Evaluating semantometrics from computer science publications

Christin Katharina Kreutz¹ · Premtim Sahitaj¹ · Ralf Schenkel¹

Received: 23 September 2019 / Published online: 18 March 2020
© The Author(s) 2020

Abstract

Identification of important works and assessment of importance of publications in vast scientific corpora are challenging yet common tasks subjected by many research projects. While the influence of citations in finding seminal papers has been analysed thoroughly, citation-based approaches come with several problems. Their impracticality when confronted with new publications which did not yet receive any citations, area-dependent citation practices and different reasons for citing are only a few drawbacks of them. Methods relying on more than citations, for example semantic features such as words or topics contained in publications of citation networks, are regarded with less vigour while providing promising preliminary results. In this work we tackle the issue of classifying publications with their respective referenced and citing papers as either seminal, survey or uninfluential by utilising semantometrics. We use distance measures over words, semantics, topics and publication years of papers in their citation network to engineer features on which we predict the class of a publication. We present the SUSdblp dataset consisting of 1980 labelled entries to provide a means of evaluating this approach. A classification accuracy of up to .9247 was achieved when combining multiple types of features using semantometrics. This is +.1232 compared to the current state of the art (SOTA) which uses binary classification to identify papers from classes seminal and survey. The utilisation of one-vector representations for the ternary classification task resulted in an accuracy of .949 which is +.1475 compared to the binary SOTA. Classification based on information available at publication time derived with semantometrics resulted in an accuracy of .8152 while an accuracy of .9323 could be achieved when using one-vector representations.

Keywords Semantometrics · Classification · Citation network · Natural language processing

✉ Christin Katharina Kreutz
kreutzch@uni-trier.de

Premtim Sahitaj
s4prsahi@uni-trier.de

Ralf Schenkel
schenkel@uni-trier.de

¹ Trier University, 54286 Trier, Germany

Introduction

With the ever growing amount of scientific publications, automatic methods for finding influential or seminal works are indispensable. A majority of research tackles the identification of important works (Gerrish and Blei 2010; Herrmannova et al. 2018; Simkin and Roychowdhury 2005; Wade et al. 2006; Whalen et al. 2005; Zhu et al. 2015). The diffusion of scholarly knowledge in a citation network is explicitly modelled by citations and references (Börner et al. 2006), ideas from referenced papers can be utilised or amended. Common approaches are based on the observation of the number of citations which publications received. As this indicator can be highly dependent on a specific dataset, it might be problematic to utilise as a measure of impact (Moed 2002; Seglen 1997). Citations need to be handled with care due to cases of self-citations (Jin et al. 2007; Schreiber 2007, 2008), varying citation practices in different areas (Cronin and Meho 2006; Jin et al. 2007; Seglen 1992, 1997; Shi et al. 2010), diverging reasons for citing (Garfield 1964), the non-existence of citations of new papers (Wade et al. 2006) and uncited influences (Garfield 1964; MacRoberts and MacRoberts 2010; Patton et al. 2016).

Distinguishing between seminal publications and popular survey papers might pose a problem as both types are typically cited often (Seglen 1997) but reviews are over-represented amongst highly cited publications whilst not contributing any new content (Aksnes 2003). Seminal papers are ones which are key to a field while surveys review and compare multiple approaches and can be comprehensible summaries of a domain. For lack of space, reviews are often referenced instead of original papers (Hou et al. 2011). Influential members of both classes can be distinguished from all other (uninfluential) publications by observing their number of citations after an initial period in which citations are accumulated. Differentiating between seminal and review papers is challenging. Therefore, methods considering more factors than the number of citations and references are required (Moed 2002; Seglen 1997; Wade et al. 2006) as these observations are no sufficient proxy in the process of measuring publication impact and scientific quality (Herrmannova et al. 2017; Seglen 1997), especially at the time a paper is first published. Preferably, an approach with the potential to measure and predict the contribution of a paper and how much it advances its field should be favoured.

Herrmannova et al. (2018) assume the classification of a paper as seminal or survey can be performed by observing semantometrics as a new method for research evaluation which uses differences in full texts of a citation network to determine the contribution or value of a publication (Knoth and Herrmannova 2014). Classification of a publication is conducted on features derived from distances between papers in its citation network. The distance between papers citing a seminal paper and its referenced papers is shown to be larger than this distance of a survey as the seminal publication advanced science by a considerable margin. Surveys are shown to reference papers from a broader field compared to seminal papers (Herrmannova et al. 2018).

Experiments of Herrmannova et al. (2017) were conducted on a multi-disciplinary dataset, where an accuracy of up to .6897 was achieved. Kreutz et al. (2019) performed similar experiments on a dataset covering the area of computer science and achieved accuracies up to .8015. We define the identification of seminal and survey papers as our core task but also want to incorporate publications into the approach which are uninfluential. This is done to broaden the methodology and to place it in a more realistic setting. Additionally, we want to predict whether a paper is seminal, survey or uninfluential based on the information

available at the time of publication. So, the usefulness of this approach is accessed on a dataset restricted to a narrower area while including a third kind of publication.

Our contribution is three-fold: First, we introduce SUSdblp, a dataset suitable for the task of classifying a publication as seminal, survey or uninfluential by providing reference and citation information of publications from the area of computer science. Second, we analyse the approach presented by Herrmannova et al. (2018) in a ternary class setting, using different document representations which encode words, semantics, topics and years of papers as well as numerous classification algorithms. Single and multiple features generated from publications of the new and homogeneous dataset are evaluated during the classification process. Third, we extend and modify the approach presented by Herrmannova et al. (2018) by combining features derived from the different aspects of documents and classify solely on features which are known as soon as a paper is published. In doing so, we introduce a prediction task which differs from the former classification task.

The remainder of this paper is organised as follows: “[Semantometrics and related work](#)” section gives an overview of the established conceptual background of semantometrics and related research. In “[SUSdblp dataset](#)” section, the SUSdblp dataset is presented. The succeeding “[Methodology](#)” section describes the used methodology in detail: utilised document vector representations, distance measures and classification algorithms to apply Herrmannova et al.’s approach (Herrmannova et al. 2018) to the new dataset in a ternary class setting as well as the extension of the methods are introduced. A detailed evaluation of our methodology is given in “[Evaluation of the approach](#)” section which is followed by the evaluation of our dataset in “[Evaluation of the dataset](#)” section. In “[Alternative approaches for classification](#)” section we compare classification based on semantometrics to alternative methods.

Semantometrics and related work

Before the methodology is described in detail, related research is covered to integrate this work in a broader context. We first introduce the concept of semantometrics and then present adjacent areas of research.

Semantometrics

Feature engineering on data through mathematical descriptors is common in medical image analysis (Gillies et al. 2016; Kumar et al. 2012). For publication networks, it was initially introduced as *semantometrics* by Knoth and Herrmannova (2014) to assess research contribution.

Herrmannova et al.’s approach (Herrmannova et al. 2018) which uses the principles described by Knoth and Herrmannova (2014) is the foundation of this work. They were the first to utilise citation networks for the classification of a publication P as seminal or survey paper. A citation network centres around P and connects it to papers referenced by P (X) and papers citing P (Y). Semantic distances that describe the relationships amongst publications were measured: the distances between titles and abstracts of X and Y are contained in group A , distances between a publication and its referenced papers X are included in group B , and group C is composed of distances between P and its citing papers Y . The semantic distances between entries of X can be found in group D , symmetrically, distances between citing publications Y are stored in E . Figure 1 visualises the different groups of

Fig. 1 Neighbourhood of publication P . Nodes symbolise publications, straight edges between papers represent citations. $X = \{x_0, \dots, x_n\}$ are papers referenced by P , $Y = \{y_0, \dots, y_m\}$ are papers citing P . Dotted edges symbolise observed relationships between publications. Group A contains distances between pairs of referenced (X) and citing papers (Y). Group B contains distances between referenced papers (X) and P . Group C contains distances from P and citing papers (Y). Group D contains distances between pairs of referenced papers (X). Group E contains distances between pairs of citing papers (Y)

relations in a citation network for a given publication P . From these five groups, twelve features each were extracted such as min, max and mean distance between papers. Herrmannova et al. (2018) show that seminal papers are associated with larger semantic distances between papers from X and Y than surveys, while in turn surveys are associated with larger distances between referenced publications compared to seminal ones.

To enable the computation of distances between publications, P and the papers contained in X and Y need to be represented by vectors. Feature $\text{sum}B$ would thus describe the summed up distance between each vector representing a reference of a paper (all papers contained in X) and the vector of paper P itself:

$$\text{sum}B = \sum_{i=0}^n \text{dist}(x_i, P), x_i \in X$$

Here, $\text{dist}(a, b)$ measures the distance between two papers a and b . Another feature $\text{min}E$ would describe the smallest distance from any pairs of vectors representing papers Y by which publication P is cited:

$$\text{min}E = \min_{\forall y_i, y_j \in Y, i \neq j} \text{dist}(y_i, y_j)$$

The applied distance measure as well as the utilised vector representation of papers are parameters.

Herrmannova et al. (2018) utilised cosine distance on tf-idf vectors of publications. A classification accuracy of .6897 was achieved in the binary classification task by using Naïve Bayes classifier on single features derived from papers in their citation network based on the TrueImpactDataset (Herrmannova et al. 2017). The dataset was created from a user study and contains publications from multiple distinctly labelled fields. Other previous work comes from Kreutz et al. (2019). They applied cosine distance as well as Jaccard distance on tf-idf vectors and Doc2Vec representations of publications in their citation network to derive the features for classification. The dataset used in this revaluation of Herrmannova et al.'s approach is based in the narrower area of computer science (<https://doi.org/10.5281/zenodo.3258164>). An accuracy of up to .7432 for single features and an accuracy of .8015 using multiple features for classifying publications were achieved in the binary classification task.

Prior research tackling the task of quality estimation of papers using semantometrics solely focuses on the two classes seminal and survey but does not observe the group of papers which are neither groundbreaking nor reviews. In contrast, the task we describe includes this third group of publications (uninfluential) for the findings to be applicable in a real-world scenario where no predefined categorisation separates the different types of publications from each other.

Related work

Relevant areas for our work besides semantometrics are scientific influence assessment, identification of important referenced papers, general citation behaviour analysis as well as document vector representation methods.

Scientific influence assessment

The fields of influence assessment of scientific papers and influence estimation of authors or author groups can be seen as alternatives to using semantometrics as an indicator for publication quality.

Several papers can be found in the field of *influence estimation of publications*. Gerrish and Blei (2010) observe topical developments in corpora and topical compositions of documents to identify influential publications, this approach is correlated with citation counts. Prediction of high-impact papers can also be done by observing similarities in full texts of citing papers. In this case, citations of topical similar papers are an indicator of wide influence (Whalen et al. 2005). The prediction of citation counts is also rooted in this domain. This influence of a paper can be accessed by using text-similarity of extracted popular terms of publications. Context-aware citation analysis on full texts and leading edge impact assessment (Patton et al. 2016) can be utilised for estimating impact of publications. Patton et al. (2017) propose an audience based measure which leverages citation counts of publications and altmetrics to describe the influence of papers on the scientific community and the general public.

There are multiple approaches for the *assessment of influence of authors or author groups* and thus indirectly the influence of publications they write. The most popular one might be the h -index for estimation of author impact based on citation counts of authored publications (Hirsch 2005). It is defined as the highest value h such that an author has written at least h publications, each of which has been cited at least h times. There are numerous works extending, modifying, complementing or improving the h -index (Banks 2006; Bornmann and Daniel 2010; Jin et al. 2007; Rousseau and Ye 2008). The g -index is an improvement which measures global citation performance of sets of articles based on citation counts (Egghe 2006). Here, publications of an author are ordered decreasingly by the received number of citations, the g -index is the largest value such that the added citation count of the top g papers is at least g^2 . An example for a complement of the h -index is the index h_a which quantifies an author's scientific leadership (Hirsch 2019). The papers contributing to the h -index of an author are observed, the number of times the author is the coauthor with the highest h -index of a publication in this set of papers is his h_a -index. Citation-based methods come with a number of drawbacks: Citations are highly influenced by the data source they are extracted from (Seglen 1997; Moed 2002). Self-citations can boost scores relying solely on numbers of citations by a great margin. As their influence and meaning depends on the field, self-citations need to be handled case-by-case (Jin et al.

2007; Schreiber 2007, 2008). Citation practices in general are highly dependent on the different areas (Cronin and Meho 2006; Jin et al. 2007; Seglen 1992, 1997; Shi et al. 2010). Another argument for caution when working with citation-based methods are diverging and unclear reasons for citing such as paying homage to pioneers, corrections of previous work or providing background reading material (Garfield 1964). In cases of non-existence of citations occurring with new papers (Wade et al. 2006) or a high amount of uncited sources (Garfield 1964; MacRoberts and MacRoberts 2010; Patton et al. 2016), citation-based methods cannot reliably perform influence estimations.

Another measure which can be utilised for influence assessment not relying on citations is the so-called research endogamy which observes fluctuation or stability of members in research teams (Montolio et al. 2013). It can be applied on different fine-grained or broad communities and venues (Silva et al. 2014) to estimate the quality and therefore influence of papers. Research endogamy can also be combined with semantic information and citation counts (Herrmannova and Knoth 2015) to classify the quality of research groups. Rocha and Moro (2016) analyse research contribution of individual authors on established links between communities as a measure of influence and productivity of authors.

All of these methods inherently estimate the quality of research in retrospect, with semantometrics one could overcome this issue as evaluated in “[Information available at publication time](#)” section.

Identification of important referenced papers

Observing content of citing papers enables identification of *important references* of publications: Hou et al. (2011) classify references of papers as closely related or less related based on the number of times they are referenced in a certain paper, additionally they consider the number of references, the referenced paper and the current paper share. While Valenzuela et al. (2015) label references as incidental or influential based on the section of a paper they occur in Zhu et al. (2015) use numerous features such as counts of occurrences of references, similarity of abstracts or context and position of occurrences of references to determine if a reference is influential or not. Pride and Knoth (2017) state that the number of in-text occurrences of particular references as well as abstract similarity are the most descriptive features in the identification of influential references.

These approaches are directly linked to semantometrics as they all underline the varying importance of different references for publications and even conclude that the majority of referenced papers is not influential at all. Semantometrics does not consider these differing impacts but should do so as described in future work in “[Conclusion and future work](#)” section.

Citation behaviour analysis

Citation behaviour analysis tries to explain the distributions of citations and references and properties associated with papers based on these counts. Price (1965) defines publications which have at least 25 references as review-type papers and publications which received at least four citations in a single year as classics. He observed citations and references in the context of time and described a span of ten years after publication as the major period in which a paper is cited. Classics as defined by Price are roughly equivalent to seminal papers. These definitions are used in “[Description and discussion](#)” section to give quantified insight in our dataset.

The *citation half-life* describes the rate of obsolescence of a corpus of research papers (Burton and Kebler 1960). Recently, the half-life of corpora increasing has been observed (Davis and Cochran 2015; Martín-Martín et al. 2016).

Citation life cycle analysis can be seen as a subtype of general citation behaviour analysis and as an alternative of the half-life referring to single publications. Differences in citedness of scientific papers can be perceived in relation to their age (Seglen 1992). A citation life cycle is the period of time in which a publication is cited. Different patterns can be found in these cycles by which papers can be clustered. Avramescu (1979) describes five citation frequency curves, of which a continuous steady low number of new publications as well as two curves with initially high amount of citations followed by a decline over the succeeding years appear the most in his dataset. While Smith Aversa (1985) found the two clusters *delayed rise - slow decline* and *early rise - rapid decline*, Cano and Lind (1991) observed the patterns *Type A* and *Type B*. Papers of *Type A* accumulated citations fast in the first few years after the publication but the citation frequency declined after this incipient peak. Publications of *Type B* had a moderate initial amount of citations in the first six years but afterwards, the number of new citations per year was steady. Later, Aksnes (2003) found several patterns for highly cited papers whereas *early rise - rapid decline* and *medium rise - slow decline* make up the highest share of his analysed papers.

As semantometrics utilises papers citing publications, awareness of citation life cycles is beneficial for constructing datasets as seen in “[Contained data](#)” section. Incomprehensive representation of these cycles in a dataset could lead to biased results of algorithms working on it which was looked at in “[Classification for different years](#)” section. Citation behaviour analysis can be used for the interpretation of evaluation results as was done in “[Classification on number of references and citations](#)” section.

Document vector representations

In order to enable the conduction of computations on documents, these documents have to be represented as vectors. There are several approaches of transforming documents to vectors, some rely on semantic information while others regard the topical composition of texts.

Semantic information of documents can be accessed by algorithms describing input data as vectors abstracting from words to meanings behind terms. The resulting vectors are dependent on the context a word occurs in. Relations between words are learned by observing the surrounding tokens in a document. Doc2Vec (Le and Mikolov 2014) embeds all words that are presented in the training documents as vectors in a distributed space of fixed dimensionality. BERT (Devlin et al. 2019) also learns language representations from text data but produces varying vectors for the same word in different contexts. Amongst others, these two algorithms can be used to represent textual data as numeric features in a vector of certain length with which NLP tasks can be tackled.

Topic modelling tries to describe the topical composition of documents with a probabilistic model. A number of topics is fixed, then the topic proportions of every document in the collection are calculated as well as the word probabilities for every topic (Blei et al. 2003). LDA is a widespread, basic topic model (Blei et al. 2003). More complex topic models which incorporate authors’ research interests (Rosen-Zvi et al. 2004), temporal aspects in topic developments (Blei and Lafferty 2006) or citation information used

for circulation of topics (Dietz et al. 2007) tend to provide better results. Usage of these topic models requires more or other data than usage of LDA and is more computationally expensive.

Doc2Vec, BERT and LDA are going to be used as document vector representations for semantometrics as described in “[Methodology](#)” section. More sophisticated topic models can be utilised as described in “[Conclusion and future work](#)” section.

SUSdblp dataset

As we strive to tackle the task of determining whether a paper is seminal, survey or uninfluential while investigating its citation network, we require a suitable dataset. Current datasets (Herrmannova et al. 2017; Kreutz et al. 2019) are only designed for binary classification tasks which leaves us with the need of constructing a new one, that is the Seminal-UninfluentialSurveyDBLP (SUSdblp) dataset.

In this chapter, first we introduce the new SUSdblp dataset and describe the contained data. Afterwards we broach the issues of numbers of references and citations as well as years of publication for the three classes. Hereinafter, sub-fields contained in the dataset are observed before we describe and discuss the assumptions used in the construction of the dataset and also shed light on alternative methods for the generation of such a dataset.

Introduction

The SUSdblp dataset contains 1980 publications and is an extension of the SeminalSurvey-DBLP dataset (<https://doi.org/10.5281/zenodo.3258164>).¹ One third of the publications are seminal (referred to as papers from class c_0), one third of the papers are surveys (referred to as papers from class c_1) and another third of the documents are uninfluential (referred to as papers from class c_2). All works are from the area of computer science and adjacent fields as they are contained in dblp (Ley 2009). For seminal publications, entries published in conferences attributed as A* at the CORE Conference Ranking (<http://www.core.edu.au/>) CORE2018 such as *SIGIR*, *JCDL* or *SIGCOMM* were collected as publications often cited (and thus important) tend to appear in high-impact venues (Aksnes 2003). We assume papers published in a seminal venue as attributed by the CORE rank are seminal themselves, or they would not have been accepted for such a venue, even if they have not yet accumulated large amounts of citations. This might be a strong assumption, as not every paper from an A* conference is seminal and seminal papers can also appear in other venues, but is a simplified approximation of truly seminal papers. Surveys were extracted from *ACM Computing Surveys*, *Synthesis Digital Library of Engineering and Computer Science* and *IEEE Communications Surveys and Tutorials*. These venues are specialised in solely publishing reviews. Every paper of class seminal and survey has at least ten citations and references. Uninfluential papers are gathered from a number of venues attributed as C at the CORE Conference Ranking (<http://www.core.edu.au/>) CORE2018. They have an arbitrary number of references which in our case surpasses five but their number of citations lies between five and ten.

¹ The SUSdblp dataset can be found at <https://zenodo.org/record/3693939#.Xl0cF0oxIEa>

Table 1 Numeric description of the SUSdblp dataset

	# Papers in class	# References (X)			# Citations (Y)		
		#	Max	Avg.	#	Max	Avg.
Seminal	660	20,858	154	31.6	50,397	1370	76.4
Survey	660	29,366	186	44.5	51,082	1365	77.4
Uninfluential	660	6629	33	10	4263	10	6.5

Contained data

For each of the papers, the citing and referenced papers were collected. Citation information and abstracts from the AMiner dataset (Tang et al. 2008) were joined with dblp data to make sure they were also from computer science or adjacent domains. The join was based on matching DOIs of dblp papers with ones from AMiner or paper title and author name matches where DOIs were not present. Full texts are not included in the AMiner dataset. Citing and referenced papers not contained in dblp were omitted so the number of links for the papers might not necessarily represent the number of linkages which a paper received in the real world. For every paper, its year of release is also enclosed. The newest publications (as P and Y) contained in the dataset are from 2017 so the citation life cycle of several publications might not be completed yet. Considered publications for P , X and Y needed to have a length of at least ten terms in their combined title and abstract (in some cases the abstract was not present). The average length of the combination of title and abstract is 172.25 terms for seminal publications, 173.03 terms for surveys and 149.57 terms for uninfluential papers.²

For all textual content, punctuation marks were omitted and lower case was used. A stemmed (S) and an unstemmed (U) version of the dataset are provided, the stemmed version contains 82,916 unique terms in the textual components while the unstemmed version holds 113,730 distinct words. The Porter stemmer (Porter 1980) was used to create the stemmed version of the dataset.

For all papers in X and Y , the number of citations the publication received from publications from the area of computer science is also contained. Additionally, for these papers a field and time normalised citation count is included in the dataset to provide the possibility of assessing overall importance of these papers.

Number of references and citations

The SUSdblp dataset is engineered to provide similar numbers of citations and references for publications of classes seminal and survey. Including this characteristic allows for methods working on this dataset to focus on hard cases. In general, seminal papers are cited numerous times while the average survey is not. Surveys typically reference a multitude of papers. Reproducing this scenario would lead to a majority of easy decisions when deciding on the class of a publication and thus divert from more challenging cases such

² Influence of the different abstract lengths on classification accuracy is evaluated in the Appendix 6.

Fig. 2 Distribution of number of references and citations for seminal (blue circles), survey (orange triangles) and uninfluential (green crosses) publications. (Color figure online)

Fig. 3 Histograms of number of seminal (blue), survey (orange) and uninfluential (green) publications over the years. (Color figure online)

as highly cited surveys, seminal papers referencing a multitude of publications or seminal papers which have not yet gained lots of citations. Such fringe cases would not occur often and presumably would therefore be neglected by most algorithms. Instead we decided to focus on such instances.

The total number of unique publications contained in the dataset is 129,443. This number includes all publications P as well as their referenced papers X and citing papers Y . Table 1 shows statistics regarding the number of citations for each type of paper. Each of the seminal and survey publications has at least ten citations and references. As the increased amount of references is assumed to be a feature of survey papers compared to seminal publications, the average and total number of references is higher and thus our dataset is unbalanced in this aspect. All papers contained in the set of uninfluential publications have between five and ten citations, the number of references was not restricted for them. Figure 2 shows the distribution of reference and citation cardinalities for all papers of groups seminal, survey and uninfluential from the dataset. Numbers of citations are distributed rather homogeneously between the classes seminal and survey, but for references, differences in the distributions can be observed. While there are fewer publications with few references for surveys, a gap in the number of references from 40 to 50 can be seen for seminal papers. Distributions of numbers of references and citations for class uninfluential highly differ from the other two classes.

Table 2 Average publication years for papers from P , X and Y as well as average distances in years between publications from P and X as well as P and Y for the three classes

	Seminal	Survey	Uninfluential
Avg. year P	2007.13	2006.45	2008.01
Avg. year X	2003.69	2002.24	2003.23
Avg. year Y	2011.66	2009.55	2011.5
Avg. distance years $P&X$	4.97	6.2	5.19
Avg. distance years $P&Y$	4.55	4.89	3.52

Fig. 4 Histograms over number of referenced (left) and citing (right) papers for seminal (blue), survey (orange) and uninfluential (green) publications over the years. (Color figure online)

Publication years

Having equal publication years for papers of the three classes seminal, survey and uninfluential as well as their references and cited publications was no priority in the construction of the dataset, so there are several differences in these features. Considering the primary focus on comparable numbers of references and citations for publications from classes seminal and survey, pursuing comparable distributions of years would have restricted the pool of publications which could be incorporated into the dataset and therefore also the number of contained papers by a considerable margin.

In Fig. 3 the distribution of publication years of seminal, survey and uninfluential papers is depicted. While seminal papers are more common from 2002 to 2009, for the period before 1993 and from 2010 to 2014, the dataset contains more surveys. Between 1995 and 2001 as well as around 2010, the number of publications from these two classes is comparable. The number of uninfluential papers resembles the number of seminal papers between 2006 and 2008.

Table 2 provides the average publication years of papers from P , X and Y for the three classes as well as the average distance between the publications P and their respective referenced and citing papers. We assume that the larger distance between surveys and their referenced papers compared to the other classes might stem from the longer time a publication takes until it is published in a journal compared to the length of periods between submission and publication of papers to conferences. All papers contained in the class of surveys appeared in journals which underlines the validity of this hypothesis.

Figure 4 shows the number of referenced and citing papers associated with the three classes of publications over the years. As the overall number of references is considerably higher for surveys, the higher numbers of references per year for surveys were expected. The number of references and citations is notably lower for uninfluential papers. It is

Fig. 5 Percentages of top five topics (and all others) for publications P from classes seminal, survey and uninfluential. (Color figure online)

observable that reference distributions differ greatly. Surveys reference more publications from years preceding 1990 and after 2003 compared to seminal papers. Between 1990 and 2003, the numbers of references from these two classes are comparable. The number of citing papers is almost identical for papers from the two categories, but the distribution over the years differs slightly. Surveys were cited more often before 2003. From 2003 to 2008 as well as from 2013 onwards the numbers of citing papers are comparable for the two classes. Between 2009 and 2012, the number of publications citing seminal papers has increased.

To test whether publication years as well as distances in years in a citation network are equally distributed, Kruskal–Wallis H tests for independent samples with $p = 0.05$ were conducted as requirements for standard statistical analysis were not met. The publication years of papers P , those from X and Y of works from the three classes are significantly different from each other. Publication years of P of seminal and survey publications are the only ones which are not significantly different from each other. Distances between papers in P and X , P and Y as well as X and Y are also significantly different for the three classes c_0 , c_1 and c_2 . The distances between seminal and survey papers are not significantly different for the three observed cases. This underlines the validity of the dataset for tackling the previously described hard cases based on differences in publication years even though the even distribution of years was no prerequisite in the construction of the dataset. Distinguishing between c_0 and c_1 based on distances between publication years of a paper P and its referenced papers, P and its citing papers as well as distances between publication years of referenced and citing papers is non-trivial.

Sub-fields and topic distributions in publications

The SUSdblp dataset gathers publications from venues from several sub-fields of computer science: real time systems (conferences RTSS, ISORC), HCI (conferences CHI, ICCV, MMM, COMSWARE, ICCHP), data mining (conferences ICDM, ACII) and software engineering (conferences ICSE, ICGSE, ICCBSS) are the major fields shared between seminal and uninfluential publications. Venues of papers in class survey do not classically target a specific area but encompass all domains of computer science. This observation leads to the

conclusion that the dataset at hand contains multiple areas of computer science, including their potentially differing writing and citing habits.

Figure 5 shows the percentages of the top five topics over all publications P in the three different classes. All other topics are contained in share *other*. Topic distributions were calculated by usage of LDA (Blei et al. 2003) trained on dblp data (Ley 2009) combined with abstracts from AMiner (Tang et al. 2008) with $k = 100$. Percentages of the different topics were added for all papers from a certain group to determine the most prevalent topics. Topics 41, 97, 62 and 93 can be found amongst all classes in the most popular topics. Although their share varies between classes, the sole existence of the topics in the top five topics hints at a topically relatively unbiased dataset.

Description and discussion

The SUSdblp dataset is specifically constructed to tackle tasks by usage of semantometrics as there is no dataset suitable for its evaluation in a ternary setting. This premise prevented the inclusion of publications in class *uninfluential* which have no citations, i.e. really do not influence any following works. A vast percentage of scientific articles is uncited (Seglen 1992) but the papers contained in our dataset do not intend to represent citation distributions found in full corpora. If we included mostly uncited publications in the dataset for c_2 , methods relying on whole citation networks of publications could degenerate as much less features could be extracted from these papers. This aspect is evaluated in “[Truly uninfluential publications](#)” section.

None of the publications of class *survey* are taken from conferences while all papers P in classes *seminal* and *uninfluential* are extracted from conferences. This might lead to longer periods between papers from groups X and P for surveys. For survey papers, it seems far more likely for them to appear in journals than in conferences. If we tried to include publications from conferences in class c_1 , they would have to be collected by hand as automatic methods could only rely on keyword search and manual verification of the assigned label.

The *uninfluential* papers were taken from conferences which were ranked as tier C in the CORE Ranking, the assumption of being not important stems from their venue of publication and their number of citations but might be completely false for some papers as citations might not be mapped correctly or could not be mapped for these publications at all. Publications appearing in a lower tier venue might also be seminal as citation counts are unaffected by prestige of venues (Aksnes 2003; Seglen 1994). Another aspect to consider for papers of class *uninfluential* might be their citation life cycle, maybe they are only on the beginning of theirs, or they are genial work which only receive citations after an initial phase of absent recognition (Avramescu 1979). Papers not cited in a certain year or period might be cited in a following year (Price 1965). Influence of the different years for the classification accuracy in terms of disruption of citation life cycles is evaluated in “[Classification for different years](#)” section.

Of the 660 seminal papers 24 have received a best paper award. Only incorporating publications which received an award would dramatically decrease the size of the dataset as a similar distribution of referenced and citing works of classes *seminal* and *survey* was prerequisite in its construction.

613 of the seminal papers were cited four or more times in a single year, making them classics (Price 1965). Of the surveys, 504 reference 25 or more papers which is said to be

Fig. 6 Simplified graphical depiction of methodology

an attribute of this type of publication (Price 1965). Of the uninfluential papers, eleven have at least 25 references and 85 are cited at least four times in 1 year.

Another way to construct a dataset fit for tackling the same task would be via an email questionnaire similar to the generation of the TrueImpactDataset (Herrmannova et al. 2017). There are several challenges associated with such a procedure. Conducting a survey is entirely dependent on the participants and their willingness as well as ability to judge publication quality. In this context, problems arise in sampling of subjects for the survey leading to bias as the answers might not be representative (Baltes and Diehl 2016). The dataset resulting from responses would most probably be unbalanced for the three classes and its size would entirely depend on the number of responses. Answers would very likely have to be omitted as the identification of referenced papers would be impossible or they could be from completely different fields. For all answers which could be mapped to real publications, metadata as well as references and citations would have to be extracted from suitable data sources. For comparison, the TrueImpactDataset contains 166 seminal and 148 survey papers which were gathered through 184 responses. The response rate of the study was 13%. The dataset holds papers from 31 distinct scientific fields (Herrmannova et al. 2017). Acquiring a number of publications comparable to the one of the presented dataset while focusing on a single domain in this manner thus would take vast efforts for study conductors. Another factor worth considering is the caused cost of invited subjects even if they did not complete the questionnaire (Baltes and Diehl 2016).

Furthermore, a dataset could be constructed by usage of only papers which received a test of time award for those in class seminal. The other classes could be constructed like they are now. As the number of distinguished publications per year is considerably low and these awards have not been handed out for a long time yet, it would be difficult to construct a sufficiently big dataset out of them. Another drawback using this method would pose the collection of seminal papers. Their titles would have to be scraped from web pages of conferences and then the associated metadata concerning abstracts, citations and references would have to be retrieved from external data sources. Additionally, such a dataset would not hold any recent publications in class seminal.

Methodology

Herrmannova et al. (2018) proposed the usage of citation networks to extract patterns from differences between texts which can be represented by distance features for making assumptions whether publications are seminal or survey. First, document vector representations

(V) of P , its referenced papers X and its citing papers Y need to be generated from a suitable dataset. In a next step, a distance measure (M) is chosen with which distances between publications for every group A to E can be calculated. From these five sets of distances, twelve features are then computed for each set: Minimum, maximum, range, mean, sum of distances in a group, standard deviation, variance, 25th percentile, 50th percentile, 75th percentile, skewness, and kurtosis. Those $12 \cdot 5 = 60$ features are named by concatenating the feature with the group it originates from, e.g. *minA* or *rangeE*. Classifiers are trained on different sets of data which either describe a publication by one feature or multiple features. We (Kreutz et al. 2019) proved the classification on multiple features as useful. For test data, the classification algorithms are then able to determine the class a publication P is most likely to be associated with. Figure 6 displays the simplified course of action from dataset to the accuracy of a classification as described above. In this pipeline, there are several interchangeable parts where different options are available, which are indicated by rectangular boxes in the figure.

As a considerable difference to previous work (Herrmannova et al. 2018; Kreutz et al. 2019), our dataset contains three classes, we modify the binary classification problem to become a ternary classification task. Class c_0 describes seminal publications, c_1 indicates the class containing surveys and c_2 is defined as group of comparably uninfluential papers.

In this work, we observe different aspects of citation networks of publications: words, semantics, topical compositions and publication years. Earlier work only focused on document vectors derived from words (Herrmannova et al. 2018) or showed the helpfulness of simple semantic features (Kreutz et al. 2019). We extend the approach by not only classifying on features derived from one document vector representation but observe the possibility of combining features describing multiple aspects of the dataset to improve overall classification accuracy.

Document vector representations

For document vector representation (V) methods working on words, semantics, topics and distances between publication years in a citation network were utilised. As a method working on words, tf-idf (Salton et al. 1975) is applied. Semantics of publications are depicted by usage of Doc2Vec (Le and Mikolov 2014) (D2V) as well as BERT (Devlin et al. 2019). Topical information of publications is constructed by using LDA (Blei et al. 2003). We refrained from using more complex topic models since we wanted to focus on experimentation of the general usefulness of utilisation of topics in the context of scientometrics. Publication years of papers are extracted from the underlying citation network.

Distance measures

As distance measure (M), cosine distance (COS) is applied as described by Herrmannova et al. (2018). Additionally, Jaccard distance (JAC) is used as a second method as seen in previous work (Kreutz et al. 2019). These two measures are applied on tf-idf, Doc2Vec and BERT vectors. On LDA document representations, Earth Mover's Distance (EMD) is applied. We apply $1 - \text{standard inner product}$ (IPD or inner product distance) on all word, semantic and topical document vector representations. On publication years of papers from the three classes, differences in years in the citation network (DIST) are calculated.

Classification algorithms

The set of selected classification algorithms (Cla) includes logistic regression (LR), random forests (RF), Naïve Bayes (NB), support-vector machines (SVM), gradient boosting (GB), k-nearest neighbours (KNN) and stochastic gradient descent (SGD) as seen in previous work (Kreutz et al. 2019). Herrmannova et al. (2018) applied SVM, LR, NB and decision trees. We wanted to include those classifiers except for decision trees which we omitted as we incorporated random forests and gradient boosting which are ensembles over decision trees and thus are able to outperform them.

Implementation

Python 3.6 and classifiers from scikit-learn (Pedregosa et al. 2011) are used in this implementation. For SVM, multi-class as one-vs.-one is calculated. For LR, GB and SGD, multi-class is calculated as one-vs.-all. Implementations for kurtosis, skewness and Wasserstein distance are used from scipy (Jones et al. 2001). Gensim (Řehůřek and Sojka 2010) is used for the Doc2Vec as well as the LDA implementation. For the generation of BERT document vector embeddings, the PyTorch (Paszke et al. 2017) framework is used. For statistical analysis, SPSS 26 is used.

As dataset, the SUSdblp dataset is used in a stemmed and unstemmed version. On this dataset, we constructed the document vector representations: the tf-idf values are computed on the 129,443 unique publications in the stemmed or unstemmed SUSdblp dataset, abstracts of all citing and referenced papers were included in the calculation of term frequencies. Vectors computed on the stemmed dataset consist of 82,916 dimensions while vectors representing the unstemmed dataset consist of 113,730 dimensions.

Weights for Doc2Vec are generated by usage of the English Wikipedia corpus from 20th January 2019. We refrained from using Doc2Vec on a stemmed corpus as this preprocessing is no prerequisite for achieving good results (Le and Mikolov 2014). The Doc2Vec model was trained so that resulting vectors consist of 300 dimensions. This size was proposed by Lau and Baldwin (2016) for general-purpose applications.

A pretrained uncased BERT base model was also used to create document embeddings. It was also only used on unstemmed publications. The BERT implementation used is only able to process input vectors of at most 512 tokens (Devlin et al. 2019), documents were cut at places where punctuation marks would have been or after half of the tokens if sentences were still too long. A sliding window was used to always input two consecutive sentences to maintain as much context as possible. The model consists of overall twelve hidden layers each having 768 features. The last four layers from these twelve layers were concatenated for each token and averaged over all tokens to receive vectors of length 4 layers * 768 features = 3072 dimensions for each publication.

We ran LDA on unstemmed and stemmed titles concatenated with abstracts from all publications contained in the dblp dataset (Ley 2009). Abstracts were extracted from AMiner (Tang et al. 2008). Following this procedure, we ensured the computed topics were from the area of computer science. The number of topics was set to 100 in both cases, resulting in the same number of dimensions for these document vector representations.

In case of years, the publication years of all papers in the citation network were extracted, resulting in one-dimensional vectors for each of the publications.

Removal of high-frequency words from titles and abstracts of publications before construction of document vector representations was out of scope for this work but we assume a

possible increase in classification accuracy from conducting this pre-processing step as related tasks typically benefit from doing so (Schofield et al. 2017).

The implementation of our approach including usage instructions can be found at GitHub under <https://github.com/dbis-trier-university/Semantometrics>.

Evaluation of the approach

Our approach is evaluated by observance of different classification modalities. Classifications were conducted based on single features and all features. Additionally, classification on combinations of features derived from multiple aspects of the dataset is observed. A following experiment evaluates the performance of the approach when trying to classify truly uninfluential publications without citations. Reclining on this experiment, we predict the class of a paper based on semantometrics derived from information that is available as soon as a paper is published.

All accuracies (Acc), the 95% confidence interval ($\pm$) and F1 scores (F1) are rounded to four decimal places. Values have been calculated by usage of ten-fold cross-validation if not specified otherwise. Accuracies marked bold in our result tables are the highest ones out of the experiments displayed in the respective table for the different settings. In all of our experiments, there is no need for a development set as we do not perform hyperparameter tuning, data the model gets trained with is always different from the data it is evaluated against.

If more than one classification algorithm achieved the highest accuracy, the algorithm with the highest F1 score is mentioned in a table. For all significance tests, we use a p -value of 0.05. Statistical analysis is conducted on the accuracies extracted from the ten folds of the cross-validation for each model. Normal distribution of values is evaluated by usage of Kolmogorov-Smirnov (Massey 1951) and Shapiro-Wilk (Shapiro and Wilk 1965) tests. Homogeneity of variances is tested with Levene's test (Levene 1960). If an independent two-sample t -test is used, data is normally distributed in the two groups and variances are homogeneous. If a Welch t -test is conducted, data is normally distributed in the two groups but variances are not homogeneous. If a Kruskal–Wallis H test is used, data is not normally distributed in the different groups or variances are not homogeneous. If ANOVA is used, data in the (more than two) different groups is normally distributed and variances are homogeneous.

The dataset used in this evaluation is the SUSdblp dataset.

Single features

The first experiments consider the whole citation network of publications for a classification based on a single feature. Each of the 60 features derived from different aspects of a publication is used on its own as input for the classification algorithms. This evaluation delivers a baseline and enables us to relate the results to previous ones (Herrmannova et al. 2018; Kreutz et al. 2019) in a ternary setting.

At first, words, semantics, topics and publication years of seminal, survey and uninfluential papers are observed. Words of publications stem from their titles and concatenated abstracts. The tf-idf vectors of stemmed and unstemmed publications are combined with COS, JAC and IPD as distance measures. Semantics of papers were derived by applying Doc2Vec and BERT on concatenated titles and abstracts. Distances between resulting vectors of papers in the different groups were computed by using the same distance measures as before. For topics of publications, LDA was applied on stemmed and unstemmed

Table 3 Best classifiers dependent on distance measures for all document vector representations

V	M	Cla	F	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	COS	LR	<i>sumA</i>	.6879 ($\pm$.0414)	.6818	.6318	.447	.9848
tf-idf S	COS	LR	<i>sumA</i>	.6884 ($\pm$.0421)	.6824	.6333	.447	.9848
D2V U	COS	LR	<i>sumA</i>	.6904 ($\pm$.0438)	.6845	.6394	.45	.9818
BERT U	COS	LR	<i>sumA</i>	.6889 ($\pm$.0472)	.6833	.6364	.453	.9773
LDA U	IPD	GB	<i>sumD</i>	.6939 ($\pm$.0723)	.6989	.2727	.5076	.9182
LDA S	IPD	LR	<i>sumA</i>	.6874 ($\pm$.0430)	.6815	.6303	.4485	.9833
Years	DIST	GB	<i>sumA</i>	.6879 ($\pm$.0372)	.6856	.0848	.9121	.9742

For the classification algorithm achieving the highest accuracy per combination, the single feature (F), the corresponding F1 score as well as accuracies for the three classes c_0 , c_1 and c_2 are displayed

Fig. 7 Box plot of value distribution of feature *sumD* derived from differences computed with inner products of unstemmed LDA document representations for seminal, survey and uninfluential publications in their citation networks

documents. Distances between topical distributions then were calculated by usage of EDM and IPD. Lastly, distances between publication years were computed.

For each of these combinations of document vector representation and distance measure, the seven classifiers were applied which each of the 60 features as input. The best results per document vector representation can be found in Table 3.³ No significant differences for the seven methods were found when applying a Kruskal–Wallis H test.

The best results can be achieved by usage of LDA on unstemmed documents combined with inner product distances and gradient boosting as classifier. The fact that feature *sumD* is the most descriptive feature in this setting is very promising, as this feature describes the sum of distances between references. This information is already available at the time when a paper is first published. Figure 7 shows the distribution of values of feature *sumD* for the three different classes.

In general, a highly descriptive feature regardless of document vector representation and distance measure is *sumA* which describes the sum of distances between referenced and citing papers. Figure 8 gives an overview of the distribution of *sumA* for the three classes at hand for tf-idf on stemmed documents and cosine distance. The relatively low values for papers from class uninfluential might be explained by their comparably low number

³ In-depth results on all combinations of document vector representation as well as distance measure can be found in the Appendix 1.

Fig. 8 Box plot of value distribution of feature sumA derived from differences computed with cosine distance of stemmed tf-idf document representations for seminal, survey and uninfluential publications in their citation networks

of references. Surveys tend to have a higher value with this feature than seminal publications which could be explained by the inherently higher number of references per paper for reviews. The best classifier might group seminal and survey papers together, resulting in the low accuracy for c_0 . A binary classifier on features derived from distances between publications which only distinguishes seminal and survey publications might lead to better results for our core task.

Compared to Herrmannova et al. (2018), we achieved an increased accuracy of .0042 which mainly only stems from the introduction of c_2 . Publications from the two other classes cannot be identified reliably using only a single feature. While features from group B , D and C were achieving good results for Herrmannova et al., they do not report on features from group A to be helpful. Kreutz et al. (2019) were able to surpass our results by .0493. For them, features from group A were also not helpful in determining the class of a paper but they also found features from group D to be performing quite reliably. Contrasting earlier work, here we introduced a third class into the classification task, it is quite comprehensible for our single features to not be as descriptive as those used in a binary setting.

Multiple features

For the following experiment, the complete citation network of every publication is used for the classification procedure: all 60 features are used for the classification. Features derived from all combinations of document vector representations and respective distance measures are used as input for the seven classifiers.⁴ This evaluation enables us to relate our results to the binary classification task on seminal and survey publications performed by Kreutz et al. (2019).

Table 4 shows the detailed accuracies for all combinations of distance measure and document vector representations.⁵ Significant differences were found in the seven methods when tested with a Kruskal–Wallis H test. The two tf-idf were significantly different from

⁴ An inferior second experiment using only the 33 features, which were found to be significant by Herrmannova et al. (2018) can be found in the Appendix 3.

⁵ Results on all different combinations of document vector representation as well as distance measure can be found in the Appendix 2.

Table 4 Best classifiers dependent on distance measures for all document vector representations

V	M	Cla	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	COS	GB	.8414 ($\pm$.0356)	.8415	.7894	.7515	.9833
tf-idf S	JAC	GB	.8439 ($\pm$.0371)	.8439	.7848	.7621	.9848
D2V U	COS	GB	.8525 ($\pm$.0309)	.8526	.7742	.8015	.9818
BERT U	IPD	GB	.8646 ($\pm$.0569)	.8647	.806	.8	.9879
LDA U	IPD	GB	.8601 ($\pm$.0367)	.8601	.8167	.7788	.9848
LDA S	IPD	GB	.8475 ($\pm$.0266)	.8477	.7803	.7773	.9848
Years	DIST	RF	.8747 ($\pm$.0401)	.8743	.8439	.7985	.9818

For the classification algorithm achieving the highest accuracy per combination, the corresponding F1 score as well as accuracies for the three classes c_0 , c_1 and c_2 are displayed

usage of years. For unstemmed tf-idf, the effect size was small ($r = .0074$), for stemmed tf-idf, effect size was medium ($r = .0103$).

When using all features derived from distances in publication years in a citation network, the best results (accuracy of .8747 with F1 score of .8743) can be achieved by usage of gradient boosting. The highest accuracy from the single feature experiments was surpassed by .1808. Accuracy values for the three classes are all relatively high, almost all unimportant papers can be reliably classified. Here, accuracy for seminal papers is higher than the accuracy for surveys. Equally distributed years of publications were no precondition in the generation of the dataset but here, instead of years, distances between years were observed. The differences between publication years already mentioned in “[Publication years](#)” section might be a characteristic of papers of the three classes. Nevertheless, the high descriptiveness of features derived from years might be unrepresentative of reality.

For the best performing combination, the five features with the highest influence on estimator performance are *sumA*, *sumE* with importance values of over .12 as well as *sumA*, *sumB* and *sumD* with importance values of around .006 for all ten instances of the random forest classifier in the cross-validation process.

Even though results for distances between years can be influenced by the construction premises of the dataset, applying gradient boosting on features derived from BERT embeddings and inner product distance leads to an accuracy of .8646 (F1 score .8647). Accuracies for the three classes are also considerably high. Here, no dataset topic bias should artificially boost the results based on distances between semantics of papers in their whole citation network. Papers stemming from different areas in computer science might have diverse citation practices (Cronin and Meho 2006; Jin et al. 2007; Seglen 1992, 1997; Shi et al. 2010) and therefore reference papers with possibly domain dependent topical compositions. These variability cannot be omitted in the automatic creation of a dataset but might hint at the approach’s suitability in terms of applicability on topical diverse citation networks.

Building upon results from Kreutz et al. (2019), we were able to improve the accuracy by .0792 when using featured derived from distances between publication years and achieved an increase of .0646 when utilising all features derived from BERT embeddings. Looking at the two classes c_0 and c_1 only also leads to higher overall accuracies compared to the best results from the binary classification task.

Combination

Experiments were conducted where all five document vector representations were utilised in concatenation. All possible combinations of stemmed and unstemmed vector representation as well as all combinations of distance measures were observed. This experiment explores the informative power of utilisation of numerous aspects of publications.

The 60 features derived from the first document vector representation with accompanying distance measure were concatenated with the following 60 features derived from document vector representations with fitting distance measure. This procedure resulted in the construction of vectors of length 300 features (5 document vector representations * 60 features) for each paper of the three classes.

An accuracy of .9247 ($\pm$.0446, F1 .9249, Acc_{c_0} .9015, Acc_{c_1} .8894, Acc_{c_2} .9833) was achieved for usage of features derived from inner product distances between stemmed tf-idf vectors, unstemmed LDA, Doc2Vec and BERT document representations as well as distances between years.⁶ Gradient boosting produced the highest accuracies. Results from the multi feature experiment were surpassed by .05, compared to the single feature baseline, improvements of .2308 were achieved in this experiment.

Comparison of the best performing single document representation (years) with usage of two, three, four or all five document vector representations in concatenation with ANOVA showed significant differences between the five groups. With Bonferroni correction (Fuhr 2017) and Scheffé's method, significant differences between utilisation of the five-vector representation approach and all other methods were found.

Truly uninfluential publications

In the SUSdblp dataset, only publications are contained that have at least five citations, as this is only an approximation of truly uninfluential publications which did not yet accumulate any citations. In the next experiment, the capacity of the approach to recognise such papers is evaluated.

As truly uninfluential publications, 112 publications from conferences which are not contained in the CORE Conference Ranking (<http://www.core.edu.au/>) CORE2018 were chosen at random from dblp. These publications have no citations but each of them has at least 5 references with concatenated titles and abstracts of length ≥ 10 tokens. Doc2Vec vector representations of the citation networks are constructed and cosine distance is applied to derive the 60 features. There are no values contained in groups A, C and E (which equals a value of 0 for all twelve features in these groups) if no citations exist for a publication.

When using all 60 features conjointly, the accuracy was 1 when using SVM. Due to this result we assume that the approach is robust to application on truly uninfluential publications which did not yet accumulate any citations.

⁶ Results of inferior experiments using two, three or four different document vector representations in concatenation can be found in the Appendix 4.

Table 5 Classification accuracy and F1 scores for features of groups *B* and *D* derived from all different vector representations from the SUSdblp dataset for the best performing distance measure and classification algorithm

V	M	Cla	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	IPD	GB	.7131 ($\pm$.0499)	.7159	.6606	.7273	.7515
tf-idf S	JAC	GB	.703 ($\pm$.0558)	.7056	.6364	.7152	.7576
D2V U	JAC	GB	.6838 ($\pm$.0417)	.685	.5833	.7197	.7485
BERT U	IPD	GB	.8152 ($\pm$.0652)	.8155	.7364	.7288	.9803
LDA U	IPD	GB	.7192 ($\pm$.0625)	.7211	.6621	.7697	.7258
LDA S	IPD	GB	.7071 ($\pm$.0599)	.7096	.6409	.7379	.7424
Years	DIST	GB	.7268 ($\pm$.0535)	.7281	.6773	.7318	.7712

Information available at publication time

When a paper is first published, only its referenced papers and the content of the publication are available. It has not yet gained any citations. In this experiment, we try to predict a class solely based on features derived from groups *B* and *D* to simulate this situation. Table 5 shows the best distance measure as well as accuracies and F1 scores resulting by usage of the best performing classifier for each document vector representation. Significant differences were found for all seven methods when tested with a Kruskal–Wallis *H* test. The BERT option differed significantly from the two tf-idf methods, Doc2Vec and stemmed LDA. The effect sizes were large for all of them (unstemmed tf-idf .7703, stemmed tf-idf .9673, Doc2Vec 1.2111, stemmed LDA .864).

The highest accuracy of .8152 (F1 score .8155) was achieved by using features derived from inner product distances of BERT embeddings.

This experiment underlines the usefulness of semantometrics for real information retrieval systems as a means of automatically estimating the quality of a publication. Instead of relying on the venue it appeared in or waiting for citations to accumulate in order to make assumptions on the importance of a paper, an independent prediction could directly assess the quality. This property of semantometrics is also especially helpful for preprints of papers published in arXiv where no information on a venue rank is available.

Discussion

The highest overall accuracy when using single features of .6939 (F1 score .6989) was achieved by usage of sumD extracted from inner product distances between unstemmed LDA document vector representations. This feature is already computable at the time a paper is first published, as it encodes the sum of differences between references of *P*. In general, for all other combinations of document vector embeddings and distance measures, sumA was the most descriptive feature resulting in accuracies around .69. Feature sumA contains indirect information on the number of citations and references of a publication as it describes the sum of distances between vector representations of all combinations of citing and referenced papers. The more papers are linked to *P*, the more distances are computed and the higher the value of sumA becomes. Typically, uninfluential papers are cited few times, so this feature would have an overall low value. Both seminal papers as well as surveys tend to be cited often so their values should be higher than the ones for uninfluential publications. In general, surveys have more references than seminal papers, so more

distances between vector representations of papers can be computed (and summed) in their citation network, leading to an even higher value for this feature.

Compared to the best single feature accuracy in the binary classification task of .7432 from our former work (Kreutz et al. 2019), our current accuracy is lower by .0528. Contrasting the prior work, we introduced the third class uninfluential into the classification problem so it is not surprising that we were unable to achieve such a high accuracy while only using single features.

Usage of all features resulted in the highest accuracy of .8747 (F1 score .8743) for distances between years in a citation network and application of random forests. Overall, the highest accuracies for all other combinations of document vector representations and distance measures lie between .84 and .86. Compared to the single feature variant, accuracies increased for all cases. In general, using all features resulted in +.1808 in accuracy compared to the best single feature variant.

The best accuracy from our former work (Kreutz et al. 2019) in a multi feature setting with binary classification was .8015, which we were able to surpass by .0732 by introducing the usage of more diverse features derived from citation networks. The newly introduced third class is not entirely responsible for the higher value, as the accuracies for c_0 and c_1 are also very high (.8439 and .7985). Due to this, one cannot argue the third class artificially boosted the accuracy as it might be too easy to distinguish uninfluential papers from seminal and survey ones.

Combining features derived from distances between all document vector representations leads to an accuracy of .9247 (F1 score .9249). The performance of the multi feature setting was increased by .05.

A Kruskal–Wallis H test showed significant differences for the best performing model based on single features, all features and the combination of features derived all five document vector representations. The three models are all significantly different from each other.

Classification on truly uninfluential publications achieved a perfect accuracy.

When predicting classes of publications based on information which is already available as soon as they are published, an accuracy of up to .8152 was reached by using inner product distance on BERT embeddings of references and the publication. This finding is highly promising as this method could easily be applied in real world scenarios in context of information retrieval systems.

Herrmannova et al.'s (2018) best performing algorithm was Naïve Bayes, in our experiments usage of logistic regression (single feature setting), random forests (multi feature setting) as well as gradient boosting (single/multi feature setting, combination of feature sets) typically achieved the highest accuracies.

As the absence of label noise cannot be guaranteed, it is unclear if the upper bound for accuracies on this dataset truly is 1. With Herrmannova et al.'s approach (Herrmannova et al. 2018) and our extensions to it, we were able to approximate this bound.

Evaluation of the dataset

As we already used the SUSdblp dataset in the evaluation of our approach, we now observe the presented dataset to substantiate the reliability of our results.

The robustness of the dataset is accessed as well as differences in classification performances when using the dataset up to or from different years.

Robustness of dataset

In a first experiment, the robustness of our dataset is evaluated. As the dataset is automatically constructed and based on some strong assumptions, results might be skewed by the generation process or the underlying biased dataset; citations and references are entirely dependent on the dataset they are extracted from (Seglen 1997; Moed 2002). In the construction of our dataset, citing and referenced publications were omitted from inclusion into the dataset if the concatenated titles and abstract were less than ten tokens. This did happen several times when dblp entries could not be mapped to entries from Semantic Scholar and therefore no abstract was found. Other cases when citing and referenced papers were omitted are instances in which the respective publications were not contained in dblp because they were out of scope. Due to these factors, the robustness of the SUSdblp dataset with respect to slight variations in references and citations of publications is evaluated to estimate the overall reliability of our findings with regards to bias in the underlying data.

On unstemmed Doc2Vec document representations, cosine distance is applied. The document vector representation was chosen as the calculation of it is relatively fast; the distance measure was chosen arbitrarily for the comparisons.

In a first experiment, one citation and one reference from each seminal, survey and uninfluential publication are omitted randomly in the calculation of distances for the five groups. In a second evaluation, five citations and references from each seminal and survey publication as well as two references and citations from all uninfluential publications were ignored when deriving features.

When classifying on single features, both omitting modes produce results comparable to the single feature baseline. Feature sumA was the most descriptive one in the best cases. No significant differences were observed when applying ANOVA.

Classification on all 60 features was significantly different for the three considered groups when analysed with ANOVA. Experiment one resulted in an accuracy of .852 ($\pm$.0341, F1 score .8522, Acc_{c_0} .7788, Acc_{c_1} .7955, Acc_{c_2} .9818) when using gradient boosting, which is $-.0005$ compared to usage of the unaltered dataset. For the second omitting mode, an accuracy of .8136 ($\pm$.0508, F1 score .8113, Acc_{c_0} .6985, Acc_{c_1} .7712, Acc_{c_2} .9712) was computed with GB, which is $-.0389$ in accuracy in comparison with the regular citation network.

Due to these slight changes in accuracies throughout the two omitting modes, robustness of the dataset is assumed. This leads to the conclusion of reliability for our findings in terms of data source bias in spite of omitting references and citations from our underlying data source.

Classification for different years

The SUSdblp dataset is biased in terms of publication years for papers of the different classes. To estimate the graveness of this bias on the used method which does not directly work on the years but instead uses distances between publication years, we restricted the SUSdblp dataset to only contain information on publications P , X and Y up to or from different years. This should simulate the effects of a possibly abrupt disruption of citation life cycle of publications on classification accuracy as has happened for papers published shortly before 2017 which was the last year from which works are included in the dataset. We observed papers up to and including 2005, 2010 as well as 2015. Additionally, we observe the performance of the algorithm when using papers P only published 2000, 2005,

2010, 2015 or later. This experiment was intended to shed light on effects of different publication years, for example surveys are included in the SUSdblp dataset from much earlier years than uninfluential publications. As document vector representation for these experiments, Doc2Vec was utilised with cosine distance as distance measure on which features were derived from. The performed classifications used all 60 features and was evaluated against the unaltered dataset.⁷

We performed a Kruskal–Wallis H test as the accuracies for the different models were not normally distributed. Although a significant difference can be observed when classifying on the eight datasets (≤ 2005 , ≤ 2010 , ≤ 2015 , ≥ 2000 , ≥ 2005 , ≥ 2010 , ≥ 2015 and unaltered), the datasets do not significantly differ from the unaltered option. This indicates the assumption, that the main results of the evaluation would not change if only publications from certain years and onwards or preceding a point in time are included. Disrupted citation life cycles of publications do not seem to negatively affect classification accuracy.

Discussion

The presented SUSdblp dataset is robust when omitting one or several references and citations. Although the classification performance decreases when doing so, accuracies comparable to the one of the unaltered dataset can be reached.

When using only publications of the dataset which are existent up until certain years, accuracy of classification drops considerably compared to the unaltered dataset. In determining the class of a publication based on all 60 features, all features from group *A*, *C* and *E* can be heavily impaired as the citing papers of a publication might not lie in the observed time frame. In the worst case, all groups are empty and thus returning 0 values for all twelve features. Usually the references of a publication are unaltered as soon as a publication is part of a certain time frame, leaving features from groups *B* and *D* unchanged when compared to the full dataset. In the single feature prediction task, sumA achieved the highest accuracy for Doc2Vec combined with cosine distance, but when restricting the years, this feature naturally seems to be less descriptive.

Overall, the dataset is appropriate for the classification task, no significant differences compared to using the unaltered and time restricted datasets were found.

Alternative approaches for classification

Although we were able to reach accuracies of over .92 by using semantometrics, the approach is computationally expensive as distances between all citing and referenced papers of publications need to be computed. Comparable or better results might be achieved by using simpler methods.

We evaluate classification based on the number of citations and references of papers as well as classification on the representation of the whole citation network of a publication in a single vector. Lastly, we evaluate how well classes of papers can be predicted based on information available at the time of publication.

⁷ A detailed table on classification accuracies can be found in the Appendix 5.

Classification on number of references and citations

In a first experiment, classification solely based on numbers of references and citations of publications was conducted. This information is relatively easy to obtain and does not require vast amounts of computation and thus could provide a good alternative to usage of semantometrics.

The best accuracy of .8126 ($\pm .0462$, F1 score .8131, Acc_{c_0} .7606, Acc_{c_1} .697, Acc_{c_2} .9803) was achieved by usage of GB. The high accuracy for class c_2 is not surprising, if the number of citations is low, publications are part of this class by definition. Even though these results are quite good, semantometrics achieved much higher accuracies. The comparably low results might be caused by the construction process of the dataset. In the dataset, not all references and citations of a publication are contained. Only those which come from the area of computer science and adjacent fields and are thus covered by dblp as well as ones for which a concatenation of titles and (if existent and assignable in the underlying dataset) abstracts has at least ten tokens. Publications not fitting these criteria are not considered in the number of references and citations of publications as they are not contained in the SUSdblp dataset.

The number of citations of publications might not be the number of citations a paper is going to accumulate until it becomes obsolete if the citation life cycle of the paper has not yet ended. As half-lives of corpora are increasing (Burton and Kebler 1960) and the average year of publications contained in the dataset lies between 2006 and 2008 for the three classes, this aspect is worth considering.

Another explanation for the comparably poor performance of this experiment could be the diverse reasons for citing (Garfield 1964) which are not covered by this method or the possibly high number of uncited influences (Garfield 1964; MacRoberts and MacRoberts 2010; Patton et al. 2016).

Classification on document vector representations

In the following experiments, classification based on the citation networks of publications is performed. The same information as with semantometrics is required but instead of calculating distances between P , X and Y , dimensions of these sets of papers are concatenated and used in combination for classification. This method is less complex and might also achieve reliable results.

All dimensions of document vector representations of publications

In this experiment, whole citation networks are used to compare the results to those from semantometrics presented in “Multiple features” section.⁸

We averaged the values of all dimensions of the document vector representations for all references as well as citing papers to obtain vectors of a length which equals the number of dimensions of a certain document representation for each. Vectors generated for references are concatenated with vector representations of the publication before vectors computed for citing papers are appended for every paper. The number of dimensions of each vector

⁸ We also performed experiments on single dimensions of P , all information contained P , X and Y which can be found in the Appendix 7.

Table 6 Classification accuracy and F1 scores for all dimensions from the different one-vector representations of publications with their citing and referenced papers from the SUSdblp dataset for the best performing classification algorithm

V	Cla	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	GB	.948 (± .0252)	.948	.9424	.9273	.9742
tf-idf S	GB	.949 (± .0219)	.949	.9515	.9182	.9773
tf-idf wo SR U	GB	.9343 (± .0399)	.934	.8955	.9273	.9803
tf-idf wo SR S	GB	.9288 (± .0348)	.9285	.8879	.9227	.9758
D2V U	SVM	.8556 (± .0517)	.8556	.8348	.903	.8288
BERT U	GB	.8788 (± .0419)	.8788	.8758	.903	.8576
LDA U	RF	.9182 (± .0225)	.9177	.8773	.9015	.9758
LDA S	GB	.9096 (± .0223)	.9095	.8833	.8848	.9606
Years	GB	.5621 (± .0687)	.5621	.5152	.5758	.5955

equals 3 * length of document vector for the publication. Thus, for tf-idf vectors, classifications become computationally expensive. As the dimension representing the words *survey* and *review* naturally tend to be highly descriptive in our task, we also performed classifications on stemmed and unstemmed tf-idf vectors where we omitted these two dimensions in publications P .

Tf-idf representations of unstemmed documents achieved an accuracy of .948. Using stemmed documents for the generation of tf-idf vectors results in the highest accuracy of .949. The downside to using tf-idf vector representations is the high number of dimensions which need to be considered. Omitting the dimensions for the words *survey* or *review* in the tf-idf vectors of P did decrease the accuracy only slightly, leading to the conclusion of these dimensions not being highly relevant for the ternary classification. Table 6 shows detailed results on accuracies of all one-vector representations, F1 scores and accuracies for the three classes. A Kruskal–Wallis H test proved the significant differences between the observed combinations, particularly the tf-idf versions differed from usage of BERT and Doc2Vec. Between tf-idf variants, no significant differences were found.

The highest values from usage of semantometrics were surpassed by .0243 proving the viability of plain document vector representations. Results from this experiment hint at the superiority of straightforward methods where no information is omitted compared to semantometrics for computer science publications.

An independent two-sample t -test is used to compare the best performing model from semantometrics and the best performing model from the one-vector representations. The two models were found to be significantly different.

The combination of multiple document vector representations might be able to produce even higher accuracies but was out of scope for this paper as we only intended to evaluate semantometrics and compare the approach to a straightforward method of classification.

Information available at publication time

Here, we again want to observe prediction performance based on information which is available as soon as a paper is published. We compare our results with those from semantometrics as described in “[Information available at publication time](#)” section. We construct one-vector representations of the references X as described before and concatenate them with document vectors of the publications P .

Usage of tf-idf vectors results in the highest accuracy of .9323 for unstemmed publications. Similarly, for stemmed papers, an accuracy of .9318 can be achieved. Semantics of

Table 7 Classification accuracy and F1 scores for concatenated on- vector representations of references and publication from the SUSdblp dataset for the best performing classification algorithm

V	Cla	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	GB	.9323 ($\pm .0193$)	.9322	.9318	.8955	.9697
tf-idf S	GB	.9318 ($\pm .0218$)	.9317	.9333	.8864	.9758
D2V U	GB	.8348 ($\pm .0513$)	.8349	.8515	.8318	.8212
BERT U	GB	.8712 ($\pm .0497$)	.8712	.8606	.8955	.8576
LDA U	GB	.8424 ($\pm .0414$)	.8419	.7864	.9	.8409
LDA S	GB	.8369 ($\pm .0391$)	.8365	.7864	.8818	.8424
Years	SVM	.5015 ($\pm .0580$)	.4884	.2773	.5621	.6652

publications seem to provide good results (Doc2Vec Acc .8348, BERT Acc .8712) which is promising as these vectors can be computed quite easily as they only consist of several hundreds of dimensions.

The models are significantly different from each other by a Kruskal–Wallis H test. The two tf-idf variants differ from BERT, LDA and years significantly [large effect size (1.3501 to .7699)]. Additionally, models based on years differ from models using BERT significantly [large effect size (.8631)].

Table 7 provides detailed results on accuracies of all one-vector representations, F1 scores and accuracies for the three classes of vectors for references X concatenated with P . The best accuracy for this task achieved with usage of semantometrics was surpassed by .1171 in this experiment. Classification on numerous concatenations of several document representations again holds the possibility of further improvements of the results but was out of scope for this paper.

Application of a Welch t -test showed significant differences between the best performing semantometrics approach and the best performing model when utilising one-vector representations of publication data, which is available as soon as a paper is published. An independent two-sample t -test on the BERT model from this experiment and the best performing one from semantometrics on information available at publication time also showed significant differences.

Using this method allows for the prediction of the quality of a publication without having to wait for citations to accumulate. The straightforward approach produces more reliable results than semantometrics.

Discussion

Using the plain number of citations and references in the classification task only resulted in an accuracy worse than the ones achieved when using semantometrics. This might be owed to the creation process of the dataset and its inability to reflect the real number of references and citations of publications.

Classifying on document vector representations seemed to be more feasible. In this scenario, no features are artificially constructed but instead, all information present is used. This prevents data potentially being relevant for the classification from being omitted.

Representation of the whole citation network in a single vector representation on which classification was performed achieved the highest overall results of .949 in accuracy for tf-idf vectors of stemmed publications. For all three classes, accuracies of over .91 were achieved, which indicates that the usage of semantometrics does not provide an advantage in this case.

When using only information which is available at the time a paper is published, accuracies as high as .9323 can be achieved when using unstemmed documents on which tf-idf vectors are constructed. This value is .1171 higher than the best accuracy which can be achieved while using features derived from semantometrics under the same premise, the difference is significant. As tf-idf vectors suffer from high dimensionality, usage of information derived from semantics of references and publications in form of BERT vectors could solve this issue. An accuracy of .8712 is achieved, which is +.056 compared to the highest value for utilisation of semantometrics. This method is highly relevant as it is able to predict research quality in real time compared to the approaches which are only applicable in retrospect as they rely on the accumulation of citations.

Our experiments show the significantly superior performance of one-vector representations compared to usage of features derived from semantometrics for our dataset. The SUSdblp dataset is restricted on papers from the area of computer science and adjacent fields as all observed publications are contained in dblp. Using other data sources might lead to different or contrasting results as citation behaviour in computer science is different from other domains, where typically papers from a narrow community are referenced (Shi et al. 2010). In the SUSdblp dataset, some references or citations which exist in reality are not contained. If linked papers were not found in dblp or their combined title and abstract was less than ten tokens, they were not considered in the construction of the dataset. Although this does not represent the real world, the restriction might be a reason for the robustness of the dataset. Our dataset holds properties of a realistic use case of estimation influence of a publication. Only papers which stem from the same discipline as the one to classify are considered. If a publication is referencing or cited by papers from other domains, they seem to be irrelevant in accessing the quality of a paper for the area of computer science.

Conclusion and future work

In this work, the two main tasks of classification a publication with its complete citation network as seminal, survey or uninfluential as well as quality prediction of new papers which did not yet receive citations were observed: We dissected the classification of publications in their citation network as seminal, survey or uninfluential papers based on semantometrics derived from our proposed SUSdblp dataset which is publicly available. We used words, semantics, topical composition and publication years as different aspects of publications and calculated distances in their publication networks. Extraction and usage of single features from the citation networks leads to a classification accuracy of up to .6904 for feature $\text{sum}D$, which describes the summed up distance between pairs of papers referenced by P , and inner product distance on LDA document vectors. Using all features resulted in a highest accuracy of .8747 if distances between publication years of papers in their citation network are used as base for the feature computation. Combining features derived from multiple aspects of a citation network increased the accuracy up to .9247 when using all five observed embeddings together. Classification based only on data which is available at

the time a paper is first published before it was able to accumulate any citations lead to an accuracy of .8152 when using features derived from inner product distances from BERT document vector representations.

We presented a new dataset, the SUSdblp dataset which contains publication years, concatenated titles and abstracts as well as referencing and citing papers for each of the 660 seminal, survey and uninfluential publications. All papers come from the area of computer science and adjacent fields. Our evaluation suggested the dataset being suitable for the ternary classification task at hand.

When comparing semantometrics to established approaches like classification based on one-vector representations of the citation networks using different document embeddings, a highest accuracy of .949 was reached for tf-idf vectors. Using only information available at the time of publication of a paper, an accuracy of .9323 was achieved, labelling utilisation of semantometrics unnecessary for computer science publications.

The following three key conclusions can be derived: First, semantometrics has high potential in estimating quality of publications, especially new ones which did not yet receive any citations. Second, usage of all information available in a citation network is significantly more potent than application of semantometrics for the two observed tasks. Feature engineering techniques such as semantometrics might perform worse than usage of all information at hand, as potentially useful information is lost. Third, assessment of the quality of a publication for the diverse discipline of computer science can apparently be performed by solely observing referenced and citing papers which are also located in the same area.

We recommend a revaluation of our results on datasets from different or multiple areas to estimate the reliability of our findings in a broader context. Although the SUSdblp dataset spans multiple sub-fields and thus represents a somewhat diverse set of publications, observance of other domains could deliver evidence of the general inferiority of semantometrics compared to more straightforward methods. A thorough automatic evaluation of our dataset or the creation of a manually evaluated dataset with even more publications and full texts spanning multiple research areas would be desirable. A new dataset which purely holds publications which received a best paper award as seminal publications could describe another interesting bibliographic perspective to analyse.

Future work focused on semantometrics could be the incorporation of more statistical features such as entropy (Gillies et al. 2016) contained in the five groups of distances. Automatic feature engineering with deep feature synthesis (Kanter and Veeramachaneni 2015) could produce more descriptive features which in turn might lead to higher accuracies.

As BERT generated better results than Doc2Vec, more sophisticated document vector representations could produce higher accuracies. A semantic representation with fast-Text (Bojanowski et al. 2017) or GloVe (Pennington et al. 2014) could contribute to better results. Topic models such as DTM (Blei and Lafferty 2006) or ATM (Rosen-Zvi et al. 2004) could prove to be more suitable than LDA. Other distance metrics could also be used. It could also help to assign referenced and citing publications different weights as not all referenced papers are equally important for a publication (Patton et al. 2016; Zhu et al. 2015). For example weights based on the number of times referenced and citing papers are cited themselves in the whole dblp corpus or based on the field and time normalised citation count which is also included in the SUSdblp dataset could lead to interesting results.

Lastly, another direction for further efforts could be hyperparameter tuning via grid search or the incorporation of more advanced machine learning algorithms such as gpt-2 (Radford et al. 2019) or even a neural network as classifier. Instead of viewing the task at hand as a ternary classification problem, one could remodel it to become a binary

classification task with usage of an abstaining classifier (Tortorella 2000; Pietraszek 2005) to describe papers which are neither seminal nor survey publications.

Acknowledgements Open Access funding provided by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendices

Appendix 1: Evaluation of the approach: single features

For the following experiment, single features from the citation network of every publication are used for the classification procedure. Table 8 shows the detailed accuracies for all combinations of document embeddings with all corresponding distance

Table 8 Observation of different distance measures for all respective document vector representations

V	M	Cla	F	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	COS	LR	sumA	.6879 (± .0414)	.6818	.6318	.447	.9848
tf-idf U	JAC	LR	sumA	.6869 (± .0431)	.6809	.6303	.447	.9833
tf-idf U	IPD	LR	sumA	.6869 (± .0431)	.6809	.6303	.447	.9833
tf-idf S	COS	LR	sumA	.6884 (± .0421)	.6824	.6333	.447	.9848
tf-idf S	JAC	LR	sumA	.6869 (± .0431)	.6809	.6303	.447	.9833
tf-idf S	IPD	LR	sumA	.6869 (± .0431)	.6809	.6303	.447	.9833
D2V U	COS	LR	sumA	.6904 (± .0438)	.6845	.6394	.45	.9818
D2V U	JAC	SVM	sumA	.6768 (± .0359)	.6661	.6727	.3788	.9788
D2V U	IPD	GB	sumA	.6833 (± .0429)	.6761	.4076	.6818	.9606
BERT U	COS	LR	sumA	.6889 (± .0472)	.6833	.6364	.453	.9773
BERT U	JAC	LR	sumA	.6869 (± .0431)	.6809	.6303	.447	.9833
BERT U	IPD	GB	sumE	.6747(± .0687)	.6801	.6561	.7121	.6561
LDA U	EMD	GB	sumA	.6838 (± .0519)	.6787	.6091	.453	.9894
LDA U	IPD	GB	sumD	.6939 (± .0723)	.6989	.2727	.5076	.9182
LDA S	EMD	GB	sumA	.6859 (± .0388)	.6776	.3909	.7076	.9591
LDA S	IPD	LR	sumA	.6874 (± .0430)	.6815	.6303	.4485	.9833
Years	DIST	GB	sumA	.6879 (± .0372)	.6856	.0848	.9121	.9742

For the classification algorithm achieving the highest accuracy, the single feature (*F*), the corresponding F1 score as well as accuracies for the three classes *c*₀, *c*₁ and *c*₂ are displayed

measures. No significant differences were found between the 17 combinations of document vector representation and distance measure when looked at with a Kruskal–Wallis H test.

Appendix 2: Evaluation of the approach: all features

For the following experiment, all 60 features from the citation network of every publication are used for the classification procedure. Table 9 shows the detailed accuracies for all combinations of document embeddings with all corresponding distance measures. Significant differences between the 17 combinations of document vector representation and distance measures were observed when looked at with a Kruskal–Wallis H test.

Table 9 Best classifiers dependent on distance measures with different D2V and BERT vector representations

V	M	Cla	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	COS	GB	.8414 ($\pm$.0356)	.8415	.7894	.7515	.9833
tf-idf U	JAC	GB	.8323 ($\pm$.0299)	.8324	.7727	.7439	.9803
tf-idf U	IPD	GB	.8298 ($\pm$.0326)	.8299	.7712	.7364	.9818
tf-idf S	COS	GB	.8343 ($\pm$.0304)	.8343	.7742	.7439	.9848
tf-idf S	JAC	GB	.8439 ($\pm$.0371)	.8439	.7848	.7621	.9848
tf-idf S	IPD	GB	.8379 ($\pm$.0308)	.838	.7742	.7561	.9833
D2V U	COS	GB	.8525 ($\pm$.0309)	.8526	.7742	.8015	.9818
D2V U	JAC	GB	.8247 ($\pm$.0445)	.825	.7348	.7561	.9833
D2V U	IPD	GB	.8253 ($\pm$.0350)	.8249	.7076	.7879	.9803
BERT U	COS	GB	.8586 ($\pm$.0575)	.8584	.7788	.8152	.9818
BERT U	JAC	GB	.8561 ($\pm$.0546)	.8562	.7909	.7924	.9848
BERT U	IPD	GB	.8646 ($\pm$.0569)	.8647	.8061	.8	.9879
LDA U	EMD	GB	.8263 ($\pm$.0495)	.8264	.7258	.7758	.9773
LDA U	IPD	GB	.8601 ($\pm$.0367)	.8601	.8167	.7788	.9848
LDA S	EMD	GB	.8187 ($\pm$.0487)	.8183	.7212	.747	.9879
LDA S	IPD	GB	.8475 ($\pm$.0266)	.8477	.7803	.7773	.9848
years	DIST	RF	.8747 ($\pm$.0401)	.8743	.8439	.7985	.9818

For the classification algorithm achieving the highest accuracy, the corresponding F1 score as well as accuracies for the three classes c_0 , c_1 and c_2 are displayed

Appendix 3: Evaluation of the approach: 33 features

For the following experiment, the complete citation network of every publication is used for the classification procedure. Classification algorithms are trained on the 33 features, which were found to be significant by Herrmannova et al. (2018). Table 10 shows the detailed accuracies for all combinations of document embeddings with corresponding distance measures. Significant differences between the 17 combinations of document vector representation and distance measure were discovered when looked at with a Kruskal–Wallis H test.

Table 10 Best classifier dependent on document vector representation and distance measure

V	M	Cla	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	COS	GB	.8293 ($\pm$.0296)	.8297	.7682	.7394	.9803
tf-idf U	JAC	GB	.8293 ($\pm$.0428)	.8295	.7712	.7364	.9803
tf-idf U	IPD	RF	.8288 ($\pm$.0204)	.8294	.747	.7591	.9803
tf-idf S	COS	GB	.8293 ($\pm$.0377)	.8295	.7667	.7364	.9848
tf-idf S	JAC	GB	.8399 ($\pm$.0316)	.8401	.7909	.747	.9818
tf-idf S	IPD	GB	.8409 ($\pm$.0208)	.8412	.7924	.747	.9833
D2V U	COS	GB	.8404 ($\pm$.0268)	.8409	.7667	.7742	.9803
D2V U	JAC	GB	.8227 ($\pm$.0357)	.8228	.7258	.7606	.9818
D2V U	IPD	GB	.8136 ($\pm$.0422)	.8128	.6909	.7682	.9818
BERT U	COS	GB	.85 ($\pm$.0461)	.8498	.7652	.8045	.9803
BERT U	JAC	GB	.8561 ($\pm$.0368)	.8562	.7955	.7864	.9864
BERT U	IPD	GB	.8636 ($\pm$.0544)	.8637	.8015	.8061	.9833
LDA U	EMD	GB	.8066 ($\pm$.0383)	.8057	.6606	.7773	.9818
LDA U	IPD	GB	.8505 ($\pm$.0431)	.8509	.7985	.7727	.9803
LDA S	EMD	GB	.8066 ($\pm$.0373)	.8066	.7182	.7182	.9833
LDA S	IPD	GB	.8318 ($\pm$.0271)	.832	.7667	.7455	.9833
Years	DIST	GB	.8712 ($\pm$.0283)	.8708	.8303	.8	.9833

For the classification algorithm achieving the highest accuracy, the corresponding F1 score as well as accuracies for the three classes c_0 , c_1 and c_2 are displayed

Table 11 Best accuracies dependent on the number of combined sets of features derived from distances between document vector representations

# Combined feature sets	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
2 (BERT, years)	.9096 ($\pm$.0342)	.9098	.8985	.8455	.9848
3 (LDA, BERT, years)	.9192 ($\pm$.0414)	.9192	.8879	.8833	.9864
4 (tf-idf, LDA, BERT, years)	.9237 ($\pm$.0345)	.9238	.8939	.8909	.9864

Appendix 4: Evaluation of the approach: combination

Experiments were conducted where two, three or four document vector representations were utilised in concatenation. Table 11 shows detailed results on accuracies of the best performing combinations for all numbers of combinations.

The combination of two feature sets achieving the highest accuracy of .9091 (F1 score .9093) were Jaccard distances of unstemmed BERT vectors and distances between years. When concatenating three feature sets, the highest accuracy of .9192 (F1 score .9192) was reached when using features derived from inner product distances from stemmed LDA and unstemmed BERT document vector representations as well as distances between years. Combining four sets of features lead to an accuracy of up to .9237 (F1 score .9238). This value was reached by using all features derived from inner product distances of stemmed tf-idf, stemmed LDA and unstemmed BERT document vector embeddings as well as distances between years. Gradient boosting was the best performing classifier in all of these cases.

Appendix 5: Evaluation of the dataset: classification for different years

Table 12 provides classification accuracies for the dataset with publications from P , X and Y until certain years and for publications P which are older than multiple years. Doc2Vec document vector representations and cosine distance were used in these calculations.

Table 12 Classification accuracy and F1 scores using all features derived from Doc2Vec vector representations for the best performing classification algorithm up until or from and after different years as well as the accuracies and number of publication per class

Year	Cla	Acc	F1	Acc _{c₀} (#P)	Acc _{c₁} (#P)	Acc _{c₂} (#P)
≤ 2005	GB	.7755 ($\pm$.0779)	.7763	.7409 (220)	.797 (197)	.7963 (162)
≤ 2010	GB	.7751 ($\pm$.0477)	.7732	.6904 (478)	.7044 (345)	.9156 (462)
≤ 2015	GB	.8331 ($\pm$.0601)	.8321	.7446 (646)	.7824 (648)	.9697 (659)
≥ 2000	GB	.8735 ($\pm$.0337)	.8736	.8156 (602)	.8067 (538)	.983 (646)
≥ 2005	GB	.8624 ($\pm$.0382)	.8626	.7857 (476)	.8025 (476)	.9801 (552)
≥ 2010	GB	.8588 ($\pm$.0483)	.857	.6835 (218)	.8782 (353)	.9841 (251)
≥ 2015	RF	.8525 ($\pm$.1814)	.8531	.8525 (61)	.84 (50)	.9091 (11)

Appendix 6: Evaluation of the dataset: abstract length bias

In the SUSdblp dataset, concatenated titles and abstracts of uninfluential papers tend to be shorter than those of seminal or survey publications. Possible classification accuracy bias caused by these different lengths is evaluated in the following experiment.

For each of the three groups, publications complete with their citation network were extracted which did not have a paper in the other two classes with the same abstract lengths. After this, 342 publications remained for each class. Doc2Vec document vector representations combined with cosine distance are utilised here to derive the 60 features on which classification was performed. The highest accuracy was reached by usage of random forests as classifier (Acc .8285 ($\pm$.0751), F1 .828, Acc_{c_0} .7398, Acc_{c_1} .7602, Acc_{c_2} .9854). Not significant differences were found in comparison to utilisation of the full SUSdblp dataset when using Welch-ANOVA. Bias due to different abstract lengths for the three classes could thus be suspended in terms of overall classification accuracy.

Appendix 7: Other approaches: all dimensions of document vector representations of publications

Single dimensions of document vector representations of publications

For tf-idf vectors, LDA document representations as well as years, single dimensions are quite understandable. The meaning of Doc2Vec and BERT dimensions cannot be explained as easily. So here, we restrict the single dimension classifications on tf-idf vectors, LDA document vector representations and publication years.

The following classifications are all performed by using one dimension of the numerous dimensions of document vector representation of publications P .

When classifying on one of the dimensions of unstemmed tf-idf document vector representations, the highest accuracy of .5359 was achieved for the dimension representing the word *survey*. Unfortunately, this classifier is completely unable to identify publications of type uninfluential. For single dimensions from stemmed tf-idf document vector representations, an accuracy of .5616 can be achieved for dimension 380 which refers to the word *survei* (stemmed version of survey) as the most descriptive one. Again, the classifier is not able to identify uninfluential papers. Classification on single dimensions from the LDA embedding of unstemmed publications lead to an accuracy of up to .5409 for dimension ut41. This dimension seems to represent a background topic AlSumait et al. (2009) which usually is contained in all documents. In vast parts of the corpus, topic 41 can be observed as displayed in Fig. 5. The top words of this topic can be seen in Table 13. When classifying on dimensions of stemmed LDA document vector representations, an accuracy of .5101 is achieved for dimension

Table 13 Ten most probable words per topic for best performing topics in the single feature classification based on features of the publication alone, in decreasing probability

Topic	Ten most probable words
ut41	the, of, and, to, in, a, is, this, as, are
st87	the, and, of, in, to, research, thi, on, their, null

Table 14 Classification accuracy and F1 scores for single dimensions DIM from different vector representations of publications *P* from the SUSdblp dataset for the best performing classification algorithm

V	Cla	DIM	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	KNN	u460	.5359 ($\pm$.0273)	.4473	1	.4394	0
tf-idf S	KNN	s380	.5616 ($\pm$.0316)	.4707	1	.4879	0
LDA U	GB	ut41	.5409 ($\pm$.0930)	.5196	.6576	.8182	.1106
LDA S	GB	st87	.5101 ($\pm$.0544)	.4991	.75	.676	.0424
Years	GB	0	.4879 ($\pm$.0452)	.4837	.3773	.4591	.6273

st87. From its most probable words, which are displayed in Table 13, this dimension again seems to describe a background topic. An accuracy of .4879 is achieved when classifying on publication years of papers *P*.

Table 14 provides detailed results on accuracies per class and F1 scores for classification based on single dimensions from the document vector representations as well as the best performing classifiers. In general, the highest accuracy decreased by .1323 when compared to the accuracy achieved with usage of the best single feature derived from citation networks of publications. The five models are significantly different from each other when looked at with Kruskal–Wallis *H* test. Utilisation of years is significantly different from the two tf-idf variants. Additionally, usage of stemmed tf-idf is significantly different from utilisation of stemmed LDA.

All dimensions of document vector representations of publications

The next experiments observe classification on all dimensions of document vector representations of the publications *P*. As the dimension representing the words survey and review naturally tend to be highly descriptive in our task, we also performed classifications on stemmed and unstemmed tf-idf vectors, where we omitted these two dimensions. Table 15 shows detailed results on all document vector representations with their accuracies, F1 scores and accuracies for the three classes. The highest accuracy from the multi feature classification of semantometrics surpassed the best result from this experiment by .0899.

Significant differences in the eight models were found when looked at with Kuskal–Wallis *H* test. The two LDA document vector representations significantly differ from all other document vector representations.

Table 15 Classification accuracy and F1 scores for all dimensions from the different vector representations of publications *P* from the SUSdblp dataset for the best performing classification algorithm

V	Cla	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	GB	.7848 ($\pm$.0598)	.7873	.6894	.8576	.8076
tf-idf S	GB	.7707 ($\pm$.0373)	.7734	.6894	.8606	.7621
tf-idf wo SR U	GB	.7722 ($\pm$.0308)	.7729	.6879	.8576	.7712
tf-idf wo SR S	GB	.7601 ($\pm$.0443)	.7605	.6879	.8591	.7333
D2V U	SVM	.7763 ($\pm$.0642)	.7766	.7152	.8636	.75
BERT U	LR	.7828 ($\pm$.0544)	.7819	.7273	.903	.7182
LDA U	GB	.704 ($\pm$.0756)	.7014	.6848	.8318	.5955
LDA S	GB	.7071 ($\pm$.0472)	.7059	.6697	.8076	.6439

Table 16 Classification accuracy and F1 scores for all dimensions from the different vector representations of referenced papers *X* from the SUSdblp dataset for the best performing classification algorithm

V	Cla	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	GB	.8778 ($\pm .0349$)	.8768	.8485	.8212	.9636
tf-idf S	GB	.8712 ($\pm .0483$)	.8699	.8318	.8121	.9697
D2V U	RF	.8152 ($\pm .0368$)	.8151	.8242	.7848	.8364
BERT U	GB	.8212 ($\pm .0373$)	.8214	.8364	.8076	.8197
LDA U	GB	.7985 ($\pm .0444$)	.798	.7273	.847	.8212
LDA S	RF	.797 ($\pm .0501$)	.7948	.6818	.8439	.8652
Years	RF	.4343 ($\pm .0871$)	.4334	.3636	.4394	.5

All dimensions of document vector representations of referenced papers

The following experiments required for vector representations of referenced papers of publications to be of equal length. For this, we averaged the values of all dimensions of the document vector representations for all referenced papers to obtain vectors of a length which equals the number of dimensions of a certain document representation. Using only referenced publications in the classification would equal using only features derived from group *D* for semantometrics. Table 16 shows detailed results on accuracies, F1 scores and accuracies for the three classes for classification based on one-vector representations of referenced papers *X* of *P*.

Usage of ANOVA showed significant differences between the seven observed groups. With Bonferroni correction and Scheffé's method significant differences between classification based on the dimensions of the tf-idf methods and all other document vector representations were found. Usage of years also results in significant differences from utilisation of all other document vector representations in the classification task.

All dimensions of document vector representations of citing publications

These experiments again required for vector representations of citing papers of publications to be of equal length. For this, we averaged the values of all dimensions of the document vector representations for all citing papers to obtain vectors of a length which equals the number of dimensions of a certain document representation. Using only citing papers in the classification would equal using only features derived from group *E* for semantometrics. Table 17 shows detailed results on accuracies, F1 scores and accuracies for the three classes for classification based on one-vector representations of citing publications *Y* of *P*.

Table 17 Classification accuracy and F1 scores for all dimensions from the different vector representations of citing papers *Y* from the SUSdblp dataset for the best performing classification algorithm

V	Cla	Acc	F1	Acc _{c₀}	Acc _{c₁}	Acc _{c₂}
tf-idf U	GB	.8121 ($\pm .0587$)	.8109	.7242	.8439	.8682
tf-idf S	GB	.8035 ($\pm .0410$)	.8022	.7152	.8288	.8667
D2V U	SVM	.7742 ($\pm .0476$)	.7742	.8136	.7833	.7258
BERT U	GB	.7515 ($\pm .0482$)	.7513	.7803	.7621	.7121
LDA U	GB	.8707 ($\pm .0429$)	.8701	.8333	.8258	.9530
LDA S	RF	.8747 ($\pm .0234$)	.8732	.8152	.8273	.9818
Years	RF	.5556 ($\pm .0692$)	.5476	.4288	.4788	.7591

Application of Kruskal–Wallis H test on the seven models showed significant differences. The two models representing words and semantics of publications were significantly different from those utilising topics. Classification on BERT embeddings was also significantly different from classification on tf-idf vectors. Utilisation of years produced significantly different accuracies than all other embeddings except Doc2Vec.

References

- Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, 12(3), 159–170.
- AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of LDA generative models. *ECML/PKDD*, 2009(1), 67–82.
- Avramescu, A. (1979). Actuality and obsolescence of scientific literature. *JASIS*, 30(5), 296–303.
- Baltes, S., & Diehl, S. (2016). Worse than spam: issues in sampling software developers. In *ESEM 2016* (pp. 52:1–52:6).
- Banks, M. G. (2006). An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics*, 69(1), 161–168.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *ICML 2006* (pp. 113–120).
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *TACL*, 5, 135–146.
- Börner, K., Penumarthy, S., Meiss, M. R., & Ke, W. (2006). Mapping the diffusion of scholarly knowledge among major U.S. research institutions. *Scientometrics*, 68(3), 415–426.
- Bornmann, L., & Daniel, H.-D. (2010). The citation speed index: A useful bibliometric indicator to add to the h index. *Journal of Informetrics*, 4(3), 444–446.
- Burton, R. E., & Kebler, R. W. (1960). The “half-life” of some scientific and technical literatures. *American Documentation*, 11(1), 18–22.
- Cano, V., & Lind, N. C. (1991). Citation life cycles of ten citation classics. *Scientometrics*, 22(2), 297–312.
- CORE Computing Research and Education. <http://www.core.edu.au/>. Accessed 18 Dec 2019.
- Cronin, B., & Meho, L. I. (2006). Using the h-index to rank influential information scientists. *JASIST*, 57(9), 1275–1278.
- Davis, P. M., & Cochran, A. (2015). Cited half-life of the journal literature. *CoRR*. [arXiv:1504.07479](https://arxiv.org/abs/1504.07479).
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT 2019* (1), 4171–4186.
- Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. *ICML 2007* (pp. 233–240).
- Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1), 131–152.
- Fuhr, N. (2017). Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3), 32–41.
- Garfield, E. (1964). Can citation indexing be automated? *Essays of an Information Scientist*, 1, 84–90.
- Gerrish, S., & Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *ICML 2010* (pp. 375–382).
- Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2), 563–577.
- Herrmannova, D., & Knoth, P. (2015). Semantometrics in coauthorship networks: Fulltext-based approach for analysing patterns of research collaboration. *D-Lib Mag.*, 21(11/12).
- Herrmannova, D., Knoth, P., & Patton, R. M. (2018). Analyzing citation-distance networks for evaluating publication impact. In *LREC 2018*.
- Herrmannova, D., Patton, R. M., Knoth, P., & Stahl, C. G. (2017). Citations and readership are poor indicators of research excellence: Introducing TrueImpactDataset, a new dataset for validating research evaluation metrics. In *Proceedings of the 1st Workshop on Scholarly Web Mining*, 2017.
- Hirsch, J. E. (2005). An index to quantify an individual’s scientific research output. *PNAS*, 102(46), 16569–16572.
- Hirsch, J. E. (2019). h_a : An index to quantify an individual’s scientific leadership. *Scientometrics*, 118(2), 673–686.

- Hou, W.-R., Li, M., & Niu, D.-K. (2011). Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. *Bioessays*, 33(10), 724–727.
- Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, 54(6), 855–863.
- Jones, E., Oliphant, E., Peterson, P., et al. (2001). SciPy: Open source scientific tools for python. Retrieved September 18, 2019, from <http://www.scipy.org>.
- Kanter, J. M., & Veeramachaneni, K. (2015). Deep feature synthesis: Towards automating data science endeavors. In *DSAA 2015* (pp. 1–10).
- Knoth, P. & Hermannova, D. (2014). Towards semantometrics: A new semantic similarity based measure for assessing a research publication's contribution. *D-Lib Magazine*, 20(11/12).
- Kreutz, C. K., Sahitaj, P., & Schenkel, R. (2019). Reevaluating semantometrics from computer science publications. In *BIRNDL@SIGIR 2019* (pp. 42–55).
- Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., et al. (2012). Radiomics: The process and the challenges. *Magnetic Resonance Imaging*, 30(9), 1234–1248.
- Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Rep4NLP@ACL 2016* (pp. 78–86).
- Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *ICML 2014* (pp. 1188–1196).
- Levene, H. (1960). Robust tests for equality of variances. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, 278–292.
- Ley, M. (2009). DBLP—some lessons learned. *PVLDB*, 2(2), 1493–1500.
- MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. *JASIST*, 61(1), 1–12.
- Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & López-Cózar, E. D. (2016). Back to the past: On the shoulders of an academic search engine giant. *Scientometrics*, 107(3), 1477–1487.
- Massey, F. J, Jr. (1951). The Kolmogorov–Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46, 68–78.
- Moed, H. F. (2002). The impact-factors debate: The ISI's uses and limits. *Nature*, 415, 731–732.
- Montolio, S. L., Dominguez-Sal, D., & Larriba-Pey, J.-L. (2013). Research endogamy as an indicator of conference quality. *SIGMOD Record*, 42(2), 11–16.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in PyTorch. In *Autodiff@NIPS 2017*.
- Patton, R. M., Hermannova, D., Stahl, C. G., Wells, J. C., & Potok, T. E. (2017). Audience based view of publication impact. In *WOSP@JCDL 2017* (pp. 64–68).
- Patton, R. M., Stahl, C. G., & Wells, J. C. (2016). Measuring scientific impact beyond citation counts. *D-Lib Magazine*, 22(9/10).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *JMLR*, 12, 2825–2830.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP, 2014*, 1532–1543.
- Pietraszek, T. (2005). Optimizing abstaining classifiers using ROC analysis. In *ICML 2005* (pp. 665–672).
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.
- Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, 149, 510–515.
- Pride, D., & Knoth, P. (2017). Incidental or Influential?—Challenges in automatically detecting citation importance using publication full texts. In *TPDL 2017* (pp. 572–578).
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.
- Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large Corpora. In *New Challenges for NLP Frameworks@LREC 2010* (pp. 45–50).
- Rocha, L. M. A., & Moro, M. M. (2016). Research contribution as a measure of influence. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 2259–2260).
- Rosen-Zvi, M., Griffiths, T. L., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *UAI 2004* (pp. 487–494).
- Rousseau, R., & Ye, F. Y. (2008). A proposal for a dynamic h-type index. *JASIST*, 59(11), 1853–1855.
- Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620.
- Schofield, A., Magnusson, M., & Mimmo, D. M. (2017). Pulling out the stops: Rethinking stopword removal for topic models. In *EACL* (vol. 2, pp. 432–436).
- Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *ELA*, 78(3), 30002.

- Schreiber, M. (2008). The influence of self-citation corrections on Egghe's g index. *Scientometrics*, 76(1), 187–200.
- Seglen, P. O. (1992). The skewness of science. *JASIS*, 43(9), 628–638.
- Seglen, P. O. (1994). Casual relationship between article citedness and journal impact. *JASIS*, 45(1), 1–11.
- Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314(7079), 498–502.
- <https://doi.org/10.5281/zenodo.3258164>
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4), 591–611.
- Shi, X., Leskovec, J., & McFarland, D. A. (2010). Citing for high impact. In *JCDL 2010* (pp. 49–58).
- Silva, T. H. P., Moro, M. M., Silva, A. P. C. d., Meira Jr., W., & Laender, A. H. F. (2014). Community-based endogamy as an influence indicator. In *JCDL 2014* (pp. 67–76).
- Simkin, M. V., & Roychowdhury, V. P. (2005). Copied citations create renowned papers? *Annals of Improbable Research*, 11(1), 24–27.
- Smith Aversa, E. (1985). Citation patterns of highly cited papers and their relationship to literature aging: A study of the working literature. *Scientometrics*, 7(3–6), 383–389.
- Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In *KDD 2008* (pp. 990–998).
- Tortorella, F. (2000). An optimal reject rule for binary classifiers. In *Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition* (pp. 611–620).
- Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Wade, A. D., Wang, K., Sun, Y., & Gull, A. (2006). WSDM Cup 2016: Entity ranking challenge. In *WSDM 2016* (pp. 593–594).
- Whalen, R., Huang, Y., Sawant, A., Uzzi, B., & Contractor, N. (2005). Natural language processing, article content & bibliometrics: Predicting high impact science. In *ASCW'15 Workshop at Web Science 2015* (pp. 6–8).
- Zhu, X., Turney, P. D., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *JASIST*, 66(2), 408–427.