by topic, and on the other hand the authors of these journals know the published articles of both journals. Hence the results are not nearly identical by accident. When comparing the results for the absolute and relative number of citations, then we recognize that quite often the places of two succeeding journals are switched.

Nerur et al. [41] analyze a citation network of 27 journals (which we described in Section 2.3.2). Among these journals are also ESA, KBS, HCI and JSS. According to their results, the relatedness between ESA and KBS is the largest. Afterwards, HCI and then JSS are most related to these journals. Based on our results according to all considered approaches, also ESA and KBS are most related to each other. According to the number of absolute citations JSS is more close to these journals in comparison to HCI. According to the relative number of citations it is one time HCI and the other time JSS.

The rankings according to common fields of study have the largest discrepancy to the three other approaches. Two journals can cover a similar area and hence have common fields of study. However, this does not automatically result in a close relatedness to each other. For example, Nerur et al. [41] discovered a clear separation between European and North American journals. Another reason is maybe that in our analysis we considered the 100 most often occurring fields of study for a journal, but did not weight the fields of study according to their frequency, i.e., we dealt with the first and the 100th most often occurring fields of study in the same way.

The consideration of the common authors is in comparison to the other strategies the most simple one. We neither need to assign fields of study to publications, nor the citations must be known. As more common authors are there, as higher is the probability of citations between them. Journals with many common authors are related by topic and the common authors are more likely familiar with the contributions of each other. Looking at the already found citations is presumably a relative precise estimation to find further citations between them, but enough already found citations are necessary for this approach. To consider the relative number of citations instead of the absolute number is reasonable in order to find citations faster. Let x be the calculated number (i.e., number of citations, authors etc.) between two journals. In no any strategy is the minimum nor the maximum value of x known, as long this value is not completely calculated (between all journals). Hence we cannot recognize how well the x value is for a concrete relatedness. However, if we consider fields of study, we can calculate a relatedness value between 0 and 100 %. The precondition for this calculation is that we express a relative number for the assignment to each field of study. For example, we may express for a journal that it is 100 % assigned to computer science on level 0, and 80 % to machine learning and 40 % to artificial intelligence on level 1.

5 Summary and Conclusions

In this paper we consider the problem of searching for citations. Searching for citations is a time-consuming task even if not a whole citation network of a large data set should be determined, but only the citations of one or several authors (e.g. of the same institute).

For the purpose of finding the citations of relative few publications (i.e., publications of one author or of an institute), we propose various search strategies, which allow to find citations earlier by ranking publications which are scheduled to be the next ones being analyzed for extraction of their references. We propose metadata strategies that rank the publications according to metadata information like author or venue, and combinations of these basic strategies. Furthermore, we propose cluster strategies, which group publications according to a precomputed relatedness of authors or venues. The precomputed relatedness may be calculated based on fields of study, already known citations or common authors.

We examine in a comprehensive experimental evaluation each of the strategies and discuss their effects

Table 9: The most related journals according to the 4 different approaches. (The journals are descendant ordered according to the previously calculated values.)

	ESA	KBS	HCI	JSS
Absolute Citations	1. KBS	1. ESA	1. JSS	1. ESA
	2. JSS	2. JSS	2. ESA	2. KBS
	3. HCI	3. HCI	3. KBS	3. HCI
Relative Citations	1. KBS	1. ESA	1. JSS	1. KBS
	2. JSS	2. HCI	2. KBS	2. ESA
	3. HCI	3. JSS	3. ESA	3. HCI
Common Authors	1. KBS	1. ESA	1. ESA	1. ESA
	2. JSS	2. JSS	2. JSS	2. KBS
	3. HCI	3. HCI	3. KBS	3. HCI
Common Fields of Study	1. KBS	1. ESA	1. ESA	1. ESA
	2. HCI	2. HCI	2. JSS/KBS	2. HCI
	3. JSS	3. JSS		3. KBS

in terms of effectiveness (how many publications are overall found) and efficiency (how many publications must be analyzed in relation to the found citations). In the experiments our best strategy finds about 75 % of the citations (for authors with not too few publications) by analyzing only a small proportion (1.54 %) of the overall publications.

REFERENCES

[1] B.-C. Bjork, A. Roos, and M. Lauri, "Scientific journal publishing: yearly volume and open access availability," *Information Research: An International Electronic Journal*, vol. 14, no. 1, 2009.

[2] Brown University, "FreeCite," http://freecite.library.brown.edu (accessed on 17.7.2017).

[3] C. Caragea, J. Wu, K. Williams, S. Das, M. Khabsa, P. Teregowda, and C. L. Giles, "Automatic identification of research articles from crawled documents," in *Proceedings of the Workshop: Web-Scale Classification: Classifying Big Data from the Web*, ser. WSDM 2014, New York, NY, USA, 2014.

[4] S.-B. Chang, K.-K. Lai, and S.-M. Chang, "Exploring technology diffusion and classification of business methods: Using the patent citation network," *Technological Forecasting and Social Change*, vol. 76, no. 1, pp. 107 – 117, 2009.

[5] P. Chen and S. Redner, "Community structure of the physical review citation network," *Journal of Informetrics*, vol. 4, no. 3, pp. 278 – 290, 2010.

[6] C. Clark and S. Divvala, "Looking beyond text: Extracting figures, tables, and captions from computer science paper," *AAAI, Workshop on Scholarly Big Data*, 2015.

[7] Cornell University Library, "arXiv API," http://arxiv.org/help/api/index (accessed on 17.7.2017).

[8] Cornell University Library, "arXiv API User's Manual," http://arxiv.org/help/api/user-manual (accessed on 17.7.2017).

[9] Cornell University Library, "arxiv bulk data access," https://arxiv.org/help/bulk_data (accessed on 17.7.2017).

[10] Cornell University Library, "arxiv help - open archives initiative," http://arxiv.org/help/oa/index (accessed on 17.7.2017).

[11] Cornell University Library, "General information about arxiv," http://arxiv.org/help/general (accessed on 17.7.2017).

[12] I. G. Councill, C. L. Giles, and M. yen Kan, "Parscit: An open-source crf reference string parsing package," in *International Language Resources and Evaluation*. European Language Resources Association, 2008.

[13] S. Dawson, D. Gašević, G. Siemens, and S. Joksimovic, "Current state and future trends: A citation network analysis of the learning analytics field," in *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, ser. LAK '14, 2014, pp. 231–240.

[14] DBLP, "DBLP XML download," http://dblp.uni-trier.de/xml/ (accessed on 17.07.2017).

[15] DBLP, "Statistics - Records in DBLP," http://dblp.uni-trier.de/statistics/recordsindblp.html (accessed on 17.7.2017).

[16] DBLP, "What is DBLP?" http://dblp.uni-trier.de/faq/What+is+dblp.html (accessed on 17.7.2017).

[17] Elsevier B.V., "Expert systems with applications," http://www.journals.elsevier.com/expert-systems-with-applications (accessed on 17.7.2017).

[18] Elsevier B.V., "International journal of human-computer studies," http://www.journals.elsevier.com/international-journal-of-human-computer-studies/ (accessed on 17.7.2017).

[19] Elsevier B.V., "Journal of systems and software," http://www.journals.elsevier.com/journal-of-systems-and-software/ (accessed on 17.7.2017).

[20] Elsevier B.V., "Knowledge-based systems," http://www.journals.elsevier.com/knowledge-based-systems/ (accessed on 17.7.2017).

[21] P. Érdi, K. Makovi, Z. Somogyvári, K. Strandburg, J. Tobochnik, P. Volf, and L. Zalányi, "Prediction of emerging technologies based on analysis of the us patent citation network," *Scientometrics*, vol. 95, no. 1, pp. 225–242, 2012.

[22] Google Inc., "Google scholar," https://scholar.google.de/intl/en/scholar/about.html (accessed on 17.7.2017).

[23] GROBID, "Grobid documentation - introduction," http://grobid.readthedocs.org/en/latest/Introduction/ (accessed on 17.7.2017).

[24] H. Han, C. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. Fox, "Automatic document metadata extraction using support vector machines," in *Digital Libraries, 2003. Proceedings. 2003 Joint Conference on*, 2003, pp. 37–48.

[25] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[26] Y. Kajikawa, J. Ohno, Y. Takeda, K. Matsushima, and H. Komiyama, "Creating an academic landscape of sustainability science: an analysis of the citation network," *Sustainability Science*, vol. 2, no. 2, pp. 221–231, 2007.

[27] M.-Y. Kan, I. G. Councill, C. L. Giles, M.-T. Luong, and H. N. H. Do, "ParsCit: An open-source CRF Reference String and Logical Document Structure Parsing Package," https://github.com/knmnyn/ParsCit (accessed on 17.7.2017), 2016.

[28] M. Khabsa and C. L. Giles, "The number of scholarly documents on the public web," *PLoS ONE*, vol. 9, no. 5, p. e93949, 2014.

[29] A. Kusserow and S. Groppe, "Getting indexed by bibliographic databases in the area of computer science," *Open Journal of Web Technologies (OJWT)*, vol. 1, no. 2, pp. 10–27, 2014. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:101:1-201705291343

[30] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01, 2001, pp. 282–289.

[31] C. Lagoze, H. V. de Sompel, M. Nelson, and S. Warner, "The open archives initiative protocol for metadata harvesting - v.2.0," https://www.openarchives.org/OAI/openarchivesprotocol.html (accessed on 17.7.2017), Jan. 2015.

[32] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing," *IEEE Computer*, vol. 32, no. 6, pp. 67–71, 1999.

[33] M. Ley, "Dblp: Some lessons learned," *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1493–1500, Aug. 2009.

[34] H. Li, I. Councill, W.-C. Lee, and C. L. Giles, "Citeseerx: An architecture and web service design for an academic document search engine," in *Proceedings of the 15th International Conference on World Wide Web*, ser. WWW '06, New York, NY, USA, 2006, pp. 883–884.

[35] P. Lopez, "Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications," in *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, ser. ECDL'09, 2009, pp. 473–474.

[36] Microsoft Corporation, "Academic knowledge api documentation," https://www.microsoft.com/cognitive-services/en-us/academic-knowledge-api/documentation/overview (accessed on 7.04.2016).

[37] Microsoft Corporation, "Cognitive services - preview pricing," https://www.microsoft.com/cognitive-services/en-us/pricing (accessed on 7.04.2016).

[38] Microsoft Corporation, "Microsoft academic search - help center," http://academic.research.

microsoft.com/About/Help.htm#4 (accessed on 18.11.2015).

[39] Microsoft Corporation, "Microsoft academic search api v1.3," http://academic.research.microsoft.com/about/Microsoft%20Academic%20Search%20API%20User%20Manual.pdf (accessed on 18.11.2015), May 2012.

[40] Microsoft Research, "Microsoft academic graph," http://research.microsoft.com/en-us/projects/mag/ (accessed on 7.04.2016).

[41] S. Nerur, R. Sikora, G. Mangalaraj, and V. Balijepally, "Assessing the relative influence of journals in a citation network," *Commun. ACM*, vol. 48, no. 11, pp. 71–74, Nov. 2005.

[42] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, p. 026113, 2004.

[43] M. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, 2004.

[44] E. Oberesch and S. Groppe, "The mf-index: A citation-based multiple factor index to evaluate and compare the output of scientists," *Open Journal of Web Technologies (OJWT)*, vol. 4, no. 1, pp. 1–32, 2017. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:101:1-2017070914565

[45] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion, 2015, pp. 243–246.

[46] A. Spitz and M. Gertz, "Breaking the news: Extracting the sparse citation network backbone of online news articles," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ser. ASONAM '15, 2015, pp. 274–279.

[47] TEI Consortium, "TEI: P5 Guidelines," http://www.tei-c.org/Guidelines/P5/ (accessed on 17.7.2017).

[48] The Open Archives Initiative, "About OAI," https://www.openarchives.org/OAI/OAI-organization.php (accessed on 17.7.2017).

[49] The Open Archives Initiative, "Registered Data Providers," http://www.openarchives.org/Register/BrowseSites (accessed on 17.7.2017).

[50] The Pennsylvania State University, "About CiteSeerX," http://csxstatic.ist.psu.edu/about (accessed on 17.7.2017).

[51] The Pennsylvania State University, "Citeseerx data," http://csxstatic.ist.psu.edu/about/data (accessed on 17.7.2017).

[52] Thomson Reuters, "Web of science fact sheet," http://thomsonreuters.com/content/dam/openweb/documents/pdf/scholarly-scientific-research/fact-sheet/wos-next-gen-brochure.pdf (accessed on 17.7.2017), 2014.

[53] S. Tuarob, S. Bhatia, P. Mitra, and C. L. Giles, "Automatic detection of pseudocodes in scholarly documents using machine learning," in *2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, August 25-28, 2013*, 2013, pp. 738–742.

[54] United States Patent and Trademark Office, "Us patent and trademark office full-text and image database," http://patft.uspto.gov (accessed on 17.7.2017).

[55] R. Van Noorden, "The true cost of science publishing," *Nature*, vol. 495, no. 7442, pp. 426–429, 2013.

[56] K. Williams, L. Li, M. Khabsa, J. Wu, P. C. Shih, and C. L. Giles, "A web service for scholarly big data information extraction," in *Web Services (ICWS), 2014 IEEE International Conference on*. IEEE, 2014, pp. 105–112.

[57] J. Wu, J. Killian, H. Yang, K. Williams, S. R. Choudhury, S. Tuarob, C. Caragea, and C. L. Giles, "Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search," in *Proceedings of the 8th International Conference on Knowledge Capture*, ser. K-CAP 2015, 2015, pp. 13:1–13:8.

[58] J. Wu, P. Teregowda, M. Khabsa, S. Carman, D. Jordan, J. San Pedro Wandelmer, X. Lu, P. Mitra, and C. L. Giles, "Web crawler middleware for search engine digital libraries: A case study for citeseerx," in *Proceedings of the Twelfth International Workshop on Web Information and Data Management*, ser. WIDM '12, 2012, pp. 57–64.

[59] J. Wu, P. Teregowda, J. P. F. Ramírez, P. Mitra, S. Zheng, and C. L. Giles, "The evolution of a crawling strategy for an academic document search engine: Whitelists and blacklists," in *Proceedings of the 4th Annual ACM Web Science Conference*, ser. WebSci '12, 2012, pp. 340–343.

AUTHOR BIOGRAPHIES

Fabian Rosenthal studied Computer Science at the University of Lbeck and the University of Oslo. He earned his bachelor degree in 2013 with a thesis about accuracy-driven time synchronization of sensor nodes in distributed networks. In 2016 he earned his master degree from the University of Lbeck. His master thesis is about a web service for purposeful searching for citations of scholarly publications. Since then he works as a software engineer.

Sven Groppe earned his diploma degree in Informatik (Computer Science) in 2002 and his Doctor degree in 2005 from the University of Paderborn. He earned his habilitation degree in 2011 from the University of Lübeck. He worked in the European projects B2B-ECOM, MEMPHIS, ASG and TripCom. He was a member of the DAWG W3C Working Group, which developed SPARQL. He was the project leader of the DFG project LUPOSDATE, an open-source Semantic Web database, and of two research projects in the area of FPGA acceleration of relational and Semantic Web databases. He is also the chair of the Semantic Big Data workshop series, which is affiliated with the ACM SIGMOD conference (so far in 2016 and 2017), and of the Very Large Internet of Things workshop in conjunction with the VLDB conference in 2017. His research interests include databases, Semantic Web, query and rule processing and optimization, Cloud Computing, peer-to-peer (P2P) networks, Internet of Things, data visualization and visual query languages.