Scientometrics (2021) 126:9779–9809
https://doi.org/10.1007/s11192-021-04103-w

The explanatory power of citations: a new approach to unpacking impact in science

Matthias Sebastian Rüdiger[1] · David Antons[1] · Torsten-Oliver Salge[1]

Received: 16 June 2020 / Accepted: 8 July 2021 / Published online: 5 August 2021

Abstract

Citation analysis has been applied to map the landscape of scientific disciplines and to assess the impact of publications. However, it is limited in that it assumes all citations to be of equal weight. Doing away with this assumption could make such studies even more insightful. Current developments in this regard focus on the evaluation of the syntactic and semantic qualities of the text that surrounds citations. Still lacking, however, are computational techniques to unpack the thematic context in which citations appear. It is against this backdrop that we propose a text clustering approach to derive contextual aspects of individual citations and the relationship between cited and citing work in an automated and scalable fashion. The method reveals a focal publication's absorption and use within the scientific community. It can also facilitate impact assessments at all levels. In addition to analyzing individual publications, the method can also be extended to creating impact profiles for authors, institutions, disciplines, and regions. We illustrate our results based on a large corpus of full-text articles from the field of Information systems (IS) with the help of exemplary visualizations. In addition, we provide a case study, the scientific impact of the Technology acceptance model. This way, we not only show the usefulness of our method in comparison to existing techniques but also enhance the understanding of the field by providing an in-depth analysis of the absorption of a key IS theoretical base.

Keywords Citation analysis · Content-based citation analysis · Topic modeling · Information systems

✉ Matthias Sebastian Rüdiger
ruediger@time.rwth-aachen.de

David Antons
antons@time.rwth-aachen.de

Torsten-Oliver Salge
salge@time.rwth-aachen.de

[1] Institute for Technology and Innovation Management, RWTH-Aachen University, Aachen, Germany

Introduction

Used as a proxy for the relevance of academic work, citation counts are among the most frequently—employed measures of impact in science. Publications with higher citation counts are deemed more influential than those with lower ones (Cronin, 1984). Citation counts unclutter the mass of publications, separating the wheat from the chaff. Additionally, they form the basis for the field of citation analysis, in which bibliographic references of research publications are used to form networks that can be explored using graph—analytical methods to uncover hidden relations and the flow of knowledge (De Bellis, 2009). Frequently—used techniques include author networks, co-author networks, co-citation networks, and bibliographic coupling (Zupic & Čater, 2015). For several decades, citation analysis has been used to map the landscape of scientific disciplines, to observe knowledge transfer across disciplinary boundaries, and to assess the impact of publications (Gläser et al., 2017). Traditional analyses based on citation counts, however, provide only a rather shallow image of the scientific landscape (Moravcsik, 1973; Zhu et al., 2015). Most importantly, they do not differentiate between the various types of citations. Yet citations differ based on the citing author's motivation, the thematic context of a citation, or a citation's argumentative relation to its subject (Amsterdamska & Leydesdorff, 1989; Cronin, 1984). Typically, all citations are assumed to be of equal weight. While this simplification might be useful for some purposes, it prevents researchers from gaining a deeper and more subtle understanding of the scientific impact generated by an individual paper, an author, a department, or an entire field. Traditional citation analyses are limited by their inability to discern the level of intellectual indebtedness (Ding et al., 2014; Phelan, 1999).

Unpacking the precise nature of citations—and doing so at scale using automated approaches—is hence important to make impact metrics more meaningful for performance assessments and more informative for the research community (Angrosh et al., 2010). Some researchers have begun to address this question by breaking down citation counts by journal and scientific field. Others have focused on the syntactic features of citations, the argumentative context (i.e. whether the author's reference to a previous work is appreciative or disparaging), and the author's motives (Ding et al., 2014; Hernández-Alvarez & Gomez, 2016). While certainly valuable, these efforts fall short of helping to understand the thematic contribution to the citing work and hence its impact. For example, some large survey-based studies may be cited for a small methodological refinement, other studies may be cited for historical reasons or because a term was coined in them, and still others to serve as cautionary tales. Existing methodological approaches are ill-equipped to reveal the thematic context of citations. While manual coding of citations using qualitative techniques might seem a solution at first glance, such an approach is limited in its practical value, given the large number of citations typically examined. Scalable methods to uncover for the context of citations are thus needed.

The present study breaks new ground in this regard. We develop and test a novel automated method to extract the context of citations, rather than simply their numbers or locations. We aim to uncover the thematic contributions of academic publications through their absorption and actual use by the scientific community. For this purpose, we employ a combination of state-of-the-art techniques from computational linguistics and data mining in order to analyze the textual environment of every citation in a given set of publications. As part of this process, we identify and extract the text environments of all citations to the focal publication. We then apply text-mining techniques to compute the set of topics a focal publication is being cited for. This topic map provides a unique impact profile for one or more publications that depicts their direct impact within a defined scientific community.

To demonstrate our method, we apply it to a document collection known as the AIS Senior Scholars' Basket, comprising eight leading journals from the field of Information systems (IS). Information systems (IS) outstandingly benefits from such efforts, as the discipline is still searching for its place in the scientific landscape and a concise identity (Grover, 2012). Developments in IS call for regular assessment of its intellectual structure, its boundaries, and its relation to its reference disciplines. Impact plays a key role in this assessment. Our study, therefore, also provides novel insights into the context of citations within the IS field. Ultimately, the proposed method may be used to simplify impact assessment in science in general. Authors of literature reviews, for instance, will be able to assess with greater ease how certain articles impact the content of others. Search engines for publication data may automatically inform users about what publications are cited for instead of simply displaying citation counts. Further, the analysis we propose can be combined straightforwardly with metadata such as publication dates to plot knowledge absorption curves over time. A more far-reaching application may consist of comparing and contrasting a publication's computed impact profile depicting its absorption with its positioning as reflected in the set of keywords the authors select for it. It could also be used for the evaluation of funding initiatives in academia, by comparing generated citation profiles before and after a funding initiative. Using our novel method, researchers can thus not only compute impact profiles, but also use them as inputs to subsequent analyses to unearth novel patterns in knowledge diffusion within and across scientific fields.

In the upcoming section, we briefly outline the evolution of citation analysis. We then explain our methodological approach in detail. We showcase the feasibility and value of our approach with exemplary citation profiles based on data from the IS field. Subsequently, we use our approach to trace the impact of the Technology acceptance model (TAM), a key concept in IS research, within the IS literature. Hereafter, we contrast the analytical approach presented in this article with text-based methods for mapping scientific disciplines. We conclude by summarizing our findings and sketching ideas for further research.

Evolution of citation analysis

Not all citations are based on the same type of intellectual involvement with the cited article. While some citations are listed in a bibliography because the original sources truly influenced the present work, others basically constitute name—dropping without any actual impact, and still others are noted purely with disapproval. Concerns about this situation have led to two research streams in citation analysis. The first stream is, above all, theory-driven, whereas the second stream is data-driven and does not rely on any theoretical framework but draws heavily from computer science and linguistics. An overview of different theories of citation and open research questions can be found in the overview article by Leydesdorff and Milojevic (2012).

The first research stream is motivated by the need to understand citation motivations and intentions. It is represented by two theories: the normative theory of citing and the social constructivist view (Cronin, 1984). The former theory argues that citations are symbolic payments of intellectual debts, a consequence of norms in science through which scholars are expected to acknowledge the use of others researchers' work (Merton, 1973). Hence, citations indicate true peer recognition. The social constructivist view, by contrast, claims that scientific knowledge is a social construct and that citations are used to persuade and to reward, and sometimes just because the authors believe their target audience considers them authoritative. Empirical evidence for both theories can be found in Baldi (1998). Whichever of these two theories is in play, analyses focusing on intentions and motivations of citing authors are usually conducted by carrying out surveys and interviews (Brooks, 1985; Vinkler, 1987). A more recent example investigating citing behavior in the discipline of communications via surveys is Case and Higgins (2000).

The second type of research enhancing traditional citation analysis is called content-based citation analysis. It does not focus on theoretical grounding but on the texts of the articles that contain the citations. The present study aims to contribute to this set of methods. These methods may be further subdivided into those that consider the syntactic level of the article texts and those that consider the semantic level. In syntactic text analysis, citations are differentiated based on structural features in the texts where they appear. Usually, the positions of the citations in the focal text are considered, or else the analysis makes distinctions between citations that are recurring and non-recurring, between those that are direct and indirect, and between those that are individual and block, that is, citations that mention several papers simultaneously. Voos and Dagaev (1976) analyzed the value of each citation by assessing its location within a document. The authors concluded that citation evaluation should be based on citation location and frequency within a document. According to Herlach (1978), citations have a greater contribution to the citing article if they are mentioned in the Introduction or Literature Review section and mentioned again in the Methodology or Discussion section. Later, Maričić et al. (1998) conducted an analysis based on the location of references in more than 350 papers. Hu et al. (2013) examined and visualized how citations are distributed across sections in papers published in the Journal of Informetrics, Ding et al. (2013) investigated where references to the most frequently-cited articles are located in a publication, and Boyack et al. (2018) investigated multiple properties of in-text citations such as reference age.

Semantic text analyses focus on the argumentative context or discourse structure of citations, that is, the semantic relation between the cited document and the citing article. Citations can be used to motivate and justify, to relate contributions to previous work, or to contrast differing positions. Many studies on citation contexts first develop a categorization scheme and then manually or semi-automatically classify the citations in a given set of publications. In one of the earliest examples of analysis on a semantic level, Garfield (1965) presented fifteen different reasons why authors make citations. Moravcsik and Murugesan (1975) analyzed the function and quality of citations by examining their surroundings and distinguished whether the citing paper was extending previous ideas, proposing a new viewpoint, or denying or confirming the cited work. Chubin and Moitra (1975) classified four types of confirmative citations and two types of negating citations. These early studies were conducted manually through labor-intensive and small-size content analysis. Thus, the results were often not generalizable due to the limited sample size. Later studies applied semi-automated or automated techniques to enhance citation analysis. Teufel et al. (2006) developed a general scheme to automatically classify a citation's argumentative relation to its subject; most of the categories in them correspond to the linguistic use of conjunctions and/or discourse markers. More recently, studies have started to use machine learning algorithms to determine citation functions, attempting to overcome the limitations of previous studies. Jha (2017) summarized the ongoing research in citation analysis powered by natural language processing to automatically derive citation

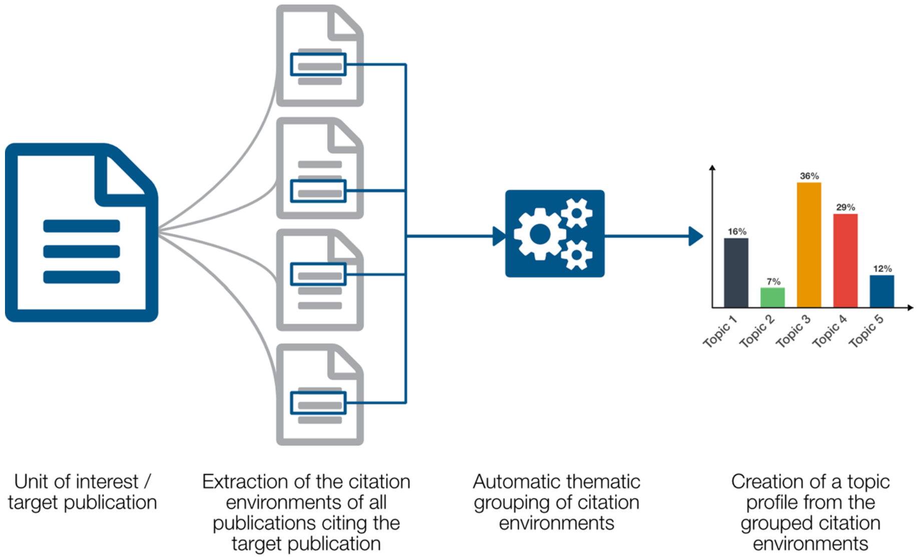
16%
7%
36%
29%
12%
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Unit of interest / target publication
Extraction of the citation environments of all publications citing the target publication
Automatic thematic grouping of citation environments
Creation of a topic profile from the grouped citation environments

Fig. 1 Outline of the overall procedure to unpack the thematic context of citations. (Color figure online)

classifications and polarity. Tahamtan and Bornmann (2019) gave a comprehensive update on the review of Bornmann's and Daniel's (2008) work on authors' reasons for citing previous works and on linguistic features that needed to be analyzed in order to properly assess the functions of citations.

All in all, researchers employed different approaches to enhance citation analysis by examining the citing articles in greater detail. Here, both syntactic and semantic text features are extracted and used. Nearly all of the works dealing with semantic content-based citation analysis focus on the argumentative context of citations. Only a few of the more recent studies attempt to extract topics from citation environments to improve citation graphs, e.g. Liu et al. (2013). However, a general approach to unpack the actual thematic context of citations is still missing.

Methodology

We propose a new method extending content-based citation analysis at the semantic level. Instead of assessing the nature of citations itself, we explore the thematic context in which citations occur. Our objective is to automatically derive meaningful thematic categories for research articles from a corpus of full-text manuscripts citing those articles to reveal their use and absorption within the scientific community. These categories are not predefined but are learned by a topic modeling algorithm. This automated procedure has significant advantages over previous manual techniques. In contrast to manually coding a large number of citations, a topic model is not limited in information processing capacities and hence may not suffer from the challenges that occur when multiple coders are required to classify large amounts of text. It might also obtain more objective results than manual analyses.

In the following, we give a short overview of the procedure behind our approach outlined in Fig. 1. First, we define the unit of analysis, i.e. the publication or publications whose impact is to be measured. In a second step, we extract the textual surroundings of citations that refer to our unit of analysis—which can be many things, such as an article, all the articles by a particular author, or even all the articles by a faculty or journal or discipline. Third, we preprocess the citations' environments and use a text clustering technique to group them. The resulting sets of text snippets represent the topics the publication of interest is connected with. In the end, the result is visualized.

Data acquisition

The selection of a suitable data set for the proposed undertaking represents a major challenge. To determine a complete citation profile for one publication, all documents that ever cited that publication are required in full text. As our approach can only work to its full potential for publications that have already been cited many times, this requires obtaining many publications in full-text, usually from many different publishers. Due to the licensing restrictions in today's publishing ecosystem, this poses a considerable hurdle. As a consequence, we consider the reception of an article or author within a particular research area only, in our case Information systems (IS). We focus our analysis on the full-text articles of the eight leading journals in IS research, known as the AIS Seniors Scholars' Basket.[1] This focus limits our citation profiles to pure IS journals, thus excluding other high-ranking multidisciplinary scientific journals related to IS. We believe, however, that the selected journals represent a comprehensive collection of work for the evaluation of IS research and constitute an appropriate choice for our purpose, despite the limitations. Accordingly, we have downloaded the complete set of full-text publications of all these journals from their first issue up to and including 2019 as PDF files. The resulting dataset forms the basis for the validation of our approach.

Data preparation

For the automatic extraction of citation environments, all the publications must be in a machine-readable format. We applied Optical character recognition (OCR) using the Tesseract[2] open-source library to all the pdf files that were non-machine-readable, which included most of them that were published before the year 2000. We also corrected some PDF files for alignment and contrast before applying OCR using OCRmyPDF,[3] These corrections were especially necessary for older publications only available as scans. We then applied Grobid,[4] an open-source Java library, to export all information contained in the pdf files into a structured format that would allow further automatic processing. Grobid uses machine learning to extract information from pdf files. It is especially suited for scientific publications, as well as patents, and comes with several pre-trained models based on Conditional random fields (CRF). The resulting Text encoding initiative (TEI) format is probably the most comprehensive encoding standard for machine-readable texts. It allows precise text annotations while maintaining readability. With it, we can not only extract the contents of all articles in a structured way but also obtain the information contained in their

[1] https://aisnet.org/page/SeniorScholarBasket.

[2] https://github.com/tesseract-ocr/tesseract.

[3] https://ocrmypdf.readthedocs.io.

[4] https://github.com/kermitt2/grobid.

bibliographies. The reference information from the bibliographies is crucial for the correct identification of references in the article texts.

Reference identification

Parsing and identifying the reference sections constitutes a difficult task due to varying formatting styles of the reference section, mistakes in spelling, and errors introduced during the optical character recognition process. As Pasula et al. (2003) show, identity uncertainty (the correct recognition and assignment of names to the corresponding authors) and citation matching (the attribution of bibliography entries to the in-text references) are still major issues in research. The authors propose a probabilistic method to enhance the matching procedure. We followed this advice by utilizing the aforementioned Grobid library and its pre-trained models based on conditional random fields. According to Tkaczyk et al. (2018), it is the best-performing out-of-the-box tool for this task. The Grobid library was configured to obtain reference information from crossref.org to increase the data quality of the detected references. Additionally, we integrated a supplementary source of information to assist the process. We downloaded all available metadata about the journals from Thomson Reuters Web of Science. The metadata download contained a list of cited references in a structured format. The metadata enabled us to ensure the correctness of the reference extraction process.

Citation environment extraction

The Grobid library enabled us to identify direct and indirect citations including their link to the bibliography. Because citations in the text do not provide a clear assignment to a publication, they must first be associated with the appropriate entry in the bibliography. In combination, we were thus capable of clearly identifying citations in the text. Hence, we were also able to identify the corresponding citation environments. As outlined above, the citation environments were taken as representatives of the thematic embedding of the citations. This raises the question of the optimal text length to be extracted from the surroundings of each citation. Ritchie et al. (2008) compare citation contexts of different lengths to enhance information retrieval. They assess different fixed or variable window sizes, that is, a number of words on either side of a citation, up to a fixed limit of words, or based on the number of sentences. They conclude that sentence-based contexts are more effective than window sizes of fixed length. Additionally, they found that using neighbor citations to delimit a citation's environment was not helpful. In our approach, we used a variable window size, taking as input the entire paragraph in which a citation is located. This way, we followed Ritchie et al.'s (2008) recommendations based on the assumption that words likely to describe the use of the cited publication occur close to the citation, whereas words further away are less likely to describe it. Hence, the sentences surrounding a citation are a good approximation of the citation's descriptive terms. In our case, we had the additional requirement that we needed to obtain enough text to feed the data mining algorithm.

Text preprocessing

After obtaining the extracted citation environments, it was necessary to convert the text snippets to a format appropriate for the subsequent text mining procedure. Our overall goal for the text preprocessing is, first and foremost, to improve the interpretability of the impact profiles so that they provide a quick overview of a publication's absorption in the scientific community. To this end, we conducted a large number of experiments to assess the effects of different combinations of preprocessing techniques. Based on the results of our experiments, and in line with the objectives of our method, we determined the optimal combination of text preprocessing techniques. They were as follows.

First, we normalized the texts to an identical character set to ensure the absence of noisy characters and artifacts. This is especially necessary since we employed OCR on some of the articles to obtain machine-readable texts, and OCR is prone to errors, such as when it confuses dirt on scanned sheets of paper with accent marks. Then, we converted all words to lowercase. We excluded terms written all in capital letters and a length of two or more characters from the lowercase conversion so as not to confuse acronyms with words, such as "IS" (for Information System) with "is." After that, we converted the stream of words to a set of meaningful tokens using a tokenization algorithm. To this end, we employed a model-based sentence boundary disambiguation algorithm for the English language. During the process, every recognized sentence is separately treated by a rule-based tokenizer. The raw text is first split using whitespace characters. Then the tokenizer processes the text from left to right and checks whether a substring matches an exception rule or if a prefix, suffix, or infix can be split off. Further details on the process can be obtained on the website of the natural language processing library[5] we employed. After acquiring a stream of tokens, we applied a filtering method to reduce the size of the dictionary and thus the dimensionality of the texts. We opted to reduce the vocabulary based on part-of-speech filtering by removing all but nouns, proper nouns, and multiword expressions, that is, idioms or compound nouns. Removing all words without meaning taken in isolation helped to sharpen the contours of the topics and to stabilize the text-clustering algorithm. It also facilitated the quick interpretation of the resulting word lists by the analyst. Additionally, we filtered out all tokens contained in less than 2.5% and more than 97.5% of the documents to further eliminate irrelevant terms. Since we are less interested in the words themselves than in their underlying meaning, we grouped inflected or derived forms of a word so they could be analyzed as a single item. This was done using a lemmatization technique available from the natural language processing library we used. Last, we utilized compound word identification to tease out possible errors introduced by the tokenization heuristics. In general, terms may occur as words or compound words. By compound words, we understand terms written as two or more separate words, such as "Information Systems." Those terms can be proper names, including brand names, or they can be phrasemes. However, without oversight, a tokenization procedure may split up terms into multiple tokens, losing their meaning. We utilized a dictionary-based approach for compound word identification using Wikipedia.[6] Extracting all Wikipedia titles and then filtering out all non-multi-word titles yielded a comprehensive list of known multi-word expressions. The wordlist we obtained from this process then served as a look-up table and was used to replace sets of tokens with the respective multi-word token for phrases comprising up to six words.

Following Salton et al. (1975), we transformed each text snippet into a numerical feature vector. In this process, each element of the vector represents a term: in our case, a word or multi-word expression. If a term occurs in a document, its value in the corresponding vector is non-zero. The size of the vectors is defined by the number of tokens in the dictionary.

[5] See https://spacy.io for more details.

[6] Wikipedia database dumps available from: https://dumps.wikimedia.org/enwiki/latest/.

Thus, texts are described based on the set of words contained in them and form a so-called document-term matrix. Ultimately, through this process, we obtained a numerical representation of the citation environments to supply the topic modeling algorithm.

Text mining procedure

After extracting and preprocessing the citation environments, we applied the actual text mining procedure to form meaningful groups with similar thematic contexts. Because these clusters needed to be learned automatically from the given texts without predefined sets of categories, we used an unsupervised method (Han et al., 2011). This kind of data mining method requires no training at all, in contrast to dictionary-based and supervised learning methods. Unsupervised learning methods automatically extract statistical and semantic features of the texts to estimate a set of categories and assign texts to those categories simultaneously. As such, they can identify categories previously unknown. In this context of grouping textual data, they are also referred to as topic modeling algorithms. When discovering topics from a text corpus, topic modeling algorithms consider each document from a document collection to be a combination of multiple topics. A topic is defined as a distribution over a fixed vocabulary. The distribution describes the probabilities of generating various words. Intuitively, the words that are characteristic of a specific topic will have a high probability. Words without special relevance, such as function words, will have roughly even probability between all distributions representing the topics. A word may occur in several topics with different probabilities and with different adjacent terms. Each text is characterized by a particular topic distribution.

In our setup, we employed Non-negative matrix factorization (NMF) to form meaningful groups of the extracted citation environments (Lee & Seung, 2001). NMF is a matrix factorization method similar to Singular value decomposition (SVD) that factorizes any matrix into the product of three separate matrices. NMF adds one additional constraint for the decomposition: the decomposed matrices must consist of only non-negative values. NMF has become popular due to its inherent clustering property and thus its ability to automatically extract sparse and easily interpretable factors. The constrained matrix factorization cannot be calculated exactly; it is commonly approximated numerically and formulated as a constrained non-linear optimization problem. The algorithm generates topics defined by wordlists representative of them, as well as a list of relationships between text and topics in the form of probabilities. The wordlists are also called topic descriptors.

We conducted a comprehensive comparison of text clustering algorithms such as Latent sematic analysis (LSA), Probabilistic latent sematic analysis (PLSA), Latent Dirichlet allocation (LDA), NMF, and SVD as well as various evaluation metrics. Based on the results, we chose NMF over other frequently—employed topic modeling algorithms like Latent Dirichlet allocation (LDA). In our study, NMF outperformed all other tested clustering procedures for datasets with only a small number of clusters. It was better at both speed and result scattering and showed comparable clustering accuracy (Chen et al., 2019). Hence, we did not require a separate testing procedure to determine the optimal run with the same parameter settings, in contrast to LDA which would require multiple executions due to the result scattering. As our approach requires high speed in several use cases, but usually only a small number of clusters have to be extracted, NMF was the topic modeling algorithm of choice. As we operated within an unsupervised setting, the number of clusters *k* in a dataset, i.e. the number of topics, was not known, and had to be determined in the process. Therefore, we needed an evaluation criterion to determine the optimal number of clusters.

The primary objective of our newly—developed instrument is to facilitate the identification of a publication's absorption in the scientific community. Consequently, we favor an evaluation metric that focuses on topic interpretability. We tested several topic modeling evaluation metrics regarding their relation to optimal clustering results. We found that for certain bandwidths of k, some of the metrics did correlate with an optimal clustering result definition. Among them was the topic coherence measure C_{NPMI}, which performed well on datasets with a small number of clusters (Bouma, 2009). Such a word-based topic coherence measure scores a single topic by measuring the degree of semantic similarity between high scoring words in the topic. C_{NPMI} uses the normalized pointwise mutual information (NPMI) to measure the semantic similarity of word pairs based on a reference corpus. The calculated score for each word pair within a topic is then aggregated for each topic and finally averaged across all the topics. This way, they can be used to assess the overall quality of a topic model. Coherence measures help to distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference. In our setup, C_{NPMI} made it possible to automatically optimize the number of topics extracted from a set of citation environments by computing a clustering using NMF for $k=2$, $k=3$, …, $k=20$ clusters. The clustering yielding the highest C_{NPMI} value was then deemed to be the most appropriate one. In the end, we obtained a clustering of the citation environments of our unit of analysis. Each cluster represented a topic and was characterized by a list of representative words, also called topic descriptors.

Result visualization

After the clustering procedure, several topic descriptors were obtained. Each topic descriptor contained several terms representative for the topic. The lists were then processed further to generate a graphical depiction of the impact profile. We chose a tree-map-like visualization (Shneiderman, 1992). Tree-map visualizations were originally developed to depict hierarchical dependencies. However, they also provide an intuitive picture of impact profiles. In this profile, each of the colored areas designates the impact the one or more analyzed publications on a specific topic. The larger its area relative to the other areas, the more impacted the topic is by the publication(s). The topics themselves are specified by lists of characteristic words shown in the colored areas, arranged according to their importance. It is important to stress that the resulting impact profiles are created in a fully automated manner. These visualizations allow scholars to quickly review the impact of one or more publications of interest. The automatic unpacking of thematic impact in this way greatly simplifies the assessment of the relevance of a paper for one's own research. It hence will serve as a useful feature of databases for scientific publications or search engines. Accordingly, tree-map-like visualizations of citation profiles may be integrated into lists such as those of search results for each article. Such visualizations could support the assessment of candidates in academic hiring committees or portray the impact of entire institutions at a glance. A detailed illustration of the entire process is shown in Fig. 2.

Results

Our corpus included 7,382 full-text articles from the AIS Seniors Scholars' Basket, nearly all the articles published in the eight journals from their first issue up to and including 2019. We filtered out some texts such as book reviews, errata, and any supplementary

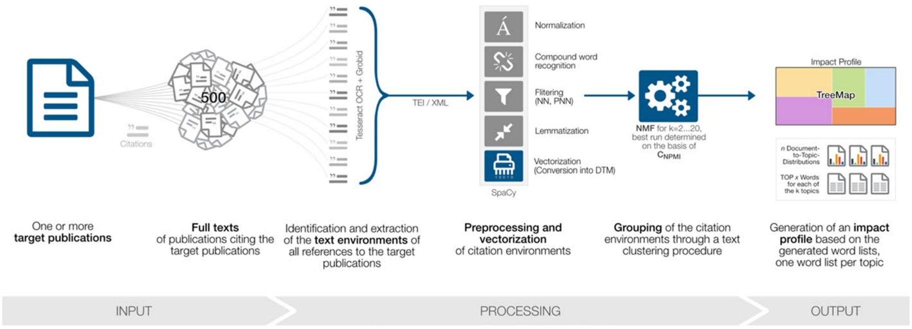
Citations
500
Tesseract OCR + Grobid
TEI / XML
Normalization
Compound word recognition
Filtering (NN, PNN)
Lemmatization
Vectorization (Conversion into DTM)
SpaCy
NMF for k=2...20, best run determined on the basis of C_{NPMI}
Impact Profile
TreeMap
n Document-to-Topic-Distributions
TOP x Words for each of the k topics
One or more **target publications**
Full texts of publications citing the target publications
Identification and extraction of the **text environments** of all references to the target publications
Preprocessing and vectorization of citation environments
Grouping of the citation environments through a text clustering procedure
Generation of an **impact profile** based on the generated word lists, one word list per topic
INPUT
PROCESSING
OUTPUT

Fig. 2 Detailed process outline to unpack the thematic context of citations. (Color figure online)

materials; most of these texts are less than one page long, and lack bibliographies, so they are of no use for our analyses. An overview of the included journals and the coverage is given in Table 1. Due to license restrictions, we were not able to acquire a small number of the earliest articles from the *Information Systems Journal* and the *Journal of the Association for Information Systems*. We identified a total of 397,077 entries in all bibliographies and 542,503 in-text references and their corresponding citation environments within the set of articles. The number of extracted citation environments may exceed the number of bibliographic entries because references can be used multiple times in an article. The citation environments were categorized into groups using the topic modeling algorithm, according to their similarity based on the co-occurrence of words within the citation's surroundings. In the end, the algorithm put out topic descriptors, i.e. lists containing terms characterizing each cluster. The lists were then taken to create a tree map impact profile. In this section, we will present three use cases for our method. The first two cases show citation profiles focusing on two units of interest: publications and authors. In the following, we present the ten most cited publications and authors within our article collection and discuss one result of each of them with the help of a visualization as an example of the citation profile. After that, we show how our approach can be utilized for the analysis of a concept instead of one or more articles. We do so with the help of an impact analysis of the method of citation analysis.

Impact analysis of a most-cited publication

The ten most-cited publications within our article collection are listed in Table 2. As our first exemplary use case, we graphically illustrate the impact profile of the most-cited paper, "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," by Fred D. Davis, Richard Bagozzi, and Paul R. Warshaw, published in *Management Science* in 1989, Vol. 35, No. 8, pp. 982–1003. We identified 176 papers from the corpus that cited the article. The article received a total of 394 in-text references yielding an equal number of extracted citation environments. Figure 3 shows the result of the overall process for the article. In this impact profile, every colored area represents one topic and contains words representative of the topic. The relative importance of each word in a topic is denoted as a percentage. The relative area of each rectangle corresponds to the number of mentions in the text associated with the topic, indicated by the number in round brackets. The colors and order of the areas, as well as the topic numbers, were assigned

Table 1 Coverage of full-text articles from the AIS senior scholars' basket

Journal	First Volume	Last Volume	First Year	Last Year	No. of Articles	No. of References	No. of Citation Environments
European Journal of Information Systems	1	10	1991	2019	1,079	58,389	73,058
Information Systems Journal	8	29	1998	2019	536	30,123	45,870
Information Systems Research	1	25	1990	2019	709	40,097	54,166
Journal of the Association for Information Systems	4	20	2003	2019	930	38,800	55,749
Journal of Information Technology	1	35	1986	2019	1,450	70,323	102,346
Journal of Management Information Systems	1	36	1984	2019	417	20,919	29,970
Journal of Strategic Information Systems	1	25	1991	2019	579	44,925	67,913
MIS Quarterly	1	43	1977	2019	1,682	93,501	113,431
Total					7,382	397,077	542,503

Table 2 Most cited publications within our set of documents from the AIS seniors scholars' basket

Authors	Title	Journal	Vol	No	No. of References	No. of Citation Environments
Fornell, Larcker	Evaluating Structural Equation Models with Unobservable Variables and Measurement Error	Journal of Marketing Research	18	1	308	403
Eisenhardt	Building Theories from Case Study Research	Academy of Management Review	14	4	231	430
Podsakoff, MacKenzie, Lee, Podsakoff	Common method biases in behavioral research: A critical review of the literature and recommended remedies	Journal of Applied Psychology	88	5	223	295
DeLone, McLean	Information Systems Success: The Quest for the Dependent Variable	Information Systems Research	3	1	185	292
Davis, Bagozzi, Warshaw	User Acceptance of Computer Technology: A Comparison of Two Theoretical Models	Management Science	35	8	176	394
Orlikowski, Iacono	Research Commentary: Desperately Seeking the IT in IT Research—A Call to Theorizing the IT Artifact	Information Systems Research	12	2	169	246
Moore, Benbasat	Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation	Information Systems Research	2	3	164	276
Malone, Yates, Benjamin	Electronic markets and electronic hierarchies	Communications of the ACM	30	6	162	285
Fisher, Chengalur-Smith, Ballou	The Impact of Experience and Time on the Use of Data Quality Information in Decision Making	Information Systems Research	14	2	161	279

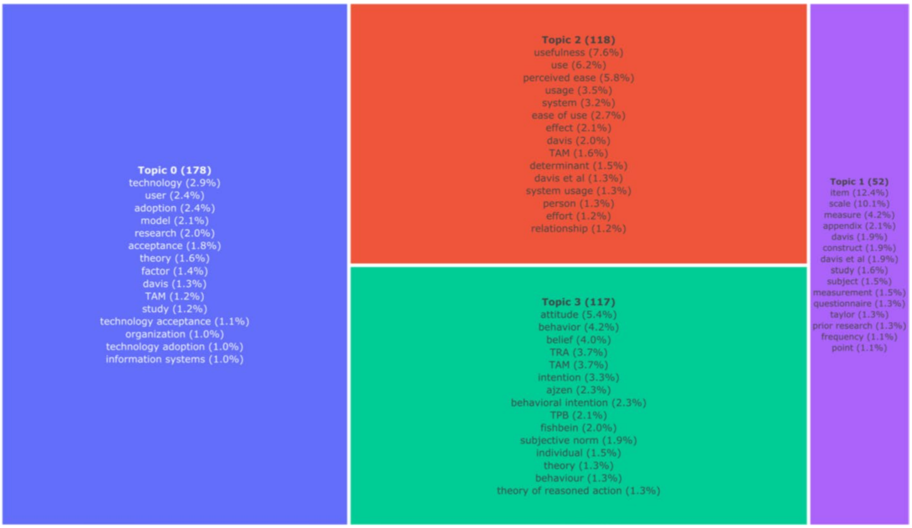

Fig. 3 Impact profile of the article "user acceptance of computer technology: a comparison of two theoretical models". (Color figure online)

randomly. As its title suggests, the article discusses two theoretical models of user acceptance of information technology, one being the Technology acceptance model (TAM) and the other the Theory of reasoned action (TRA). Topic 0 and Topic 2 represent the contribution of the paper to the further development of the TAM, whereas Topic 3 denotes its contribution to the TRA. Topic 1 shows the article's use as a source for its constructs, items, and scales. From this aggregated view, Topics 0 and 2 seem similar; in fact, it may be necessary to look at the citation environments directly to tell the precise difference. Since machines are unable to understand real meaning—they merely exploit statistical features of texts—the uncovering of fine-grained differences at this level is left to the user, as is the labeling of each topic. To assist with this, we devised an interactive variant of the impact profile showing the extracted text snippets as well as bibliographic information on the originating publication. We can make this variant available upon request.

Impact analysis of a most-cited author

Instead of focusing on individual articles, it is also possible to aggregate them and thus calculate citation profiles on a higher level. One obvious variant would be to aggregate them by author. Of course, this may be taken further and extended to departments, faculties or entire institutions. Other types of aggregation may yield valuable insights, too, e.g. grouping articles by country, journal, or discipline. In this section, we aggregate by author. The ten most-cited authors within our set of articles are listed in Table 3. We chose to graphically illustrate the impact profile of the second most cited author, Detmar Straub. Similar analyses of the other top 10 IS authors can be found in the appendix to this article. We identified 1063 references to him and 2194 corresponding citation environments. Figure 4 depicts his impact profile. In the following, we will provide exemplary citations by topic to support our interpretation. The references can be obtained from the source article

Table 3 Most cited authors within our set of documents from the AIS seniors scholars' basket

Author	No. of. references	No. of. Citation Environments
Benbasat, Izak	1217	2547
Straub, Detmar	1063	2194
Orlikowski, Wanda	1037	2154
Robey, Daniel	871	1719
Lyytinen, Kalle	783	1698
Markus, M. Lynne	856	1658
Zmud, Robert	845	1618
Davis, Fred	656	1577
Hirscheimer, Rudy	707	1471
Venkatesh, Viswanath	556	1397

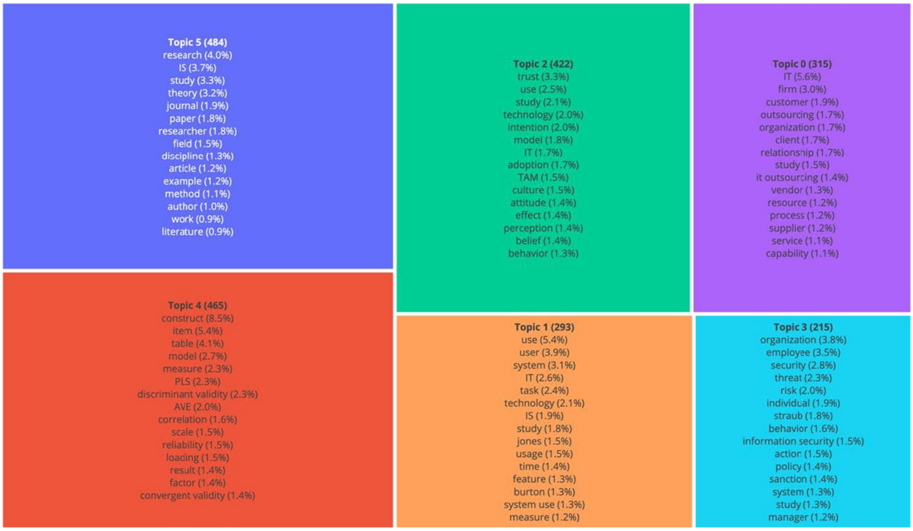

Fig. 4 Impact profile of Detmar Straub. (Color figure online)

list of citation environments our procedure produces. Topics 4 and 5 emerge as the most impactful ones. Topic 4 shows his influence on the further development of structural equation modeling and regression analysis (Gefen & Straub, 2005; Gefen et al., 2000). Topic 5 illustrates his impact on the development of the IS discipline (Straub, 1989; Straub et al., 2004). Topic 0 reflects his contributions to the field of it outsourcing (Ang & Straub, 1998; Koh et al., 2004). Topic 1 and 2 are linked to his work on technology adoption and technology acceptance and trust (Karahanna & Straub, 1999; Karahanna et al., 1999). Finally, topic 3 refers to Straub's research on computer abuse in organizations and IS security (Straub, 1990; Straub & Welke, 1998). Visualized in this way, an impact profile allows for a straightforward identification of the topics that a scientist has a significant influence on. It may pose a valuable addition to author profiles in library services, where the information may be further supplemented with the referenced and referencing articles.

Impact analysis of the method of the technology acceptance model

In the following, we showcase the potential of our approach for an entire area of research. The present procedure differs from the previous analyses in that it does not build on an impact profile in the first place but on a search in the citation environments. In this way, it enables the analysis of the impact, the influence, of a specific topic, and serves as a starting point for the search for impact profiles of publications that impact the topic in question. This approach is therefore particularly suitable for integration into search engines for academic literature or library services.

To support the presentation of our results and to facilitate the interpretation of the result visualization, we will briefly introduce the Technology Acceptance Model in the following few sentences. It is one of the most prominent theories in Information Systems (Davis, 1986). Drawing on the Theory of reasoned action (TRA) (Fishbein, 1975), the TAM explains new technology's adoption on the individual and organizational levels. According to the TAM, "perceived ease of use," "perceived usefulness," "attitude toward use," and "behavioral intention" predict the actual usage of new technology. Since its inception, the TAM has been the basis of numerous further developments, such as the Unified Theory of the Acceptance and Use of Technology (Venkatesh et al., 2003). The TAM is one of the most popular theories in Information Systems and has been acknowledged in many fields. However, the large amount of literature in IS makes it difficult to assess its impact manually. Hence, it serves as an ideal showcase to demonstrate the usefulness of our method. However, as the main purpose of this article is not to analyze the TAM but to present a new instrument for impact analysis, we will not provide an in-depth analysis here.

Usually, as shown by Mortenson and Vidgen (2016), among others, impact is assessed based on the number of citations an article has received. Table 4 shows the 10 most-cited articles obtained from the Thomson Reuters Web of Science using the search terms "technology acceptance model," sorted descending by citation count. The articles may be further aggregated by journal or author to identify the titles and researchers that have the highest impact on the matter in question. This data usually constitutes the basis for a quantitative or qualitative impact analysis or a literature review. However, this procedure implies that the analysis of the impact is based on articles that *contain* certain subject-specific keywords in their title or abstract and that have been cited frequently. It is therefore not based on articles that have been *cited for* the respective topic. Our newly—developed method may be utilized to enhance impact analyses. Instead of searching for keywords in titles or abstracts, we searched for keywords in citation environments of the citing articles. This way, we identified all articles that had been cited in the direct context of the desired keywords. We are hence able to unveil articles that contribute to the respective topic, but do not necessarily focus on it and may use diverging terminology. This way, we base our analysis on the level of topical indebtedness of citing and cited work.

Table 5 provides an overview of the ten most cited articles that have been referenced (up to 2019) in direct proximity to the key term "technology acceptance model." Half of the articles also appear in Table 4, but the other publications found in Table 5 are not present in Table 4. Those publications are marked bold and are directly related to the discourse on the TAM within IS, but they do not necessarily mention the TAM. Especially interesting in this regard, and illustrative for the strength of our method, is the article by Ajzen (1991), "The Theory of Planned Behavior." The article does not refer to

Table 4 Most cited TAM publication, ordered by citation count

Authors	Title	Journal	Vol	No	Cites
Venkatesh, Morris, Davis, Davis	User acceptance of information technology: Toward a unified view	MIS Quarterly	27	3	8,947
David, Bagozzi, Warshaw	User acceptance of computer technology—A comparison of 2 theoretical models	Management Science	35	8	7,428
Venkatesh, Davis	A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies	Management Science	46	2	5,696
DeLone, McLean	The DeLone and McLean model of information systems success: a ten-year update	Journal of Management Information Systems	19	4	3,260
Taylor, Todd	Understanding Information Technology Usage—A test of competing models	Information Systems Research	6	2	2,971
Gefen, Karahanna, Straub	Trust and TAM in online shopping: An integrated model	MIS Quarterly	27	1	2,439
Venkatesh, V	Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model	Information Systems Research	11	4	2,161
Bhattacherjee, A	Understanding information systems continuance: An expectation-confirmation model	MIS Quarterly	25	3	2,129
Compeau, Higgins	Computer Self-Efficacy—Development of a Measure and Initial Test	MIS Quarterly	19	2	2,049
Venkatesh, Thong, Xu	Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology	MIS Quarterly	36	1	1,787

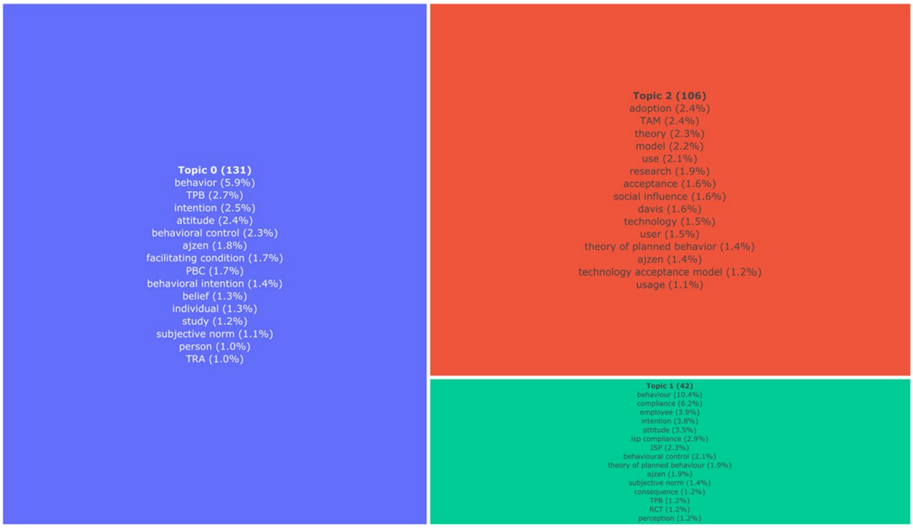

Fig. 5 Impact profile of the article "the theory of planned behavior". (Color figure online)

the TAM, nor does it mention it, so it does not show up in the traditional impact analysis (Table 4). However, it is directly related to the discourse on the TAM within IS. The Theory of planned behavior (TPB) is an alternative model to the TAM and was designed to predict behavior across many settings (Ajzen, 1985); it has often been applied to the uptake of information technology. According to the TPB, behavior is determined by intention and intention is predicted by the factors "attitude toward the behavior," "subjective norms," and "perceived behavioral control." The TPB addresses behavioral intentions—for example, to adopt IT innovations—and is hence regularly mentioned along with, and compared, to the TAM. This last fact is clearly discernible through investigating the citation environments of the article "The Theory of planned behavior", and may be further investigated by the citation profile of the article, as shown in Fig. 5. Topic 1 of the citation profile represents the discussions where TAM and TPB are compared or related to each other, supporting our interpretation. The other two topics denote the use of the TBP in several scenarios.

Our exemplary analysis the discourse on the TAM within IS illustrates how our method can extend traditional citation analysis by identifying research that truly impacts a topic of interest. Our method helps researchers to identify impactful literature with ease. We also developed an interactive variant of the illustration. It may serve as a starting point to dive deeper into the literature on the concept and can display citation profiles for the contained articles. Again, we can make this variant available upon request. Accordingly, we propose to enhance search engines for academic literature or library services in such a way.

Comparison with established content-based approaches

In the following, we compare our approach with other established content-based methods to unveil the scientific impact of authors or publications and discuss the merits of our approach. In this section, however, we only compare approaches that operate on the actual texts of publications. In general, these approaches can all be subsumed under science

Table 5 Most cited publications with TAM in their citation environments within our set of documents from the AIS seniors scholars' basket

Authors	Title	Journal	Year	Vol	No	No. of Cites
David, Bagozzi, Warshaw	User acceptance of computer technology—A comparison of 2 theoretical models	Management Science	1989	35	8	171
Venkatesh, Davis	A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies	Management Science	2000	46	2	102
Taylor, Todd	Understanding Information Technology Usage—A test of competing models	Information Systems Research	1995	6	2	69
Mathieson	Predicting User Intentions: Comparing the Technology Acceptance Model with the Theory of Planned Behavior	Information Systems Research	1991	2	3	57
Moore, Benbasat	Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation	Information Systems Research	1991	2	3	42
Ajzen	The Theory of Planned Behavior	Organizational Behavior and Human Decision Processes	1991	50	2	40
Venkatesh	Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the Technology Acceptance Model	Information Systems Research	2000	11	4	37
Benbasat, Barki	Quo vadis, TAM?	Journal of the Assoc. for Information Systems	2007	8	4	29
David, Bagozzi, Warshaw	Extrinsic and Intrinsic Motivation to Use Computers in the Workplace	Journal of Applied Social Psychology	1992	22	14	28
Szajna	Empirical Evaluation of the Revised Technology Acceptance Model	Management Science	1996	42	1	72

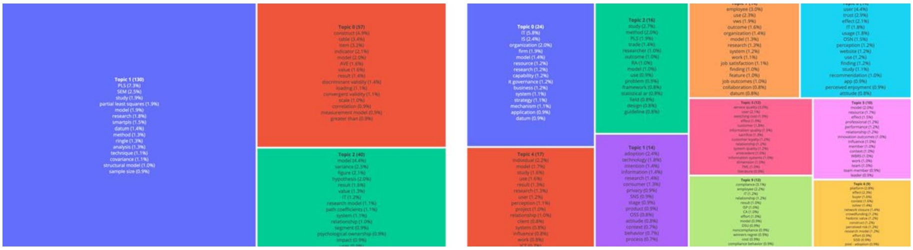

Fig. 6 Side-by-side comparison of Christian M. Ringle's citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

mapping (Gläser et al., 2017), that is, the thematic exploration and delineation of scientific disciplines. We do not include citation analytical approaches that are based solely on the graph—or network structure established from the references, without taking into account the publications' texts.

Established content-based methods from science mapping usually rely on abstracts or full texts of the citing publications and group them into meaningful clusters or construct semantic networks of characteristic terms, generating a map of the topics that appear in the abstracts or full texts (van Eck & Waltman, 2010). This is different from the topics for which a publication or author is cited. At the heart of this distinction is a more precise definition of impact. While methods based on abstracts or full texts of citing publications primarily reveal the topics *in which* an author or a publication is used, our approach reveals the topics *for which* an author or a publication is cited. The two approaches are therefore not to be considered substitutive but complementary; they each measure something different. Of course, there are also cases in which both analysis methods may produce identical results. This is the case when the topics found in the citation environments correspond to the topics in the abstracts or full texts. To further explicate this important difference, in the following three examples, we contrast both analytical approaches: a clustering of the citation environments and the clustering of the abstracts of the citing papers. Both approaches are each based on the dataset already presented here and use the same clustering and visualization technique to ensure proper comparability.

The left depiction in Fig. 6 shows the citation profile of the author Christian M. Ringle, who is well known for his works on structural equation modeling, computed from the 227 citation environments found in the publications in our dataset. The right depiction is a treemap representation of the clustering of the 142 abstracts of publications citing his papers. It is clearly evident that the thematic diversity in the illustration on the right is much greater than in the illustration on the left. This discrepancy is due to the fact that the left figure summarizes the topics *for which* the author is cited, whereas the right figure summarizes the topics *in which* the author is cited. Thus, the left figure represents the direct impact of the articles of Christian M. Ringle in the scientific community whereas the right figure reflects the thematic diversity of the publications the author is referenced in. To show that such a picture does not only emerge in the case of mainly methodological contributions, we replicate this analysis to assess the impact of an influential theory article. Figure 7 depicts the citation profile and the clustering of citing articles' abstracts of the article "Firm Resources and Sustained Competitive Advantage" published in *Journal of Management* by Jay Barney in 1991. Again, the side-by-side comparison of the two results shows the different focus of the analytical approaches. Our citation profile represents the themes the work is *cited for* and the clustering of abstracts shows the themes it is *cited in*.

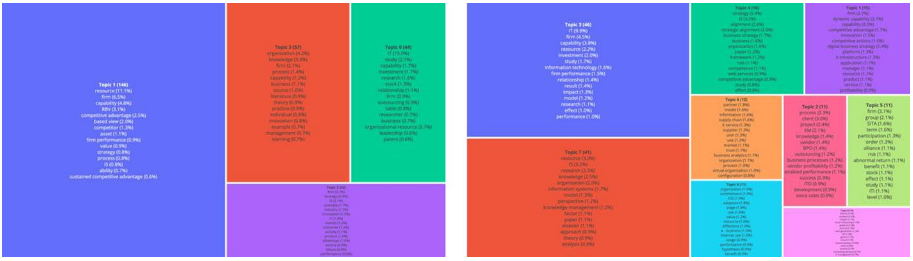

Fig. 7 Side-by-side comparison of the citation profile of the article “firm resources and sustained competitive advantage” (left) and a clustering of the papers’ abstracts citing this publication (right). (Color figure online)

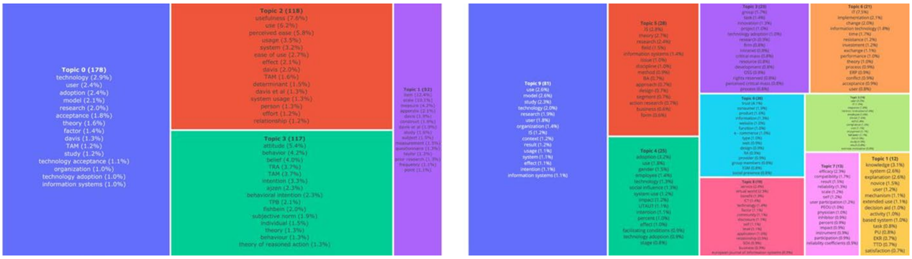

Fig. 8 Side-by-side comparison of the citation profile of the article “user acceptance of computer technology: A comparison of two theoretical models” (left) and a clustering of the papers’ abstracts citing this publication (right). (Color figure online)

In the appendix to this paper, we included similar side-by-side comparisons of the top 10 IS authors from our dataset.

Finally, we return to the paper mentioned at the beginning of this section to provide a somewhat more complex example of the method’s application and to discuss the relation of the two deptictions in more detail. Figure 8 contains the already shown citation profile of the article “User acceptance of computer technology: A comparison of two theoretical models” published in *Management Science* by Fred D. Davis et al. (1989) from Fig. 3 in addition to the clustering of the citing articles’ abstracts. Again, the side-by-side comparison of the two results shows the different focus of the analytical approaches. As mentioned above, some topics do not seem to be clearly distinguishable in the citation profile shown on the left-hand side. In this case, it is necessary to take a direct look at the citation environments to be able to interpret the results. All three topics 0, 2, and 3 have, not surprisingly, some connection to the Technology acceptance model (TAM), the Theory of reasoned action (TRA) and the constructs underlying those models. Topic 0, however, seems more ambiguous and less clearly defined. An examination of the citation environments underlying this topic reveals that the articles in this topic embed TAM in the broader context of IT adoption and IT acceptance, whereas in topic 2 and 3, the articles instead the model directly along with related theories such as TRA and/or its constitutive constructs. A different thematic picture emerges in the right depiction showing the treemap representation of the clustering of the papers’ abstracts citing the focal article by Davies et al. (1989). The thematic map is much more diverse and mainly shows areas of application of the TAM. A closer look at the citing articles’ abstracts reveals that the articles use the TAM directly, for example in the context of trust and online applications (topic 0) and online/virtual communities (topic 8). It is widely used in the context of IT implementation and change management (topic 6) and even more generally in the context of IS evolution, where the TAM is discussed as one of the most prominent IS theories, e.g., related to the nature of theory in IS research (topic 5).

In summary, the combined citation profile shown helps to unearth the specific nature of the article's theoretical contribution (left) and the thematic scope of its application (right). Such a comparison can be of value in that it allows researches to distinguish the direct intellectual impact of an article and the theory proposed therein on the one hand and the areas of application on the other hand. Importantly, a sole focus on the clustering results on the right might researchers to underestimate the article's significant contribution to subsequent scale and construct development as well as to broader theoretical discussions with emphasis on TAM, TRA, and TPB. In cases such as this, methods exclusively based on abstracts or full texts of citing publications risk revealing application areas only. The approach we advance complements such efforts by helping to unearth articles' intellectual contribution to scholarly conversations. Overall, this example demonstrates that the method presented is indeed capable of extracting the topics *for which* a publication is cited, as opposed to the topics *in which* a publication is used.

Conclusion

Impact assessment of extant research is central to all academic activities. Researchers new to a field must quickly identify the most relevant literature for the research question at hand. Academic disciplines are concerned about the impact of journals and specific topics examined. Hiring committees demand informative and meaningful academic profiles. And policy makers are interested in assessing the effectiveness of their measures. However, the increasing volume of academic output makes keeping track of the impact of research a challeging endeavor, especially if one relies on manual assessment only.

Contribution to research and practice

In this paper, we propose a new method to enhance citation analysis by focusing on the thematic context of citations. Our method differs from previous content-based citation analysis approaches which employ classification schemas to find the reasons and functions for citations (Ding et al., 2014). Our method automatically derives and visualizes impact profiles by exploring the topical references in the full texts of the citing works. In an experiment to demonstrate the method, we obtained a profile of a focal paper's impact and absorption within the scientific community. The data generated can be used to visualize an impact profile, as shown in the previous section. This way, our method may assist in impact assessments of publications in a scalable and automated fashion. The keyword set characterizing each topic in the citation profiles may then serve as a starting point for further exploration of academic literature. Temporal citation profiles may help uncover reinvigorating and weakening topics in a scientific field in general or the changing influence of a paper or paper set in particular. Another use case may be contrasting the authors' positioning of a paper and its actual absorption within the community by comparing the author's keywords and the obtained thematic citation profile. All in all, the proposed approach has the potential to serve as a starting point for a considerable number of meaningful follow-up analyses.

In our case study of the Technology Acceptance Model (TAM), we show how to exploit our approach to facilitate impact analysis of a concept within an entire field of research. Here, we searched for the topic of interest within the citation environments, instead of in article metadata, as is usually done. We identified articles that have been cited in the direct context of the TAM and thus were able to reveal articles that contribute to the discourse around TAM but do not mention it by name or discuss it explicitly. Apart from the use cases we presented to enhance citation analysis, our method might also prove useful for funding bodies, ranking agencies, and, more broadly, anyone interested in the impact of scholars, research groups, or institutions. It may be used to make funding activities more targeted and more accessible to systematic impact assessment. It also makes it possible to compare the impact of scientific institutions on a thematic level, which might also render university rankings more insightful.

Contribution to IS research

From the perspective of an IS researcher, this study offers an effective way to identify significant contributions to IS topics. It gives a preliminary insight into how impact analyses may be improved. And it may begin to help IS researchers to better consolidate a generally-accepted body of IS knowledge based on the most significant IS research in a particular area. Following up on Walstrom and Leonard (2000) and Whitley and Galliers (2007), it also facilitates the development of reading lists for scholars for artifacts, concepts, and theories. A search engine could integrate our approach to allow researchers to perform analyses similar to our case study. Future work could expand our showcase analyses to contribute to a better understanding of the IS field. Additionally, our selection of journals may be complemented with additional titles in the future to capture an even clearer portrait of impact within the IS community.

Limitations and future research

A general limitation of our approach stems from the fact that the full texts of all citing papers are necessary to compute the impact profile for one or more publications. It might not always be possible to obtain the data due to license restrictions or the technical efforts necessary for its acquisition. However, we expect the ongoing process of digitalization to ease the problem in the future. Some publishers already provide machine-readable full-text data and/or reference data. Additionally, more and more articles are made available open access. In order to further explore the possible application areas of the presented method and also to point out the differences to other methods, we plan to undertake a more comprehensive comparison of methods for determining impact including methods that are not based on full papers or abstracts but on graph-analytical methods. The approach itself may be further developed in many respects and we have identified several other areas for improvement. The citation matching and reference identification procedures we employed yield acceptable results, but an increase in reference identification accuracy will further enhance analytical capabilities, as the quality and generalizability of the analyses depend on the references identified. The text preprocessing may be further enhanced to account for synonyms. The presentation of the results may be improved by optimizing font sizes and color schemes. Additional visualization techniques may enhance usability (Chaney & Blei, 2012; Chuang et al., 2012; Sievert & Shirley, 2014). Automatic labelling of the

Table 6 Top 10 IS authors within our set of documents from the AIS seniors scholars' basket

Author	No. of references	No. of citation environments
Benbasat, Izak	1217	2547
Straub, Detmar	1063	2194
Orlikowski, Wanda	1037	2154
Robey, Daniel	871	1719
Lyytinen, Kalle	783	1698
Markus, M. Lynne	856	1658
Zmud, Robert	845	1618
Davis, Fred	656	1577
Hirscheimer, Rudy	707	1471
Venkatesh, Viswanath	556	1397

derived thematic contexts may greatly enhance the comprehensiveness of the visualizations (Chang et al., 2009; Chuang et al., 2014; Hindle et al., 2013; Mei et al., 2007; Nolasco & Oliveira, 2016). Another useful extension would be to track changes of thematic impact over time, similar to what is already possible in the science mapping tools. This would be feasible by including the timestamps available in the publication metadata. (van Eck & Waltman, 2010) and would make it possible to plot knowledge absorption curves over time. Finally, complementing our approach with already-existing semantic and syntactic content-based citation analysis techniques, especially those analyzing the argumentative context (Bertin, 2008; Ritchie et al., 2008; Valenzuela et al., 2015), would connect the dots between the distinct current developments in citation analysis.

Appendix

In the following, we include citation profiles of the top 10 IS authors from our dataset as further examples to the analyses presented in the publication. The analyses are presented in the form of side-by-side comparisons of citation profiles and clustering result depictions based on the abstracts of publications that cite the author. The figures may help assess the authors' immediate impact in the IS community (*for which* topics they are cited) as well as the subject areas in which their work is used (*in which* topics they are cited). The following Table 6 gives an overview of the included authors, their no. of references and no. of identified citation environments within our corpus of 7,382 full-text articles from the AIS Seniors Scholars' Basket.

The citation profiles and clustering results shown below include recurring themes, such as *IS research* and *IS theory*, which is due to the fact that a number of well-known IS researchers were analyzed here, all of whom have published overarching contributions to the research field itself. Other recurring themes include the *Technology Acceptance Model*, one of the most prominent theories in Information Systems, and higher-level themes such as the *IT artifact*, *technology adoption*, as well as *IT governance*, *IT outsourcing*, *IS alignment*, *team performance*, *project management, virtual teams* and *trust*.

Benbasat, Izak

See Fig. 9

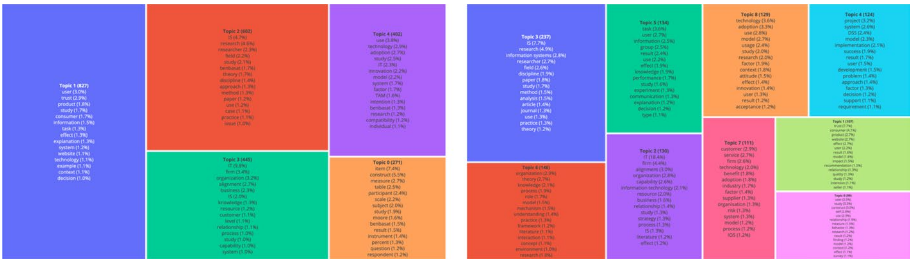

Fig. 9 Side-by-side comparison of Izak Benbasat's citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

Straub, Detmar

See Fig. 10

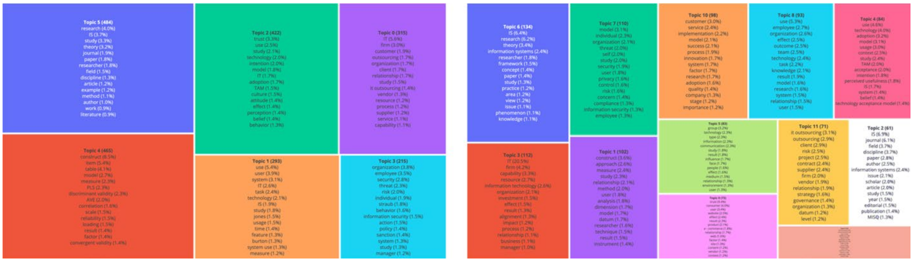

Fig. 10 Side-by-side comparison of Detmar Straub's citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

Orlikowski, Wanda

See Fig. 11

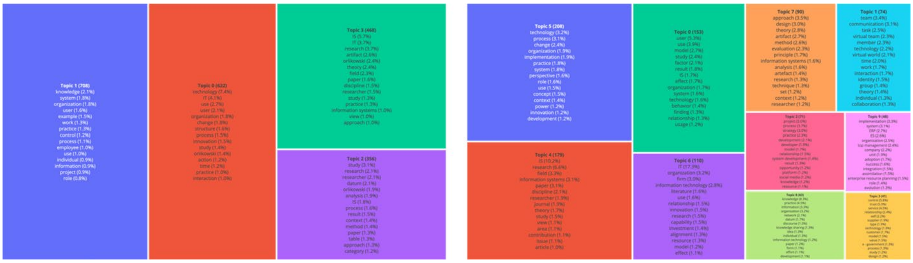

Fig. 11 Side-by-side comparison of Wanda Orlikowski's citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

Robey, Daniel

See Fig. 12

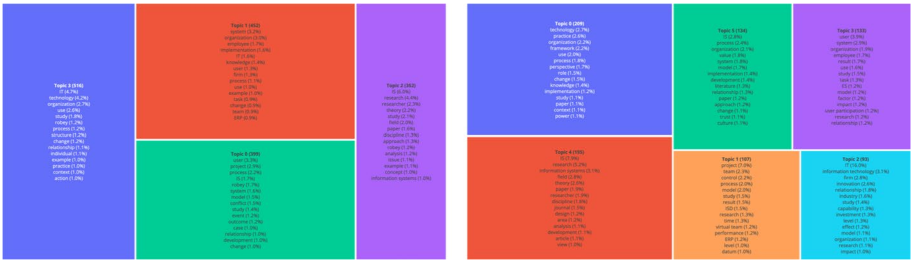

Fig. 12 Side-by-side comparison of Daniel Robey's citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

Lyytinen, Kalle

See Fig. 13

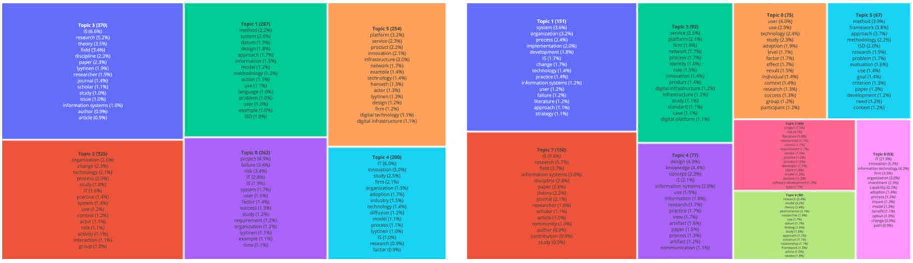

Fig. 13 Side-by-side comparison of Kalle Lyytinen's citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

Markus, M. Lynne

See Fig. 14

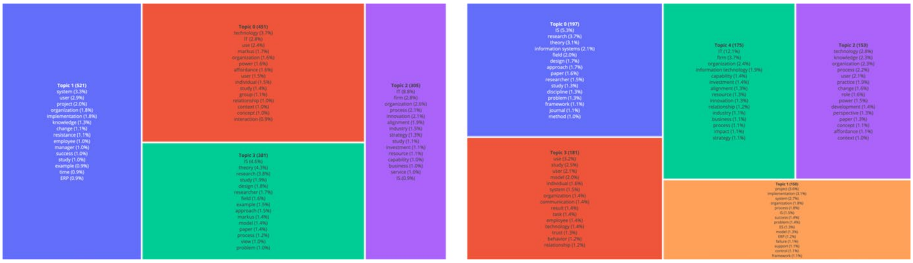

Fig. 14 Side-by-side comparison of M. Lynne Markus' citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

Zmud, Robert

See Fig. 15

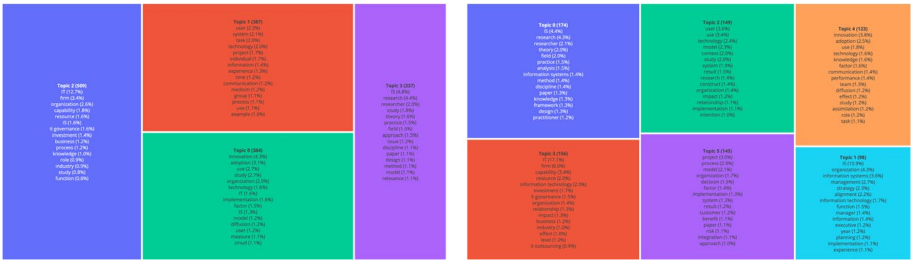

Fig. 15 Side-by-side comparison of Robert Zmud's citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

Davis, Fred

See Fig. 16

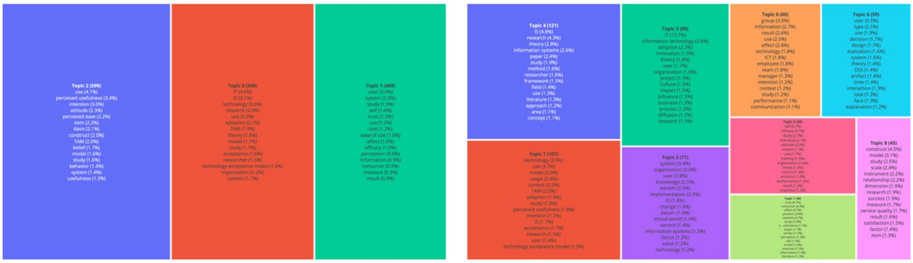

Fig. 16 Side-by-side comparison of Fred Davis' citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

Hirscheimer, Rudy

See Fig. 17

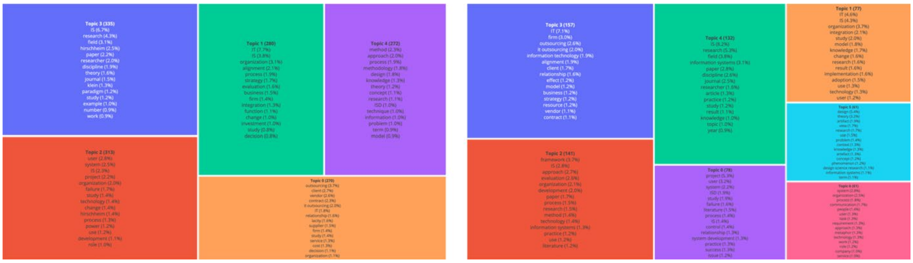

Fig. 17 Side-by-side comparison of Rudy Hirschheimer's citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

Venkatesh, Viswanath

See Fig. 18

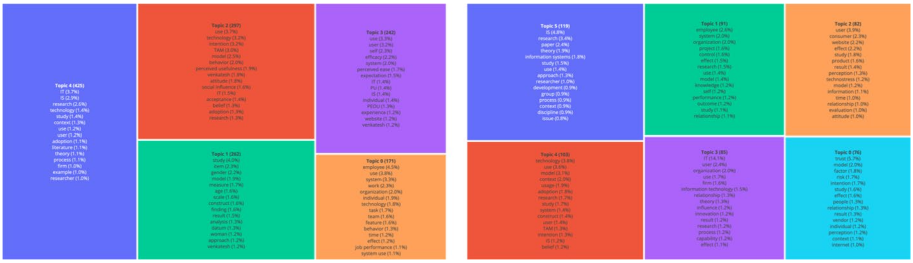

Fig. 18 Side-by-side comparison of Viswanath Venkatesh's citation profile (left) and a clustering of the papers' abstracts citing his publications (right). (Color figure online)

Authors contribution All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Matthias Rüdiger. The first draft of the manuscript was written by Matthias Rüdiger. Torsten-Oliver Salge edited previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by Federal Ministry of Education and Research (BMBF) Germany, Grant No. 01PU17020.

Code availability The software developed for the purposes of the analyses presented in the article is available on request via e-mail. It will be made publicly available at the end of the funding measure.

References

Ajzen, I. (1985). From intentions to actions: A Theory of Planned Behavior. In *Action control*, Berlin, Heidelberg: Springer.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179–211.

Amsterdamska, O., & Leydesdorff, L. (1989). Citations: Indicators of significance? *Scientometrics, 15*(5–6), 449–471.

Ang, S., & Straub, D. W. (1998). Production and transaction economies and IS outsourcing: A study of the US banking industry. *MIS Quarterly, 22*(4), 535–552.

Angrosh, M. A., Cranefield, S., & Stanger, N. (2010). Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries. *Proceedings of the 10th annual joint conference on Digital libraries*, p. 293–302.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review, 63*(6), 829–846.

Bertin, M. (2008). Categorizations and Annotations of Citation in Research Evaluation. *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference*, p. 456–461.

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of German Society for Computational Linguistics & Language Technology*, p. 31–40.

Boyack, K. W., Van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics, 12*(1), 59–73.

Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science, 36*(4), 223–229.

Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the Association for Information Science and Technology, 51*(7), 635–645.

Chaney, A. J.-B., & Blei, D. M. (2012). Visualizing topic models. *Sixth International AAAI Conference on Weblogs and Social Media*, p. 419–422.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems, 22*, 288–296.

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems, 163*, 1–13.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of the International Working Conference on Advanced Visual Interfaces*, p. 74–77.

Chuang, J., Wilkerson, J. D., Weiss, R., Tingley, D., Stewart, B. M., Roberts, M. E., Poursabzi-Sangdeh, F., Grimmer, J., Findlater, L., & Boyd-Graber, J. (2014). Computer-assisted content analysis: Topic models for exploring multiple subjective interpretations. *Advances in Neural Information Processing Systems Workshop on Human-Propelled Machine Learning*, p. 1–9.

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science, 5*(4), 423–441.

Cronin, B. (1984). *The Citation Process: The Role and Significance of Citations in Scientific Communication*. Taylor Graham.

Davis, F. (1986). *A technology acceptance model for empirically testing new end-user information systems: Theory and results* [PhD Thesis]. Massachusetts Institute of Technology.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982–1003.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Scarecrow Press.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583–592.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology, 65*(9), 1820–1833.

Fishbein, M. (1975). *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Longman Higher Education.

Garfield, E. (1965). Can citation indexing be automated. *Statistical Association Methods for Mechanized Documentation, Symposium Proceedings*, p. 189–192.

Gefen, D., & Straub, D. W. (2005). A practical guide to factorial validity using PLS-Graph: Tutorial and annotated example. *Communications of the Association for Information Systems, 16*(1), 5.

Gefen, D., Straub, D. W., & Boudreau, M.-C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems, 4*(1), 7.

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics, 111*, 981–998.

Grover, V. (2012). The Information Systems Field: Making a Case for Maturity and Contribution. *Journal of the Association for Information Systems, 13*(4), 254–272.

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the Association for Information Science and Technology, 29*(6), 308–310.

Hernández-Alvarez, M., & Gomez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering, 22*(3), 327–349.

Hindle, A., Ernst, N. A., Godfrey, M. W., & Mylopoulos, J. (2013). Automated topic naming. *Empirical Software Engineering, 18*(6), 1125–1155.

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics, 7*(4), 887–896.

Jha, R., Jbara, A.-A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering, 23*(1), 93–130.

Karahanna, E., & Straub, D. W. (1999). The psychological origins of perceived usefulness and ease-of-use. *Information & Management, 35*(4), 237–250.

Karahanna, E., Straub, D. W., & Chervany, N. L. (1999). Information technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs. *MIS Quarterly, 23*(2), 183–213.

Koh, C., Ang, S., & Straub, D. W. (2004). IT outsourcing success: A psychological contract perspective. *Information Systems Research, 15*(4), 356–373.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, p. 556–562.

Leydesdorff, L., & Milojevic, S. (2012). Scientometrics. *ArXiv*, abs/1208.4566.

Liu, X., Zhang, J., & Guo, C. (2013). Full-text citation analysis: A new method to enhance scholarly networks: Full-Text Citation Analysis: A New Method to Enhance Scholarly Network. *Journal of the American Society for Information Science and Technology, 64*(9), 1852–1863.

Maričić, S., Spaventi, J., Pavičić, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. *Journal of the Association for Information Science and Technology, 49*(6), 530–540.

Mei, Q., Shen, X., & Zhai, C. (2007). Automatic labeling of multinomial topic models. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 490–499.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.

Moravcsik, M. J. (1973). Measures of scientific growth. *Research Policy, 2*(3), 266–275.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86–92.

Mortenson, M. J., & Vidgen, R. (2016). A computational literature review of the Technology Acceptance Model. *International Journal of Information Management, 36*(6), 1248–1259.

Nolasco, D., & Oliveira, J. (2016). Detecting Knowledge Innovation through Automatic Topic Labeling on Scholar Data. *Proceedings of the 49th Hawaii International Conference on System Sciences*, 358–367.

Pasula, H., Marthi, B., Milch, B., Russell, S. J., & Shpitser, I. (2003). Identity uncertainty and citation matching. *Advances in Neural Information Processing Systems*, 1425–1432.

Phelan, T. J. (1999). A compendium of issues for citation analysis. *Scientometrics, 45*(1), 117–136.

Ritchie, A., Robertson, S., & Teufel, S. (2008). Comparing citation contexts for information retrieval. *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 213–222.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics, 11*(1), 92–99.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, p. 63–70.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly, 13*(2), 147–169.

Straub, D. W. (1990). Effective IS security: An empirical study. *Information Systems Research, 1*(3), 255–276.

Straub, D. W., Boudreau, M.-C., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems, 13*(1), 24.

Straub, D., & Welke, R. J. (1998). Coping with systems risk: Security planning models for management decision making. *MIS Quarterly, 22*(4), 441–469.

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics, 121*(3), 1635–1684.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, p. 103–110.

Tkaczyk, D., Collins, A., Sheridan, P., & Beel, J. (2018). Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, p. 99–108.

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*, 523–538.

Venkatesh, M., Davis, & Davis. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly, 27*(3), 425–478.

Vinkler, P. (1987). A quasi-quantitative citation model. *Scientometrics, 12*(1–2), 47–72.

Voos H, Dagaev KS (1976). Are all citations equal? Or did we op cit. your idem?. *Journal of Academic Librarianship*, *1*(6): 19–21.

Walstrom, K. A., & Leonard, L. N. (2000). Citation classics from the information systems literature. *Information & Management, 38*(2), 59–72.

Whitley, E. A., & Galliers, R. D. (2007). An alternative perspective on citation classics: Evidence from the first 10 years of the European Conference on Information Systems. *Information & Management*, *44*(5), 441–455.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology, 66*(2), 408–427.

Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods, 18*(3), 429–472.