Master's Thesis

Semantic approaches to citation recommendation

Tarek Saier

Examiners: Prof. Dr. Georg Lausen
Prof. Dr. Christian Schindelhauer

Albert-Ludwigs-University Freiburg

Faculty of Engineering

Department of Computer Science

Chair of Databases and Information Systems

April 30th, 2019

Writing Period

15. 10. 2018 – 30. 04. 2019

First Examiner

Prof. Dr. Georg Lausen

Second Examiner

Prof. Dr. Christian Schindelhauer

Supervisor

Dr. Michael Färber

Master-Thesis

Semantic approaches to citation recommendation

Tarek Saier

Gutachter: Prof. Dr. Georg Lausen
Prof. Dr. Christian Schindelhauer

Albert-Ludwigs-Universität Freiburg

Technische Fakultät

Institut für Informatik

Lehrstuhl für Datenbanken und Informationssysteme

30. April 2019

Bearbeitungszeit

15. 10. 2018 – 30. 04. 2019

Erstgutachter

Prof. Dr. Georg Lausen

Zweitgutachter

Prof. Dr. Christian Schindelhauer

Betreuer

Dr. Michael Färber

Declaration

I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work. I hereby also declare, that my Thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Place, Date

Signature

Abstract

New research is being published at a rate, at which it is infeasible for many scholars, to read and assess everything possibly relevant to their work. In pursuit of a remedy, efforts towards automated processing of publications, like semantic modelling of papers to facilitate their digital handling, and the development of information filtering systems, are an active area of research. In this thesis, we investigate the semantic modelling of citation contexts for the purpose of citation recommendation. For this, we generate a large data set with accurate citation information from publications' LaTeX sources on arXiv.org. Using this data set, we develop semantic recommendation models based on entities and claim structures. To assess the effectiveness and conceptual soundness of our models, we perform a large offline evaluation on our own as well as several established data sets and furthermore conduct a user study. Our findings show, that the models can outperform a non-semantic baseline model and do, indeed, capture the kind of information they're conceptualized for.

Zusammenfassung

Angesichts der Menge und Frequenz neuer wissenschaftlicher Publikationen, ist es für viele Forschende nicht praktikabel, jede Veröffentlichung zu lesen, die potentiell relevant für die eigene Arbeit ist. Dieser Form der Informationsüberflutung entgegengesetzt stehen Bestrebungen zur automatisierten Verarbeitung akademischer Arbeiten, wie die semantische Modellierung wissenschaftlicher Publikationen zur Erleichterung derer maschineller Handhabung, sowie die Entwicklung akademischer Empfehlungsdienste. In dieser Thesis untersuchen wir die semantische Modellierung von Zitierkontexten zum Zwecke der Zitierempfehlung. Hierfür generieren wir einen großen Datensatz mit akkurater bibliographischer Vernetzung aus LaTeX-Quelldaten von Publikationen auf arXiv.org. Mithilfe dieses Datensatzes entwickeln wir semantische Modelle von Zitierkontexten, basierend auf Entitäten und Behauptungsstrukturen. Zur Prüfung unserer Modelle auf Eignung für die Zitierempfehlung und konzeptionelle Richtigkeit, führen wir eine umfassende Offline-Evaluation anhand unseres eigenen und anderer Datensätze, sowie eine Nutzerstudie durch. Unsere Ergebnisse zeigen, dass die Modelle bessere Ergebnisse als eine nicht-semantische Referenzmethode erzielen können, und in der Tat die Informationen erfassen, für die sie konzipiert wurden.

Contents

1 Introduction

1.1 Motivation

Citations are a central building block of scholarly discourse. They are the means by which scholars relate their research to existing work—be it by backing up claims, criticising, naming examples or engaging in any other form. Citing in a meaningful way requires an author to be aware of publications relevant to their work. Here, the ever increasing rate of new research being published, poses a serious challenge. Even with academic search engines like Goolge Scholar and CiteSeer^{x} at our disposal, identifying publications that are worthwhile to examine and appropriate to reference remains a time consuming task.

It is therefore not surprising that methods to aid researchers in these tasks have been and are still being actively researched. While diverse in nature, the common core of these efforts is the goal to utilize the automated processing of publications. This can be achieved by either extracting information from publications as they are [1, 2], or by semantically modelling the publications or their contents in order to facilitate automated processing [3–6]. Once processed, a prevalent practice is to harvest human made citations, analyze them [7, 8] and use them for example to recommend papers [2] or aid in document exploration [9]. Although systems like this have existed for over 20 years [2, 10], there is not a lot of work looking into the use of semantic models of citations for the recommendation of papers. Therefore we will investigate this approach. More specifically, we will concentrate on the task of recommending papers for citation—as opposed to, for example, discovery. What this encompasses will be described in more detail in the following section.

1.2 Problem setting

In the broadest sense, recommending papers for citation means given an input text, suggest publications that can be referred to from within that text. In scale this can vary from specific recommendations for a section of a sentence (called *local* or *context-based*), to general recommendations for a whole input document (*global*). The task can also include deciding whether or not the input contains parts that would justify inserting a citation in the first place [11]. In this thesis, we will focus on local citation recommendation with the assumption that the input always allows for or requires a citation to be put in.

Another distinction to be made is between personalized and general citation recommendation. Some approaches make use of user specific information such as an author's prior citations. Collaborative filtering approaches by nature include a user model and therefore fall into this category. While personalization can improve recommendation, it limits the approach to users that are willing to share personal information. We therefore limit ourselves to purely content based filtering approaches.

A last clarification has to be made with regards to the semantic modelling of citations. Focussing on *local* citation recommendation, it is our goal to generate and utilize *explicit semantic representations* of the text passages in which authors reference existing work. "Explicit semantic" is to be understood as a differentiation from the mere use of unstructured text and implicit semantic information like the concepts identified by latent semantic indexing [12].

The problem setting can be summarized as follows. To investigate is, the applicability of and requirements for the use of explicit semantic representations for content based, local citation recommendation. The following section will outline how this investigation is performed.

1.3 Method

In order to assess if and how explicit semantic representations can benefit citation recommendation, we define several models that encode different semantic aspects of citations, and evaluate them against a non-semantic baseline. For the development and evaluation of these models we generate a data set with accurate citation information, that is also large enough for a realistic assessment. As far as possible, the models are furthermore evaluated on existing data sets.

Our models are based on named entities as well as claim structures contained in citation contexts, and are evaluated against a bag-of-words baseline. The assessment encompasses an offline evaluation on several data sets as well as a user study. Evaluating over several data sets ensures wide applicability and increases comparability with other approaches. The additional user study gives further insight into the data used, and provides an alternative way of judgement of performance, albeit on a smaller scale. Viewing our evaluation results through a selection of differently natured metrics, we can draw conclusions on the applicability of our models for various tasks.

1.4 Contributions

The contributions of this thesis are as follows

1. A large data set with accurate citation information, suited for the development and evaluation of approaches to a range of citation based tasks, like local and global citation recommendation and document discovery. Because the data set spans several disciplines it furthermore enables comparative analyses in e.g. citation behaviour. (See Chapter 4.)
2. Entity and claim based models for citation recommendation that can outperform non semantic models and mark a first step towards truly semantic citation recommendation. (See Chapter 5.)

3. Insights into open problems with building semantic models around citations, such as prevalent citation styles that hamper the grammatical parsing of sentences. (See Chapter 5.)

1.5 Document structure

The remainder of this document is structured as follows. Chapter 2 gives an overview of related work, thereby establishing a frame of reference for this thesis. This is followed in Chapter 3 by the definition of terms and explanation of concepts making up the theoretical background. Chapter 4 is dedicated to our new data set; discussing the reasoning for creating it as well as the generation process and result. The development and resulting definitions of our entity and claim based models are detailed in Chapter 5. In Chapter 6 we evaluate our models after a short but necessary discussion of special considerations when evaluating citation recommendation approaches. Final and overarching observations are offered in Chapter 7. Chapter 8 concludes the thesis with an outlook on future work.

2 Related Work

To the best of our knowledge there is, so far, almost no work investigating (1) the use of explicit semantic representations for (2) the task of local citation recommendation. We will therefore present related work in two areas. First, semantic approaches to *global* citation and paper recommendation. Second, *non-semantic* approaches to local citation recommendation.

Note that SemCir [13] (see below) is the only case we are aware of, that can be regarded as a semantic approach to local citation recommendation. The explicit semantic representations are, however, not generated from citation contexts (local), but from papers (global), that are textually (not necessarily semantically) similar to the citation contexts. To avoid complicating the document structure unnecessarily, we therefore list SemCir as a semantic but global approach.

2.1 Semantic approaches to citation recommendation

The key difference between the following approaches and our work, is that they recommend publications *globally* and not *locally.* This means the input for a recommendation is not a citation context but a whole paper, meaning semantic similarity between input and recommendation candidates can be assessed by paper specific aspects like authors, title, abstract, etc.

At a point in time, when publishing research papers online was an emerging trend, Middleton et al. [14] propose a system for paper recommendation making use of a topic ontology. Based on classifying papers into topics and recording, which papers a researcher would access on the web, they employ content-based filtering, collaborative filtering and a feedback mechanism to suggest papers from new topics to users. Comparing the topic ontology to a flat list of topics in two user studies, they report 7–15% more user satisfaction for the ontology case.

In a similar vein, Zhang et al. [15] propose a hybrid recommender system for papers based on semantic concept similarity. They derive concepts from CiteULike[1] tags and use these to measure the semantic similarity of papers and users' interest. In their evaluation they compare different settings of the approach but do not compare to other work or alternative techniques.

Jiang et al. [16] use CiteULike tags as academic concepts to build a topic model applied to paper abstracts. In a content-based recommendation setting they let volunteers judge the relevance of recommendations for a test set of 30 papers. The evaluation includes a TFIDF baseline, latent Dirichlet allocation (LDA) and an approach combining LDA with their concept model. The report MAP@5 and NDCG@5 values to be best for the LDA+concept method.

In [17] Zarrinkalam et al. enrich their metadata on research papers using multiple Linked Open Data (LOD) sources to drive a hybrid recommender system. They compare a purely content-based method using only text similarity with a second method additionally utilizing collaborative filtering and a third method furthermore using the LOD enriched data. They report recall, co-cited probability and NDCG values for various cut-off values for which the LOD enriched method consistently achieves the best performance.

With SemCiR [13] Zarrinkalam et al. introduce a content-based, global citation recommendation approach that utilizes a semantic distance measure between papers. They

[1] See `http://citeulike.org/`.

furthermore introduce a method for extending the measure to determine the semantic distance between an input text and a paper, which is achieved by representing the input by textually similar papers. The distance measure suggested builds on six different relational features including shared authors, venue, and overlapping in- and outgoing citations. The approach is evaluated on a 12,500 paper subset of CiteSeer^x [18] in a citation re-prediction setting, using as input a paper's title, abstract and contexts in other papers where it was referred to. An evaluation of different scenarios measuring recall, co-cited probability and NDCG leads the authors to conclude, that recommendation results can be improved by using their semantic distance measure and including citation contexts in the measurement of textual similarity.

2.2 Local citation recommendation

The following approaches recommend publications for a specific citation context (local). The key difference to our work is, that they solely rely on lexical and syntactic features (n-grams, part-of-speech tags, word embeddings etc.) and do not attempt to semantically model citation contexts.

Probably one of the first investigations into local citation recommendation is the work of He et al. [19]. They propose a two-step system that first identifies recommendation candidates and then re-ranks them by concept similarity. While also discussing global citation recommendation in detail, for the local case they compare recommending for a single context and recommending for all contexts within a document simultaneously. In an evaluation on the CiteSeer^x data set measured by recall, co-cited probability and NDCG they find that the single context task is harder, but also, that their approach to the all contexts task achieves results comparable to and even better than some global citation recommendation methods.

In a follow-up work Huang et al. [20] build upon above work by swapping out the computationally complex concept based re-ranking method with a translational model. In this model citation contexts are treated as the source language and cited papers as words in the target language. The resulting system, RefSeer, is evaluated on two smaller data sets (CiteULike and a CiteSeer subset) and one large one (all of CiteSeer). The authors report precision, recall, Bpref and MRR values for the two smaller data sets and conclude that their system can give correct recommendations in a realistic setting—such as when only the top 10 recommendations are shown.

Huang et al. improve RefSeer with a neural probabilistic model that learns distributed representations of words and documents in [21]. They evaluate their model for local citation recommendation on the whole of CiteSeer, splitting between train and test set at the year 2011 (9M contexts train, 1.5M contexts test). Measuring MAP, MRR and NDCG they show that their model outperforms 4 different state-of-the-art approaches. An analysis on the influence of papers' citation counts on recommendation performance shows that their approach especially exceeds other work in case of lesser cited papers (<100 citations).

In [22] Duma et al. test the effectiveness of using a variety of document internal and external text inputs with a TFIDF model. Their data set is built from the ACL Anthology and contains 5,446 citations. In a re-prediction setting the authors measure how reliably their models rank the correct paper at the top position. They conclude that a mixture of internal and external inputs outperforms either of the aforementioned used on their own.

The work presented in [23] by Duma et al. focusses on the rhetorical function of sentences. The authors classify sentences using the Core Scientific Concepts (CoreSC) scheme and investigate how their distinction can be used to improve recommendation. Evaluating on one million papers from the PubMed Central Open Access Subset they measure NDCG values and find that for several classes of input sentences significant gains in recommendation quality can be made by focussing on certain rhetoric passages of candidate documents when ranking text similarity.

The Neural Citation Network (NCN) proposed by Ebesu et al. in [24] is inspired by neural machine translation, learns citation context representations as well as author representations and includes an attention mechanism. For their evaluation, the authors use 4.5 million citation contexts from the RefSeer data set and report NDCG, MAP, MRR and recall values. They compare against a BM25 baseline, a citation translation model as well as two variations of their model that do not make use of author representations. While BM25 is outperformed by all of the other approaches to some degree, the NCN's results lead the evaluation by a distinct margin.

Kobayashi et al. [25] describe a variation of local citation recommendation they call *context-based co-citation recommendation*. The input here is a citation context *plus* one publication referred to in that contexts. The goal then is to recommend other publications that also can be used as citations in that contexts. By classifying text sections into the discourse facets "objective", "method", and "result", the authors are able to learn distributed vector representations per facet, which are then used for the recommendation. They evaluate their approach on contexts from the ACL Anthology containing "enumerated co-citations" (e.g. *[27,42]*) and report NDCG values at a cut-off of 100. In comparison with two baseline methods their discourse facets are shown to be effective.

In [26] Jeong et al. introduce an approach to local citation recommendation using Graph Convolutional Networks (GCN) and Bidirectional Encoder Representations from Transformers (BERT). The GCN is used to capture information from the citation relationships between papers, while the pre-trained BERT is applied on the citation contexts themselves. The authors evaluate their approach on a subset of 6,500 papers from the ACL Anthology and a self-created data set of close to 5,000 papers. They report MAP, MRR and recall values at different cut-offs, demonstrating that their BERT+CGN approach outperforms several reduced versions of the aforementioned as well as a baseline model.

3 Background

In this chapter we will define important terms and lay out the theoretical background for the remainder of the thesis.

3.1 Definition of terms

Citation The term *"citation"* can refer to both the act of citing as well as the occurrence of being cited. This can be illustrated with the phrase *"an author's citations"*. In [2] Beel et al. write *"McNee et al. assumed that an author's citations indicate a positive vote for a paper [93]."* (the author cites), while Myers [27] writes *"Thus, the number of an author's citations, in this study, means the number of articles in which one or more of his publications were cited."* (the author is being cited). Like the latter example we will make an effort to use the term unambiguously.

Citation marker A citation marker is a string of text that identifies an entry within a document's reference section. Examples are "[27]", "[Bol98]" and "(He, 2010)". These are placed within the document whenever the author refers to one of the documents in the reference section.

Table 1: Examples of citations and their categorization into integral/non-integral (values left of split) as well as syntactic/non-syntactic (values right of split).

Context excerpt (citation marker highlighted)	Swales [28]	Hyland [29]	Thompson [30]	Okamura [31]	Lamers et al. [32]	Whidby et al. [33]	Abujbara et al. [34]
"Swales (1990) has argued that ..."	i	i	i	i	i	n	?
"Swales (1990) has argued that ..."	i	i	i	i	n	s	s
"Swales [42] has argued that ..."	i	i	i	i	i	n	n
"Swales has argued that ... [42]"	i	i	i	i	i	n	n
"It has been argued (Swales, 1990) that ..."	n	n	n	n	n	n	n
"It has been argued [42] that ..."	n	n	n	n	n	n	n
"According to (Swales, 1990) it is ..."	?	?	?	?	n	s	s
"According to [42] it is ..."	n	n	n	n	n	s	s
"... has been shown (see (Swales, 1990))."	n	n	n	n	n	s	n

Citation context The context of a citation is the text surrounding its citation marker. Typical sizes are 1–3 sentences. The sentence containing the marker is also sometimes referred to as "citing sentence".

Integral and syntactic citations There are two somewhat similar, and at first glance easily confused notions considering a citation's role within its context. They are referred to as "integral"—in the adjectival sense close in meaning to "essential" or "inherent", not what we denote in calculus with $\int$—and "syntactic". Integral citations were first defined by Swales [28] in 1990 and are a frequently used [29–32] measure in discourse analysis. An integral citation is, in Swales' own words, *"one in which the name of the researcher occurs in the actual citing sentence as some sentence-element"*. Thompson [30] rephrases the definition as *"citations that [...] play an explicit grammatical role within a sentence"*. While what Thompson refers to as "citations" might be confused with the notion of citation markers, the examples given in [30] clearly indicate that a "citation" is to mean an author's name in their definition. The second notion, "syntactic" (as used in [33] and [34]), is concerned with whether or not a *citation marker* has a grammatical role within its context. In other words, if removing the citation marker would make the citing sentence ungrammatical, then it is syntactic. Table 1 gives an overview of examples for both concepts. Note that Lamers et al. [32] provide a classification algorithm for integral and non-integral citations that slightly differs from Swales' original definition depending on the interpretation of a citation marker's scope, but also gives a clear classification in an edge case where Swales's definition is unclear. Furthermore note that the two ways for distinguishing syntactic and non-syntactic citations found in literature are not identical. This is in part because the method given in [34] is kept rather simple. For the intents and purposes of this work we can follow the definitions of Lamers et al. and Whidby et al. for "integral" and "syntactic" respectively.

Reference References are the entries in a document's reference section. Each reference should unambiguously identify another document that it points to. In the context of parsing reference sections we will, at times, refer to references as "reference strings". An example of a reference is *"W. Huang, P. Mitra, and C. L. Giles, 'RefSeer: A citation recommendation system,' in IEEE/ACM Joint Conference on Digital Libraries, pp. 371–374, Sep. 2014."*.

Reference resolution The act of parsing a reference string and matching it to a document identifier is what we refer to as "reference resolution". Examples can be found in Section 4.2.4.

Named Entity In natural language processing the term "Named Entity" (NE) describes unambiguous abstract or physical entities which have a proper name. Examples are people (Tim Berners-Lee), places (the city Taipei), organizations (the Free Software Foundation) and concepts (Okapi BM25). For a more detailed discussion of the term see [35].

Noun phrase A noun phrase is a group of words (or a single word) functioning grammatically as one unit, that has a noun at its head. Examples are "*example*", "noun *phrase*" and "context-based co-citation *recommendation*" (head of the phrase *highlighted*).

Claim For the scope of this thesis we define a "claim" as an ideally verifiable and at least debatable statement in writing. The motivation for this definition is to capture those parts of publications, that are backed by citations. Examples are "The boolean satisfiability problem (SAT) is NP-complete.", "CiteSeer was introduced in 1998." and "Parsing LaTeX sources is a non trivial task."—the latter statement showcasing a not necessarily verifiable but at least debatable statement.

Citation recommendation We define "citation recommendation" as the task of recommending publications for purpose of referencing them. The input for a recommendation is at least a citation context (optionally more, like information on the author) and the output a list of publications that can or should be referred to from within the context. To give an example, a good recommendation for the input "CiteSeer was introduced in 1998." would be "K. D. Bollacker, S. Lawrence, and C. L. Giles, *'CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications'*, AGENTS '98".

3.2 Theoretical background

3.2.1 Recommending scientific publications

Recommending scientific publications can be approached as a variety of tasks, such as link prediction, machine translation (see Section 2.2) and, of course, recommendation. In the following, we will take the latter perspective. Generally speaking, recommender systems are tools of discovery. In situations where a user can not reasonably be assumed to consider all available options, a recommender system's task is to identify those items that maximize a user's satisfaction, service provider's revenue or similar goal. This is typically done by calculating a score for each available item, ranking them by score and finally presenting the top ones to the user. Common domains of recommendation are shopping products and media (movies, news articles, etc.). A coarse distinction of techniques for recommendation—ignoring hybrid and less conventional cases—can be made between *content-based* and *collaborative filtering*. The latter exploits the similarity of user profiles for recommendation, while the former operates on the similarity of the items. [36]

Scientific publications as a recommender domain has some atypical characteristics. One of these is the ratio of users to items. Comparing, for example, to movie recommendation, the average number of new movies listed on the Internet Movie Database (IMDb) for the last 10 years (2009–2018) is about the same as the average number of new publications in the field of high energy physics on arXiv.org for the same time span—about 9,500 new items per year in both cases. Needless to say, there are more people watching movies than doing research in high energy physics. This can mean fewer users to compare to, when using collaborative filtering methods to recommend scientific publications. Another peculiarity in this domain is the interconnectedness of items. While movies may have aspects like actors, directors, locations, etc. in common, scholarly writings not only share authors, venues, journals, etc. but are furthermore interconnected through citations. A consequence of this is the feasibility to utilize citation networks and citation contexts for recommendation.

Recommendation of scientific publications can be done for several purposes, like introducing researchers to papers they might want to read, identifying works that are relevant to a draft of a new paper or suggesting publications that can be cited in a given context. While in all of these cases the nature of the items remains the same, input and constraints for the recommendation process differ. For the discovery of papers a researcher might find worthwhile reading, the focus lies on the researcher's interest and there are no externally imposed hard constraints. In contrast, when recommending publications for the use of citation, the context of the citation plays a considerably larger role than an author's interest. There might be some leeway for personal citation style, but by and large the adequacy

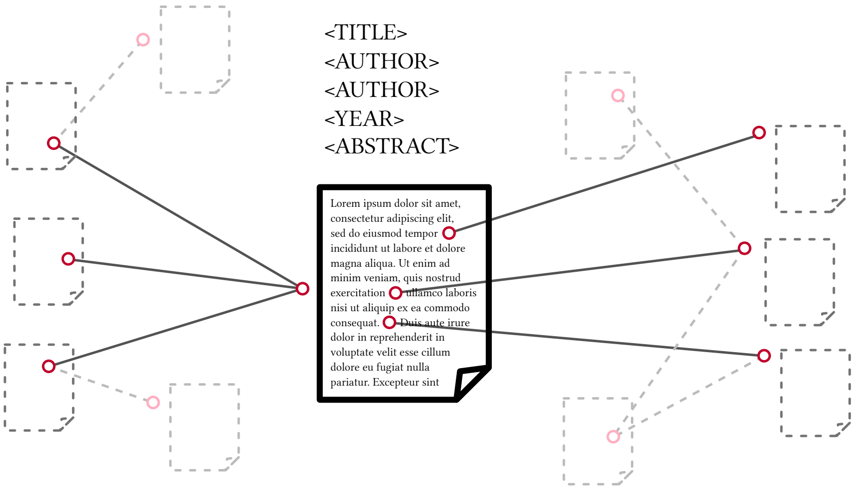


Figure 1: Schematic view on a document, its metadata and embedding in the citation graph.

of a citation is dictated by its target context and the contents of candidate items. In this way, the task of citation recommendation can also be seen as a more general information retrieval problem where the citation context is a query for which adequate citations have to be identified.

Having established the relative position of the field of citation recommendation with respect to related fields, we will lay out the core concepts relevant to our approach, *content based local citation recommendation by use of citation contexts*, below.

As mentioned above, the focus in the case of citation recommendation lies on the candidate items—which are scientific publications. Figure 1 shows the types of information that are available to describe them. First and foremost each publication has, of course, its textual contents. Alongside this it is also part of a network of citations, meaning is has documents it is being cited by as well as documents it is citing. Lastly, there may also be metadata describing the document, such as the title, authors, year of publication etc. In order to study the viability of using explicit semantic representations of citation contexts for citation recommendation, without introducing any confounding variables, we will describe our documents *only* by citation contexts. As citation contexts have been shown to contain information similar to and sometimes more extensive and focussed than abstracts [37], we

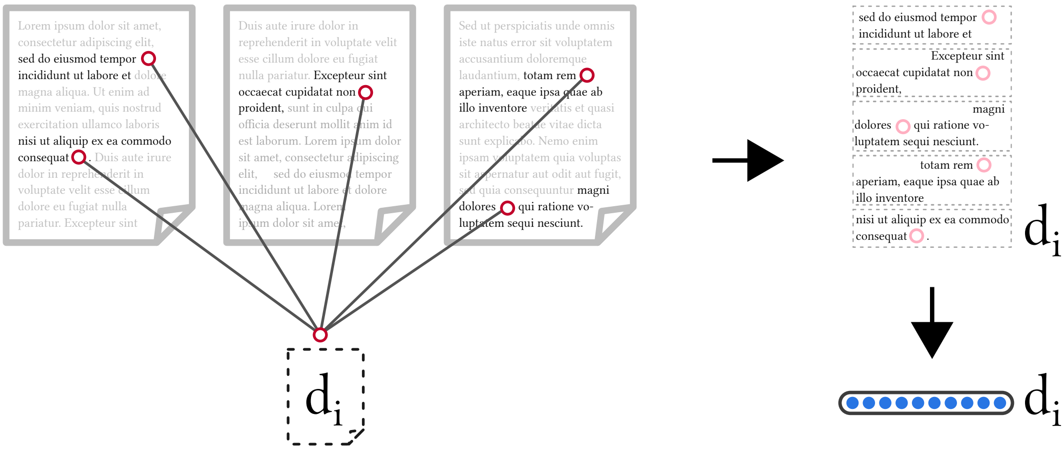


Figure 2: Schematic visualization of a document's representation using citation contexts.

can expect decent recommendation results even though we're not taking any metadata associated with or textual content contained in the candidate documents into account. Figure 2 shows how a document representation can be generated from citation contexts. Document d_i is cited by several other documents. By aggregating the contexts of these citations and deriving an abstract representation, d_i can be represented by how it is being referenced in existing literature.

Recommendation itself is then done by taking a citation contexts c_i as input, deriving an abstract representation in the same manner as was done for the aggregated contexts

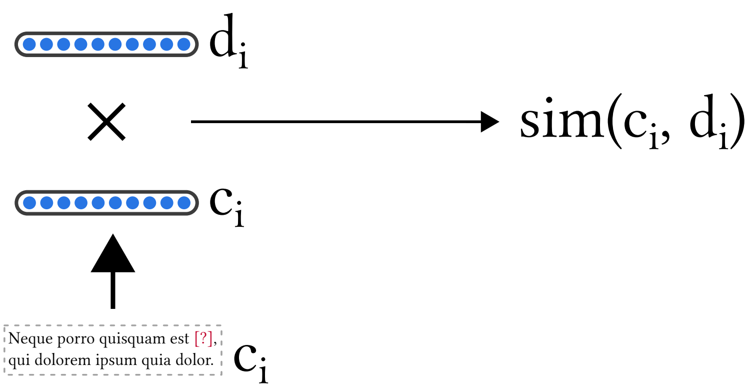


Figure 3: Schematic visualization of determining the similarity between an input citation contexts and a document representation.

referencing a document, and then comparing the input's representation against those of all the documents. One such comparison is schematically depicted in Figure 3. When every document—i.e., recommendation candidate—is associated with a similarity score, the system can output the most similar candidates as recommendations. With this, the overall picture looks as follows. The recommender system is trained on citation contexts, from which representations of the cited documents are learned. These documents are the publications the system is able to recommend. For a new context (without attached citation) given as input, the system identifies the documents most similar in the abstract representation space to the input and outputs these as recommendations.

3.2.2 Evaluating citation recommendations

Evaluation of recommender systems can be divided into three types, online evaluation, user study and offline evaluation. Online evaluation refers to monitoring naturally occurring user interaction with a system deployed online and drawing conclusions from metrics like click-through rates. User studies are conducted in a more controlled manner where users are aware of the evaluation and may be asked for explicit feedback on parts of the system. Offline evaluation differs from the other two types in that it does not directly involve human judgement. Rather, part of the data the system would, for online deployment, be trained on, is withheld and used as testing data. In doing this, past user judgement (like rating a movie or placing a citation) is used as a ground truth to test the system's output against. Because the human judgement is implicitly given within the data and can be processed automatically, offline evaluations can easily be performed on a large scale. [38]

When recommending citations, abovementioned three types of evaluation can be realized as follows. Online evaluation can be performed by measuring which, if any, of the suggested publications a user actually inserts into their text. In a user study, participants can be asked to judge the relevance of recommended items while the system's input could be chosen by the user or given by the creators of the study. Lastly, offline evaluation as outlined above results in a re-prediction setting. That is, existing citation contexts are stripped of their citations, input into the system, and the resulting output compared to the original.

To aggregate relevance judgements of a large number of recommendations into a single number in order to compare different approaches, metrics are necessary. We use Recall, MRR, MAP and NDCG measured at different cut-off values. A cut-off, often denoted as *Metric@k*, means only taking the top k recommendations into consideration. This is done because, realistically, a user will not go through a list of hundreds of suggested items, but only look at the top three, five or maybe ten. Recall, MRR, MAP and NDCG with a cut-off k can be defined as follows (see [38] for a more detailed, general discussion of these metrics):

Let $\mathcal{R} = r_1, r_2, ..., r_n$ be the ordered list of recommended items (highest to lowest) and $rel(r_i) \in [0, 1]$ denote the relevance of item r_i at rank i. Furthermore let $\mathcal{R}_{rel} = \{r_i | rel(r_i) = 1\}$ be the set of relevant items in an evaluation where $rel(r_i) \in \{0, 1\}$—that is, an item is either relevant or not, with no partial relevance in between. Note that in a re-prediction setting as described above, $|\mathcal{R}_{rel}| = 1$.

Recall measures the fraction of the number of relevant items that was retrieved. If a cut-off k is applied, the fraction's denominator is the minimum of the number of relevant items and the cut-off value.

$$Recall@k = \frac{\sum_{i=1}^{k} rel(r_i)}{min(|\mathcal{R}_{rel}|, k)}$$

For an evaluation with multiple test inputs, the mean over all Recall@k values is taken.

MRR (mean reciprocal rank) is a metric often used when there is just one relevant item (e.g. re-prediction of a citation) or (when there are multiple relevant items) defined to just consider the first one. It takes into account the position of a recommended item, by dividing its relevance by its rank.

$$MRR@k = \frac{\sum_{i=1}^{k} \frac{rel(r_i)}{i}}{min(|\mathcal{R}_{rel}|, k)}$$

Again, when evaluating over multiple test inputs, the mean over all MRR@k values is calculated.

MAP (mean average precision) is the mean of average precision (**AP**) values of multiple queries. AP measures the average of all precision values at the ranks where relevant items are found. To define this, let $\rho(k)$ denote the number of relevant items within the first k ranks. AP is then defined as

$$AP@k = \frac{\sum_{i=1}^{k} \frac{\rho(i)}{i} \times rel(r_i)}{max(\rho(k), 1)}$$

where $rel(r_i) \in \{0, 1\}$. The value of MAP@k is the mean AP@k over all evaluated queries.

NDCG (normalised discounted cumulative gain) is a metric compatible with non-binary relevance values that, similar to MAP, places higher weight on top ranks. It is calculated as the DCG (discounted cumulative gain) over the IDCG (ideal discounted cumulative gain)

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

where DCG is a sum of discounted relevance values

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel(r_i)} - 1}{log_2(i + 1)}$$

and IDCG is a normalization factor to get values between zero and one, derived from an ideal recommendation $\mathcal{R}_{ideal}$ in which all items are ordered from the most relevant to the least

$$IDCG@k = \sum_{r \in \mathcal{R}_{ideal}} \frac{2^{rel(r)} - 1}{log_2(i+1)}$$

Likewise to the other metrics, the mean is calculated when evaluating over a set of test inputs.

4 Data Set

Recommender systems rely on data for their development, training and evaluation. It is therefore important to properly assess potential data sets in terms of their strengths and shortcomings—especially with regards to the task at hand. In citation recommendation, the goal is to identify papers relevant to a user input. Because of the large amount of available research, this means a recommender has to be able to find relevant publications in a large set of possible candidates in order to be considered fit for the task. As a consequence, evaluation results are likely to be more meaningful when a large data set is used. Apart from the size, the quality of data is also crucial. For local citation recommendation, this means that a clean citation context, precise position of citation markers and valid citation information are desirable. With these criteria in mind we assessed existing data sets and come to the conclusion that, for the relatively new task of local citation recommendation, the creation of a dedicated data set will bring considerable benefits.

The following sections describe the details of our assessment as well as the creation process and evaluation of our new data set.

4.1 Existing data sets

Table 2 gives an overview of relevant existing data sets. While various recommendation domains have established quasi standard data sets, this is not yet the case in citation recommendation. CiteSeer^{x} is currently the most used in the field [2]. It is comparatively

Table 2: Overview of existing data sets.
Listed are the number of papers, nature of citation contexts, covered disciplines of citing papers and the type of global reference identifiers.
(*extractable** is to indicate that extraction might be error-prone due to papers only being available in PDF format.)

Data set	#Papers	Cit. context	Disciplines	Ref. IDs
CiteSeer^x [18] / RefSeer [20]	5M	400 chars	(unrestricted)	internal
PubMed Central OAS[1]	2.3M	extractable	Biomed./Life Sci.	mixed
Public Library of Science[2]	200k	extractable	mostly Biol./Med.	-
arXiv CS [40]	90K	1 sentence	CS	DBLP
Scholarly [41]	100k	extractable*	CS	-
ACL-AAN [42]	18K	extractable*	CS/Comp. Ling.	-
ACL-ARC [43]	11K	extractable*	CS/Comp. Ling.	-

large, but many approaches only use subsets and generate them with varying filtering criteria. It includes pre-extracted citation contexts of 400 characters in length, whereby references are resolved to an internal set of identifiers. Unfortunately there are several quality issues with the data set. The main ones being inaccurate citation information, noisy citation contexts and cut off words at the borders of citation contexts [39].

The PubMed Central Open Access Subset (PMC-OAS) is another large data set that has been used for citation based tasks [23, 44–46]. Contained publications are already processed and available in the JATS [5] XML format. While the data set overall is comparatively clean, heterogeneous annotation of citations within the text and mixed usage of global reference identifiers (PubMed, MEDLINE, DOI, ...) make it difficult to retrieve high quality citation interlinkings of documents from the data set[3] [44].

Similar to the PMC-OAS, albeit smaller, the Public Library of Science (PLOS) offers publications in the same XML format and has seen application for citation based tasks [47, 48]. In

[1] See `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`.
[2] See `https://www.plos.org/text-and-data-mining`.
[3] To be more precise, the heterogeneity makes the usage of the data set *as is* problematic. Resolving references retrospectively would be an option but comparatively challenging in the case of PubMed because of the frequent usage of special notation in publication titles; see also: `http://www.sciplore.org/files/citrec/CITREC_Parser_Documentation.pdf`.

contrast to the PMC-OAS, references are in general not resolved to any global identifier.

Consistent global reference identifiers are given in the arXiv CS data set in the form of DBLP IDs. Linking to an existing repository of publication (meta) data has the advantage that information about cited papers in readily available. The choice of DBLP restricts resolved references to the field of computer science though. Citations to, for example, publications in maths or statistics can not be resolved to a DBLP ID. A strength of the data set is that it was generated from LaTeX source files, which makes it possible to get very clean data.

For the remaining the data sets—Scholarly, ACL-AAN and ACL-ARC—citing papers are only available in PDF format and references are not resolved. The two ACL sets have the additional drawback of being very small.

Above observations lead us to the conclusion that it would be worthwhile to tackle the creation of a data set that is large (in the order of CiteSeer^{x}/RefSeer/PMC-OAS), clean (like the PMC-OAS and arXiv CS) and also offers consistent global reference IDs that don't restrict the data set to citations within the same discipline. The creation and evaluation of this data set is described in the following sections.

4.2 Data set creation

Scientific publications are usually distributed in formats targeted at human consumption (e.g. PDF) or, in cases like arXiv, also as source files *for* the aforementioned (e.g. LaTeX sources for generating PDFs). Citation-based tasks, such as local citation recommendation, in contrast, require automated processing of the publications' textual contents as well as the documents' interlinking through citations. The creation of a data set for such tasks therefore encompasses two main steps: extraction of plain text and resolution of references. In the following we will describe how we approached these two steps using arXiv publication sources and the Microsoft Academic Graph (MAG) [49].

4.2.1 Used data sets

The following two resources are the basis of the data set creation process.

arXiv.org hosts over 1.4 million submissions from August 1991 onward [50]. They are available not only as PDF, but (in most cases) also as LaTeX source files. The discipline most prominently represented is physics, followed by mathematics, with computer science seeing a continued increase in percentage of submissions ranking third (see Figure 7). The availability of LaTeX sources makes arXiv submissions particularly well suited for extracting high quality plain text and accurate citation information. So much so, that it has been used to generate ground truths for the evaluation of PDF to text conversion tools [51]. Approaches to automatically extract citation interlinks from arXiv sources by parsing LaTeX files have existed for over 20 years [52]. Nevertheless we are not aware of any existing data sets for citation based tasks generated from the whole of arXiv.

Microsoft Academic Graph (MAG) is a very large[4], automatically generated data set on publications, related entities (authors, venues, etc.) and their interconnections through citation. While citation contexts of varying length are available to some degree, full text documents are not. The size of the MAG makes it a good target for matching reference strings against it, especially given that arXiv spans several fields of study.

4.2.2 Pipeline overview

As depicted in Figure 4, we start out with arXiv sources to create the data set. From these we generate, per publication, a plain text file with the document's textual contents and a set of database entries reflecting the document's reference section. Association between references and citations in the text are preserved by placing citation markers in the text. In a second step, we then iterate through all references in the database and match them

[4] At the time of writing the MAG contains data on over 200 million publications.

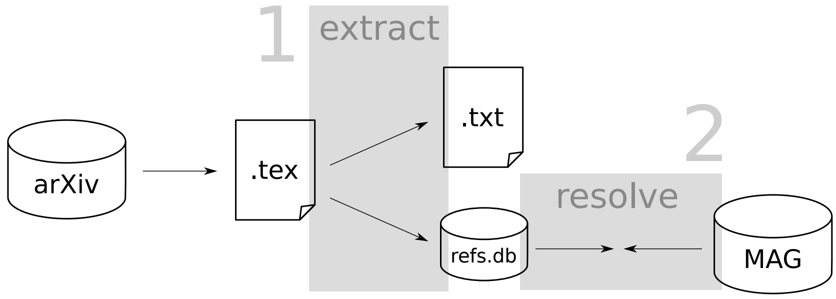


Figure 4: Schematic representation of the data set generation process.

against paper metadata records in the MAG. The result of this process are MAG paper records associated with one or more references, which in turn are associated with citation contexts in the plain text files. In other words, we end up with cited documents described by their MAG metadata and a distributed description of the document, consisting of citation contexts over many citing documents.

4.2.3 LaTeX parsing

In the following we will describe the tools considered for parsing LaTeX, the challenges we faced in general and with regard to arXiv sources in particular, and our resulting approach.

Table 3: Comparison of tools for parsing LaTeX.

Tool	Output	Robust	Usable w/o modification
plastex[5]	DOM	no	yes
TexSoup[6]	document tree	no	yes
opendetex[7]/detex[8]	plain text	no	yes
GrabCite [40]	plain text + resolved ref.	yes	no
LaTeXML[9]	XML	yes	yes
Tralics[10]	XML	yes	yes

[5]See `https://github.com/tiarno/plastex`.
[6]See `https://github.com/alvinwan/texsoup`.
[7]See `https://github.com/pkubowicz/opendetex`.
[8]See `https://www.freebsd.org/cgi/man.cgi?query=detex`.
[9]See `https://github.com/brucemiller/LaTeXML`.
[10]See `https://www-sop.inria.fr/marelle/tralics/`.

Tools

We took several tools for the conversion from LaTeX to plain text or to intermediate formats into consideration and evaluated them. Table 3 gives an overview of our results. Half of the tools failed to produce any output for a large amount of arXiv submissions we used as test input and were therefore deemed not robust enough. *GrabCite* is able to parse 78.5% of arXiv CS submissions but integrates resolving references against DBLP into the parsing process and therefore would require significant modification to fit our system architecture. *LaTeXML* and *Tralics* are both robust and can be used as LaTeX conversion tools as is. On subsequent tests we note that *LaTeXML* needs on average 7.7 seconds (3.3 if formula environments are heuristically removed beforehand) to parse an arXiv submission while *Tralics* needs 0.09. Because the quality of their output seemed comparable we chose to use *Tralics*.

Challenges

Apart from the general difficulty of parsing LaTeX due to its feature richness and people's free-spirited use of it, we especially note difficulty in dealing with extra packages not included in submissions' sources[11]. While *Tralics* is supposed to for example deal with *natbib* citations[12], normalization of such citations lead to a decrease of citation markers not being able to be matched to an entry in the document's reference section from 30% to 5% in a sample of 565,613 citations we tested.

[11] arXiv specifically suggest the omission of such (see `https://arxiv.org/help/submit_tex#wegotem`).
[12] See `https://www-sop.inria.fr/marelle/tralics/packages.html#natbib`.

Resulting approach

Our LaTeX parsing solution consists of two steps. First, we flatten each arXiv submission's sources to a single LaTeX file using *latexpand*[13,14] and normalize `\cite` commands to prevent parsing problems later on. In the second step, we then generate an XML representation of the LaTeX document using *Tralics*, replace formulas, figures, tables and intra-document references with replacement tokens and extract the plain text. Furthermore, each entry in the document's reference section is assigned a unique identifier, the reference string stored in a database and corresponding citation markers are placed in the plain text.

4.2.4 Reference resolution

Resolving references to globally consistent identifiers (e.g. detecting that the references (1), (2), and (3) in Listing 4.1 all reference the same document) is a challenging and still unsolved task [1]. Given it is, by itself, the most distinctive part of a publication, we base our reference resolution on the title of the cited work and use other pieces of information (e.g. authors' names) only in secondary steps. In the following we will describe the challenges we faced, matching arXiv submissions' reference strings against MAG paper records and how we approached the task.

Listing 4.1: Examples of reference strings.

```
(1) V. N. Senoguz and Q. Shafi, arXiv:hep-ph/0412102
(2) V.N. Senoguz and Q. Shafi, Phys. Rev. D 71 (2005) 043514.
(3) V. N. Şenoğuz and Q. Shafi, ‘‘Reheat temperature in supersymmetric
     hybrid inflation models,’’ Phys. Rev. D 71, 043514 (2005) [hep-ph
    /0412102].
(4) V.Sauli, JHEP 02, 001 (2003).
(5) Aaij, Roel, et al. "Search for the $B^{0}_{s} \to \eta^{\prime}\
    phi$ decay" Journal of High Energy Physics 2017.5 (2017): 158.
(6) According to the numerous discussions with my colleagues <removed>
     and <removed> an experimental verification of our theoretical
    predictions is feasible.
```

[13] See `https://ctan.org/pkg/latexpand`.

[14] We also tested flatex (`https://ctan.org/pkg/flatex`) and flap (`https://github.com/fchauvel/flap`) but got the best results with `latexpand`.

Challenges

Reference resolution can be challenging when reference strings contain only minimal amounts of information, when formulas are used in titles or when they refer to non publications (e.g. Listing 4.1 (4)-(6)). A further problem we encountered was noise in the MAG. Our mechanism matched 13,041 reference strings like `K. Kondo, hep-th/0303251.` and `T. Heinzl, hep-th/9812190.` to a MAG paper with the title *"hep-th."* with one of the author's names being *"He"* (paper ID 2811252340).

Resulting approach

Our reference resolution procedure can be broken down in two steps: title identification and matching. If possible, title identification is done by arXiv ID or DOI (where we retrieve the title from an arXiv metadata dump or via Crossref[15]); otherwise we use Neural ParsCit [53]. The identified title is matched against the normalized titles of all publications in the MAG. Resulting candidates are considered, if at least one of the author's names is present in the reference string. If multiple candidates remain, we judge by the citation count given in the MAG. The last step particularly helped mitigate rouge almost-duplicate entries in the MAG that often have few to no citations.

4.2.5 Result format

Listing 4.2 shows some example content from the data set. While the data set in and of itself consists of plain text files and a references database, citation contexts of successfully resolved references are straightforward to extract and use as input for a recommender. The bottom part of Listing 4.2 shows an example of a 3 sentence context with `cited doc MAG ID`, `MAG IDs of adjacent citations`, `citing doc arXiv ID` and `text` in a CSV file.

[15] See `https://www.crossref.org/`.

Citations are deemed adjacent, if they are part of a citation group or are at most 5 characters apart (e.g. *"[27,42]"*, *"[27], [42]"* or *"[27] and [42]"*).

Listing 4.2: Excerpts from (top to bottom) a plain text file, corresponding data base entries in the references DB, entries in the MAG and extracted citation context CSV.

```
It has over 79 million images stored at the resolution of FORMULA .
    Each image is labeled with one of the 75,062 non-abstract nouns in
     English, as listed in the Wordnet{{cite:9ad20b7d-87d1-47f5-aeed
    -10a1cf89a2e2}}{{cite: 298db7f5-9ebb-4e98-9ecf-0bdda28a42cb}}
    lexical database.

[uuid]          [in_doc]   [mag_id]     [reference_string]
9ad20b7d-87d1   1412.3684  2081580037   George A. Miller (1995). WordNe
-47f5-aeed-..                           t: A Lexical Database for Eng..
298db7f5-9ebb   1412.3684  2038721957   Christiane Fellbaum (1998), ""W
-4e98-9ecf-..                           ordNet: An Electronic Lexical..

[paperid]    [originaltitle]                             [publisher]  ...
2038721957   WordNet : an electronic lexical database   MIT Press    ...
2081580037   WordNet: a lexical database for English    ACM          ...

2038721957|2081580037|1412.3684|It has over 79 million images stored
    at the resolution of FORMULA . Each image is labeled with one of
    the 75,062 non-abstract nouns in English, as listed in the Wordnet
     CIT MAINCIT lexical database. It has been noted that many of the
    labels are not reliable CIT .
```

Table 4: Confidence intervals for a sample size of 300 with 297 positive results as given by Wilson score interval and Jeffreys interval [54].

Confidence level	Method	Lower limit	Upper limit
0.99	Wilson	0.9613	0.9975
	Jeffreys	0.9666	0.9983
0.95	Wilson	0.9710	0.9966
	Jeffreys	0.9736	0.9972

4.3 Evaluation of reference resolution

To evaluate the quality of our reference resolution results, we take a random sample of 300 matched reference strings and manually check if the correct record in the MAG was identified by our method. Given the 300 items, we obtained 3 errors, giving us an accuracy estimate of 96% at the worst, as shown in Table 4.

The three incorrectly identified references were as follows (MAG IDs in square brackets):

1. *"Eddy, J.A.: 1983, The maunder minimum - a reappraisal. Solar Phys. 89, 195. ADS."*
    - matched: `[2024069573]`
      "The Maunder Minimum" (John A. Eddy; 1976)
    - correct: `[2080336740]`
      "The Maunder Minimum: A reappraisal" (John A. Eddy; 1983)
2. *"J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. In Advances in Neural Information Processing Systems (NIPS), volume 16, pages 49–56, 2004."*
    - matched: `[2249237221]`
      "Support Vector Machines" (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani; 2013)
    - correct: `[2130698119]`
      "1-norm Support Vector Machines" (Ji Zhu, Saharon Rosset, Robert Tibshirani, Trevor J. Hastie; 2003)
3. *"D. T. Limmer and D. Chandler. The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water. The Journal of Chemical Physics, 135(13):134503, 2011."*
    - matched: `[2599889364]`
      "The Putative Liquid-Liquid Transition is a Liquid-Solid Transition in Atomistic Models of Water" (David Chandler, David Limmer; 2013)
    - correct: `[1977410206]`
      "The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water. II" (David T. Limmer, David Chandler; 2011)

In all three cases the misidentified document's title is contained in the correct document's title, and there is a large or complete author overlap between correct and actual match.

This shows that authors sometimes title follow-up work very similarly, which leads to hard to distinguish cases. Another observation that can be made, is that longer titles are more distinctive. As certain publication titles might be sub-strings of other publications' titles, a matching mechanism should always try to first match a long title before trying shorter candidates. The full details of the evaluation can be found in Appendix A.

4.4 Statistics and key figures

4.4.1 Creation process

We used an arXiv source dump containing all submissions up until the end of 2017 (1,340,770 documents). 100,240 of these were only available in PDF format, leaving 1,240,530 sources. Our pipeline output 1,151,707 (92.8%) plain text files, 1,018,976 (82.1%) of which contained citation markers. The number of reference strings identified is 35,053,329, for which 56,077,906 citation markers were placed within the plain text files. This first part of the process took 59 hours to run, unparallelized on a 8 core Intel Core i7-7700 3.60GHz machine with 60 GB of memory.

Of the 35,053,329 reference strings, we were able to match 14,046,239 (40.07%). For 33.14% of the reference strings we could neither find an arXiv ID or DOI, nor was Neural ParsCit able to identify a title. For the remaining 26.79% a title was identified but could not be matched with the MAG. Of the matched 14 million items' titles, 50.67% were identified via Neural ParsCit, 29.67% by DOI and 19.66% by arXiv ID. Of the identified DOIs 26.8% were found as is while 73.2% were heuristically determined[16]. The matching process took 103 hours, run in 10 parallel processes on a 64 core Intel Xeon Gold 6130 2.10GHz machine with 500 GB of memory.

[16] This was possible because the DOIs of articles in journals of the American Physical Society follow predictable patterns.

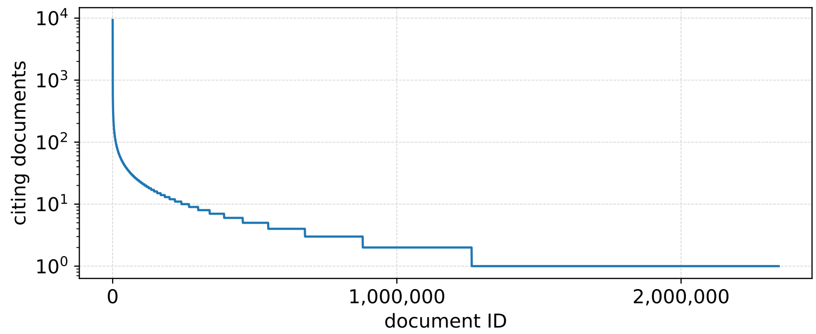


Figure 5: Number of citing documents per cited document.

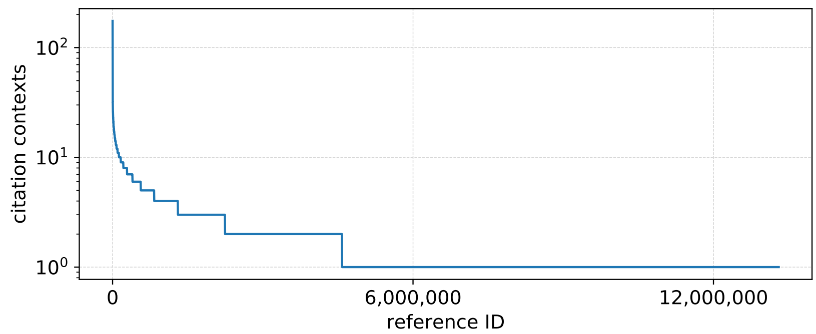


Figure 6: Number of citation contexts per reference.

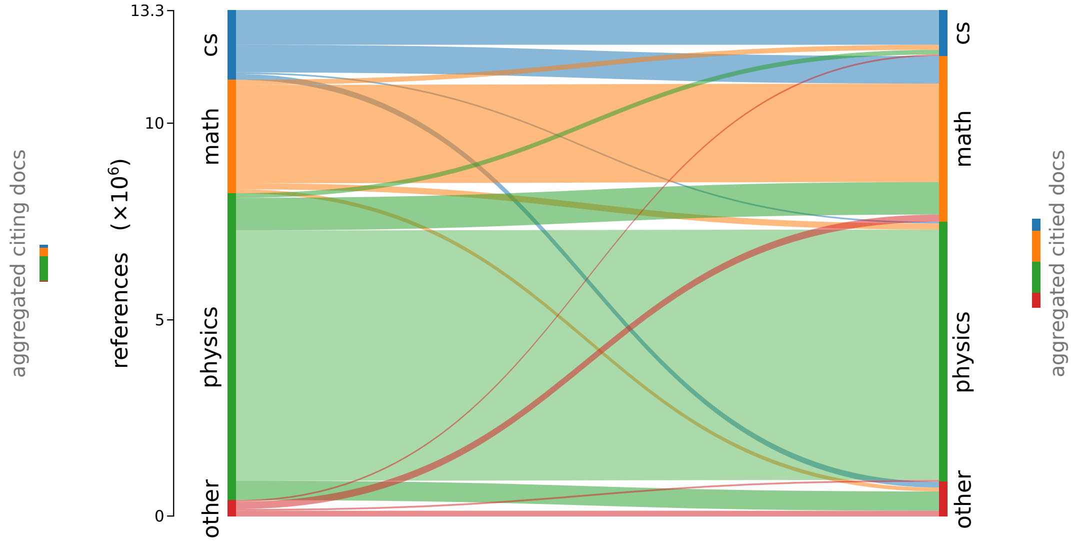


Figure 7: Citation flow by field of study for 13.3 million references. The number of citing and cited documents per field of study are plotted on the sides.

4.4.2 Resulting data set

The resulting data set consists of *2,343,585 cited papers, 926,644 citing papers, 13,303,373 references and 24,558,560 citation contexts.* Note that references with no citation markers (due to parsing errors) are not counted here.

Figure 5 shows the number of citing documents for all cited documents. There is one document with close to 10,000 citations, another three with more than 5,000 and another ten with more than 3,000. 1,262,861 (53.89%) of the documents have at least two citations, 547,036 (23.34%) have at least five. The mean number of citations is 5.68 (SD 26.82). Figure 6 shows the number of citation context per reference. 8,722,795 (65.57%) references have only one citation context, the maximum is 278, the mean 1.85 (SD 2.02). This means a cited document is described by on average $1.85 \times 5.68 \approx 10.5$ citation contexts.

Figure 7 depicts the flow of citations by field of study for all 13.3 million matched references. Fields of study with very small numbers of references are combined to *other* for legibility reasons. For the citing document's side, these are economics, electrical engineering and systems science, quantitative biology, quantitative finance and statistics. Combined on the cited document's side are chemistry, biology, engineering, materials science, economics, geology, psychology, medicine, business, geography, sociology, political science, philosophy, environmental science and art. To no surprise, publications in each field are cited the most from within the field itself. Notable is, however, that the incoming citations in mathematics are the most varied (physics and computer science combined make up 38% of the citations).

By generating our data set from LaTeX sources we were able to ensure clean text output as well as accurate citation information and exact citation marker positions. In terms of size it is closer to CiteSeer^{x} and PMC-OAS than the smaller data sets available. The fact that the data set spans multiple disciplines also allows for comparisons in citing behaviour between these disciplines. Because references are already resolved to MAG IDs, the data set is readily usable for recommendation tasks and allows for the use of rich metadata on both the citing and cited document side. Lastly, the embedding of citation markers in the full plain text of papers instead of pre-extraction enables users of the data set to choose and compare citation context lengths and variations at will.

5 Approach

In order to investigate the use of explicit semantic representations for the task of local citation recommendation we first need to decide which kinds of semantic constructs we want to model. As a starting point for this we looked into the field of citation context analysis [55]. A common task in this area is the classification of citation contexts by their polarity (positive/neutral/negative) and function (often based on the four dimensions identified by Moravcsik et al. [56], conceptual/operational, evolutionary/juxtapositional, organic/perfunctory, conformative/negational). Such approaches are primarily concerned with the *intent* of the author rather than the *content* of the citation context. We can therefore not expect to derive types of semantic constructs directly from citation functions. Starting from an established typology of citation functions will, however, ensure that we consider a wide range of different citations rather than cherry picking those that fit our preconceptions.

Table 5 lists categories of citation functions along with the kinds of semantic constructs that can be found in such citation contexts. The list of citation functions is taken from [59] (and therein built upon [30]). This study was selected because it gives an overview of previous attempts to classify citations, presents their new typology with extensive explanation as well as example contexts, and does not mix polarity into its function categories. Examining contexts from each of the eight functions, we identify two types of semantic constructs: named entities and claims. The rationale behind these two is as follows. Named entities can identify reference publications for a certain data set/tool/concept (see *Attribution*, *Statement of use* and *Application* in Table 5) as well as a method/field of study common to a selection

Table 5: Semantic constructs in citation contexts from a range of citation functions used in the field of citation context analysis.

Function	Construct	Examples (semantic construct *highlighted*)
Attribution	claim	"Berners-Lee et al. [57] argue that *structured collections of information and sets of inference rules are prerequisites for the semantic web to function*."
	NE	"A variation of this task is '*context-based co-citation recommendation*' [25]."
	-	"In [22] Duma et al. test the effectiveness of using a variety of document internal and external text inputs with a TFIDF model."
Exemplification	NE	"We looked into approaches to *local citation recommendation* such as [19–26] for our investigation."
Further reference	-	"See [58] for a comprehensive overview."
Statement of use	NE	"We use *CiteSeer^x* [18] for our evaluation."
Application	NE	"Using this mechanism we perform '*context-based co-citation recommendation*' [25]."
Evaluation	-	"The use of DBLP in [40] restricts their data set to the field of computer science."
Establishing links between sources	claim	"A common motivation brought forward for research on citation recommendation is that *finding proper citations is a time consuming task* [11, 19, 24, 25]."
	-	"Lamers et al. [32] base their definition on the author's name whereas Thompson [30] focusses on the grammatical role of the citation marker."
Comparison of own work with sources	claim	"Like [40] we find that, albeit written in a structured language, *parsing LaTeX sources is a non trivial task*."

of publications (see *Exemplification* in Table 5). Claims can identify publications that can be cited to back or support the very claim contained in a citation context. Note that the example contexts listed to have no construct ("-" in the *Construct* column) may contain named entities and claims as well (e.g. "DBLP" or "Lamers et al. base their definition of the author's name"), but these are (in the case of NEs) not representative of the cited work or (in the case of claims) just statements *about* a publication rather than statements being backed by the cited work. A third semantic construct that can be considered, but would require considering a larger citation context, is argumentative structures. To keep the scope of this thesis at a reasonable level we will limit our investigation to named entities and claims (see also Chapter 8).

The following sections will describe our investigation of entity based and claim based models for local citation recommendation.

5.1 Entity based recommendation

The intuition behind an entity based approach is, that there exists a reference publication for a named entity. Examples would be a data set ("CiteSeer^{x} [18]"), a tool ("Neural ParsCit [53]") or a concept ("Semantic Web [57]"). In a more loose sense this can also include publications being referred to as examples ("approaches to local citation recommendation [19–26]"). For the identification of such named entities we take two approaches. A more strict one based on the fields of study given in the MAG, and a more loose one based on noun phrases.

5.1.1 Fields of study in the MAG

Along with papers, authors, venues etc. the MAG data schema also includes fields of study (FoS) that are associated with papers and interlinked in a child-parent manner. At the time of writing there are 229,716 FoS at 6 levels of granularity. They range from the most coarse level 0 (example entries being *mathematics*, *sociology*, *computer science*) to more and more fine grained entities (*information retrieval*$_{(1)}$→*search engine*$_{(2)}$→*web search query*$_{(3)}$→*ranking (information retrieval)*$_{(4)}$→*Okapi BM25*$_{(5)}$). The levels of granularity don't seem to follow a globally consistent pattern though. *WordNet*, for example, being a particular piece of data in the same way *Okapi BM25* is a particular function, is not at level 5 but level 2 (*computer science*$_{(0)}$→*artificial intelligence*$_{(1)}$→*WordNet*$_{(2)}$). Another noteworthy aspect is that a FoS can have multiple parents (*search engine* for example has a second parent in *World Wide Web*).

Model The entity based representation of a citation context is the set of FoS that appear within the context. Formally, let $\mathcal{F}$ denote the set of FoS; then the entity based representation of a citation context c is the set of terms t defined as $R_{\text{FoS}}(c) = \{t | t \text{ appears in } c \wedge t \in \mathcal{F}\}$. Because of the hierarchical structure of FoS we experiment with augmenting the set by including the set members' parents into the representation. We find that his leads to worse results, presumably because a context's description becomes more vague which is detrimental to identifying reference publications or exemplifications. We furthermore look into only using a FoS when it directly precedes the citation marker—as it might be more relevant to the citation then—, but notice that such cases are too rare. In a preliminary test with 900k citation contexts, only 0.14% of our 180k test set items have a FoS in the required position matching any of the representations learned from the training set.

Recommendation For recommending documents based on R_{FoS} we use the Jaccard similarity between the input citation context and the aggregated citation contexts describing each candidate document. Formally, let c_i denote the input citation context and $\mathcal{D}$ be a set of documents with members $d \in \mathcal{D}$. Furthermore let $\varrho(d)$ be the set of citation contexts referencing d; $\varrho(d) = \{c | c \text{ references } d\}$. The Jaccard similarity then is defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, where $A = R_{\text{FoS}}(c_i)$ and $B = \bigcup\limits_{c \in \varrho(d)} R_{\text{FoS}}(c)$.

5.1.2 Noun phrases

For our second entity based model we take a more loose approach and treat noun phrases extracted from the arXiv data set as named entities. By filtering out items that appear only once, we end up with 2,835,929 noun phrases (NPs).

Model Similar to the FoS representation, we look at the NPs appearing within a citation context. To ensure a high descriptiveness, we only take into account maximally long matches. A context *"This has been done for language model training [27]"*, for example, would have "language model training" in its representation, but not "language model". Formally we can define $R_{\text{NP}}(c) = \{t|t \text{ appears in } c \land t \in \mathcal{P} \land t^{+pre} \notin \mathcal{P} \land t^{+suc} \notin \mathcal{P}\}$ where $\mathcal{P}$ is our set of NPs while t^{+pre} and t^{+suc} denote an extension of t using its preceding or succeeding word respectively. As an alternative representation we furthermore define $R^{2+}_{\text{NPmrk}}(c)$ as a subset of $R_{\text{NP}}(c)$ containing, if present, the NP of minimum word length 2 directly preceding the target citation marker in c. Formally, $R^{2+}_{\text{NPmrk}}(c) = \{t|t \in R_{\text{NP}}(c) \land len(t) \geq 2 \land t \text{ directly precedes } m\}$ where m is the citation marker in c that a prediction is to be made for.

Recommendation Recommending documents based on R_{NP} and R^{2+}_{NPmrk} is done using a vector space model (VSM) in which NP representations, treated like a bag-of-words (BoW), are compared by their cosine similarities. Representations of candidate documents are, likewise to the FoS based model, aggregates over all contexts referencing the document. Formally, the vector representation of a context is given by $V(R(c)) = (t_1, t_2, ..., t_{|\mathcal{P}|})$ where $\mathcal{P}$ is the set of all NPs and t_i is a non-negative integer representing a quantity with regards to the ith term in $\mathcal{P}$. Aggregated context representations for candidate documents are calculated by adding up all vector representations of the contexts referring to a document. I.e., let $\varrho(d)$ be the set of citation contexts referencing d, then d's vector representation is $\sum_{c \in \varrho(d)} V(R(c))$. The similarity between an input context c_i and a candidate document $d \in \mathcal{D}$ can then be calculated as the cosine θ between the two vector representations, their dot product over the product of their magnitudes, $\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$ where $A = V(R(c_i))$ and $B = \sum_{c \in \varrho(d)} V(R(c))$.

5.2 Claim based recommendation

Our claim based approach is motivated by the fact that citations are used to back up claims (see Table 5). These can, for example, be findings presented in a specific publication ("It has been shown, that ... [27].") or more general themes found across multiple works ("A common argument for X is, that ... [3-7]."). The following sections will describe our evaluation of tools for extracting claims from citation contexts and the model we developed for recommendation.

5.2.1 Tools for extracting claims

For the extraction of claims we consider a total of four state of the art [60] information extraction tools: PredPatt [61], Open IE 5.0 [62], ClausIE [63] and Ollie [64]. All of these can identify information in unformatted text, which they output in some form of predicate-argument tuples. Listing 5.1 shows the output of all four tools for the simple sentence *"The paper shows that context-based methods can outperform global approaches."*.

Listing 5.1: Information extraction tool output examples.

```
# PredPatt
?a shows ?b
    ?a: The paper
    ?b: SOMETHING := context-based methods can outperform global
        approaches
?a can outperform ?b
    ?a: context-based methods
    ?b: global approaches

# Open IE 5.0
0.90 Context(The paper shows,List([0, 15))):(context-based methods;
   can outperform; global approaches)
0.78 (The paper; shows; that context-based methods can outperform
   global approaches)
```

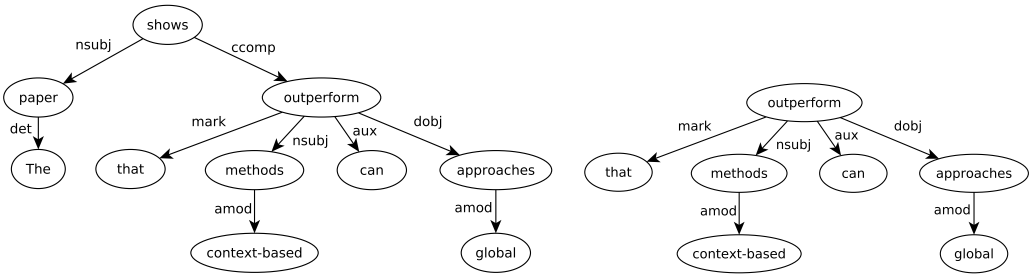
```
# ClausIE
"The paper" "shows" "that context-based methods can outperform global
    approaches"
"context-based methods" "can outperform" "global approaches"

# OLLIE
0.659: (context-based methods; can outperform; global approaches)[
    attrib=The paper shows]
```

Because the output for simple sentences is comarable for all of the tools, we tested them on 5 non trivial citation contexts, each consisting of 3 sentences. Based on this evaluation—the details of which are given in Appendix B—we deem PredPatt to perform the best, which is consistent with the observations in [60].

5.2.2 A predicate-argument model

Starting out with the predicate-argument tuples as PredPatt outputs them, we tested if there was enough of an overlap betweeen the components among citation contexts to facilitate recommendation. Taking the example from Listing 5.1, we would note the predicates *"show"* and *"can outperform"* as well as the arguments *"The paper"*, *"context-based methods can outperform global approaches"*, *"context-based methods"* and *"global approaches"*. Even when not preserving the associations between predicates and their arguments—e.g. looking for other contexts where *"global approaches"* was identified without requiring it to be in conjunction with *"can outperform"*—, we note very little overlap between citation contexts. We believe this to be the case because the predicates and especially arguments that PredPatt identifies can get very long, like *"can outperform"* (including the auxiliary verb *"can"*) and *"context-based methods can outperform global approaches"* (unlikely to appear in another citation context with the exact same wording).

To overcome this problem we develop a derivate model based on PredPatt's internal representation of sentences. Internally, PredPatt represents sentences in a tree structure based on Universal Dependencies (UD) [65] (i.e. a dependency grammar as opposed to a constituency grammar). This means words are connected by directed edges where the source is referred



(a) Main clause. **(b)** Sub clause.

Figure 8: UD trees as generated by PredPatt.

to as *governor*, the target is referred to as *dependant* and the edge itself is associated with a type of *relation*. Looking at Figure 8, depicting PredPatt's internal view on the example sentence given in Listing 5.1, we can, for example, see that the *"can"* is a dependant to and governed by *"outperform"* in an auxilliary relation. How our (predicate-argument based) claim model is constructed from this tree structure and used for recommendation will be discussed in the following sections.

Algorithm 1 Construction of $R_{\text{claim}}(c)$

```
c ← strip_quotation_marks(c)                    ▹ remove quotation marks
c ← merge_citation_markers(c)                   ▹ e.g. "CIT , CIT"→"CIT"
pp_trees_c ← predpatt(c)                        ▹ get PredPatt output
output ← []
resolve_rels ← ['name', 'goeswith', 'mwe',      ▹ (b1) UD relations
                'compound', 'conj', 'amod',
                'advmod']
foreach t ∈ pp_trees_c do                       ▹ for all claims identified
    root ← get_root(t)
    pred ← identify_predicate(t)                ▹ (a) resolve copula if present
    pred ← lemmatize(pred)
    foreach n ∈ traverse(t) do
        if pos_tag(n) == 'NOUN' then
            arg ← resolve_all(n, resolve_rels)  ▹ (b2) resolve compounds etc.
            output.append(pred + ':' + arg)     ▹ build pred:arg tuple
        end if
    end for
end for
return output
```

Model The construction of our claim representation $R_{\text{claim}}(c)$ for a citation context c is shown in Algorithm 1. There are two preprocessing steps prior to applying PredPatt. Quotation marks matching the regular expression `[‘“”’‚„"«»‹›《》〈〉]` are removed to avoid parsing errors, and groups of citation marker replacement tokens (e.g. *“CIT , CIT , CIT”* originating from e.g. *“[3,27,42]”*) are merged into one (i.e. *“CIT”*) to get a cleaner and easier to parse sentence structure. After applying PredPatt, the UD tree of each of the identified claims is traversed and predicate-argument tuples are generated. A claim is always centered around a predicate, which is the root of the corresponding UD tree unless a compula (*be*, *am*, *is*, *are*, *was*) is used, in which case the predicate is a dependant of the root node with the relation type *“cop”* (this is handled at marker **(a)** in Algorithm 1). Once identified, predicates are lemmatized using NLTKs WordNetLemmatizer. For the identification of useful arguments (markers $\mathbf{(b_1)}$ and $\mathbf{(b_2)}$ in Algorithm 1), we look at all nouns within the UD tree and resolve compounds (*“compound”*, *“mwe”*, *“name”* relations), phrases split by formatting (*“goeswith”*), conjunctions (*“conj”*) as well as adjectival and adverbial motifiers (*“amod”*, *“advmod”*). To give an example, the noun *“methods”* in both trees in Figure 8 has the adjectival modifier *“context-based”*. In such a case our model would not choose *“methods”* as an argument to *“outperform”* but *“context-based methods”*. Listing 5.2 shows the complete representation generated for the example sentence.

Listing 5.2: An example of $R_{\text{claim}}(c)$.

```
[
 'show:paper',
 'show:context based methods',
 'show:global approaches',
 'outperform:context based methods'
 'outperform:global approaches',
]
```

As becomes apparent when looking at Listing 5.2, we do not preserve the order of arguments. I.e., instead of constructing a triple (e.g. `context based methods:outperform:global approaches`) we build two tuples (`outperform:context based methods` and `outperform:global approaches`). This sequential invariance means a certain loss in precision, but also allows us to capture the similarity between active voice and passive voice sentences without the need to ensure reliabe parsing of such grammatical constructs. The input *"The paper shows that global approaches can be outperformed by context-based methods."*, for example, also has the claim representation shown in Listing 5.2.

We also experiment with models taking into account the citation marker's role within the sentence. The motivation behind this is, that there are cases where just a certain subclause is descriptive of a cited work. To give an example, in a sentence of the form *"While A et al. argue that X is well suited for Y [27], in [42] B et al. come to the conclusion that Z is a promising application area for X."* we ideally would only consider *"X is well suited for Y"* for the representation of *[27]* and *"Z is a promising application area for X"* for *[42]*. Unfortunately, markers of non-syntactic citations (*[27]* in above example) are not correctly parsed by PredPatt and could therefore not reliably be associated with a part of the parse tree. There also seems to be no straightforward way to distinguish syntactic from non-syntactic citations, which would have made a preprocessing step possible. Given these circumstances we leave the development of marker-aware claim models as future work.

Recommendation To make use of R_{claim} for recommendation we use a VSM where each dimension is a predicate-argument tuple. Similarities within this VSM are calculated as cosines between TFIDF-weighted vectors. Formally, the vector representation of a context is given by $V(R(c)) = (t_1, t_2, ..., t_{|\mathcal{T}|})$ where $\mathcal{T}$ is the set of all predicate-argument tuples and t_i is a non-negative integer representing a quantity with regards to the ith tuple in $\mathcal{P}$. TFIDF values for each t_i in $V(R(c))$ with respect to a collection of contexts C are calculated as $TFIDF(t, c, C) = tf_{t,c} \times idf_{t,C}$, where $tf_{t,c}$ denotes the number of occurences of t in c, and $idf_{t,C} = log\frac{|C|}{|\{c \in C | t \text{ appears in } c\}|}$ is the log of the number of contexts over the number of contexts containing t.

Likewise to R_{NP}, aggregated context representations for candidate documents are caluclated by adding up all vector representations of the contexts referring to a document. Again, let $\varrho(d)$ be the set of citation contexts referencing d, then d's vector representation is $\sum_{c \in \varrho(d)} V(R(c))$. The similarity between an input context c_i and a candidate document $d \in \mathcal{D}$

can then be calculated as the cosine θ between the two vector representations $\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$ where $A = TFIDF(V(R(c_i)))$ and $B = TFIDF\Big(\sum_{c \in \varrho(d)} V(R(c))\Big)$.

6 Evaluation

To measure the effectiveness of our models, we evaluate them in an offline evaluation as well as a user study. The details of these can be found in Section 6.2 and 6.3 respectively. Before that, we will discuss the peculiarities of evaluating approaches to citation recommendation in Section 6.1.

6.1 Special considerations for citation recommendation

The goal in many recommendation domains is to identify items a user is likely to engage with, like finding products a user might buy or media a user might want to consume. The relevance judgement of a recommended item in such a case is merely based on the user's subjective taste. With citations there is more to it than just preference. For a recommended citation to be useful to a researcher it has to be appropriate. To give an example, if the video platform PeerTube recommends watching *Paywall: The Business of Scholarship*, there is no ground for arguing this recommendation could never be valid. However, if a citation recommendation system suggests to cite *Alan Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem", 1937* in the context *"We use CiteSeer^x [] for our evaluation."*, then this is arguably not usable. This circumstance makes the evaluation of approaches to citation recommendation comparably difficult, because the accurate judgement of relevance (and by extension applicability) of a recommended item for a given input requires expert knowledge. This has to be kept in mind when automatically generating ground truths for offline evaluations and also when interpreting evaluation results.

A common source of ground truths for automatically evaluating recommender systems is existing relevance judgements (e.g. movies with user ratings logged on a website). For evaluation these are used as an input to the system and the output is compared with the actual judgement. The equivalent in citation recommendation is to harvest citation contexts from existing publications and trying to re-predict the citations originally contained (see Section 3.2.1). While in the case of a movie recommendation, the user's rating at the point in time it was made is *the* truth—within the bounds of the user's introspective capabilities—, a citation, albeit in a published work, can be flawed or may be one of several possibly valid ones. Awareness of this is helpful for interpreting evaluation results in re-prediction scenarios and also the motivation behind alternative evaluation metrics.

Another aspect of automatically harvested ground truth is change over time. In our movie recommendation example this can manifest in a user's taste changing. In scholarly discourse, a cited documents role—that is, how or if it is cited—can develop over time [66,67]. Examples of causes are concepts becoming obsolete through new discoveries (e.g. the expanding Earth [68]), discoveries becoming named retrospectively (e.g. Lotka's law [69]) and important work within a field becoming established knowledge and not being referred to anymore (e.g. physicists not citing Einstein's 1905 paper *"Zur Elektrodynamik bewegter Körper"* when talking about the theory of relativity). This observation motivates harvesting citation contexts for ground truths preferably from recent publications which can be achieved through a temporal split of training and test data.

A last important consideration has to be made concerning the split of training and test data. A naïve way of doing this would be to group citation contexts by cited document and then split each group into, for example, 80% training and 20% test data. This would ensure all cited documents are tested for. However, it would also lead to cases where citation contexts from a single publication are split. This would mean an author's word choice, writing style, etc. could be an unwanted signal informing the recommendation. Fortunately, a temporal split of training and test data as mentioned in above paragraph (i.e. splitting training and test data by publication date of citing documents) automatically solves this problem.

6.2 Offline evaluation

Our offline evaluation, in total, encompasses seven different models and four data sets. First, we will introduce the data sets used. We employ citation contexts from arXiv, the MAG, RefSeer and ACL (see also Chapter 4). Note that in all four cases we do not use all contexts available in the data set but rather apply filter criteria, the details of which can be found in Appendix C. Table 6 gives an overview of key properties of the resulting data. Given are the temporal split between train and test set, the number of candidates to rank (documents), number of these tested for (documents), the number of test set items (contexts) and the mean number of citation contexts trained on per recommendation candidate (contexts per document). This last measure is furthermore visualized in Figure 9. The amount of candidates to rank for each recommendation (#Cand.) and how well these candidates are described (CC/RC) gives insight into how difficult the recommendation task for each of the data sets is. Comparing, for example, arXiv and MAG under the assumption the data is equally clean (which is not necessarily the case), recommendation using the MAG data should be the easier task, because it involves a comparable number of candidates to rank, but they are significantly better described. The number of cited documents tested for (#Tested) and number of test set items (#Test set items) are indicative of how comprehensive

Table 6: Key properties of data sets used for evaluation.

Data set	Train/test split	#Cand.	#Tested	#Test set items	Mean CC/RC (SD)
arXiv	≤2016 / ≥2017	63,239	49,980	490,018	21.7 (51.2)
MAG	≤2017 / ≥2018	81,320	19,037	141,631	104.1 (198.6)
RefSeer	≤2011 / ≥2012	184,539	17,323	53,401	18.2 (47.0)
ACL-ARC	≤2005 / =2006	2,431	1,089	3,881	6.8 (9.5)

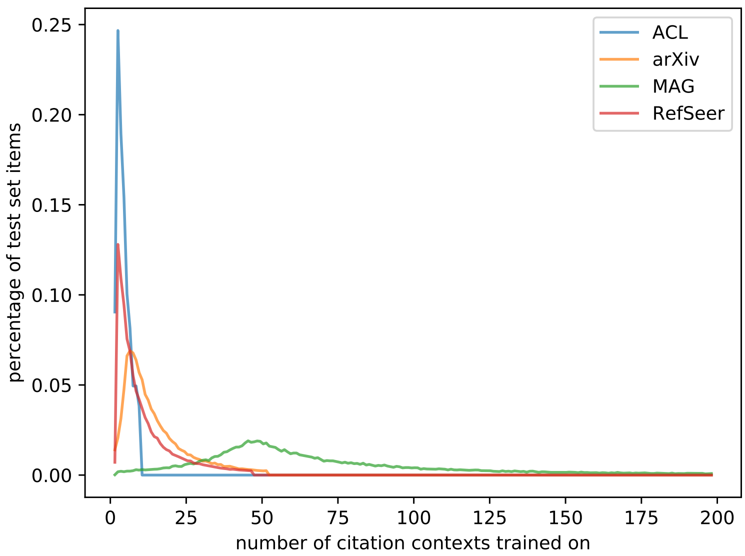


Figure 9: Number of citation contexts per recommendation candidate.

the evaluation is. Testing for 50k cited documents using 490k contexts (arXiv) can, for example, be considered more thorough than testing for 2.4k cited documents using 3.8k contexts (ACL-ARC).

The models used for recommendation are the four presented in Chapter 5, R_{FoS}, R_{NP}, R^{2+}_{NPmrk} and R_{claim}—which we will refer to as FoS, NP, NPmarker, and Claim from here on for simplicity—as well as a bag-of-words baseline (BoW) and two combined models (BoW+fosboost and Claim+BoW). The BoW baseline involves removal of punctuation as well as stop words and uses TFIDF for ranking. We choose TFIDF over BM25 because the former shows better performance in our evaluations. BoW+fosboost is constructed by taking the BoW ranking and then moving all candidates, where the overlap in fields of study ($\left|R_{\text{FoS}}(c_i) \cap \left(\bigcup_{c \in \varrho(d)} R_{\text{FoS}}(c)\right)\right|$, cf. Section 5.1.1) is maximal, one position upwards in the ranking. Ranking in Claim+BoW is done by linearly combining similarity values $S(A, B) = \sum_{m \in \mathcal{M}} \alpha_m S_m(A, B)$ of the models $\mathcal{M} = \{\text{Claim}, \text{BoW}\}$ with heuristically determined coefficients $\alpha_{\text{Claim}} = 1$ and $\alpha_{\text{BoW}} = 2$.

Table 7: Evaluation scores at cut-off 5 for all models using the arXiv data.

Model	Recall@5	MRR@5	MAP@5	NDCG@5
Claim+BoW	**0.27420**	0.17530	**0.34416**	**0.21829**
BoW+fosboost	0.25824	0.16264	0.32980	0.20682
BoW	0.25817	0.16256	0.32961	0.20673
NPmarker	0.24931	**0.17764**	0.30769	0.17677
Claim	0.11750	0.07477	0.14441	0.08337
NP	0.10558	0.06775	0.13950	0.07914
FoS	0.00933	0.00420	0.01999	0.00921

Table 7 shows Recall, MRR, MAP and NDCG values at cut-off 5 for all models evaluated on the arXiv data. While FoS alone performs very poorly, the combined model BoW+fosboost consistently performs better then BoW alone. Nevertheless, because the performance gain is minute, we see no reason for a more detailed investigation of the models FoS and BoW+fosboost. Overall Claim+BoW shows the best performance, with NPmarker achieving the highest value for the MRR metric. To investigate the performance of these models further, we look at above metrics at different cut-off values and using multiple data sets. Note that arXiv is the only of the data sets where the NPmarker model is applicable, because the exact position of the citation marker is needed. Also note that for the arXiv data, we count adjacent citations (see Section 4.2.5) as relevant when calculating MAP scores and give them a relevance of 0.5 in the NDCG calculation.

Figures 10–13 show the results of abovementioned evaluations. For each of the data sets Claim+BoW outperforms the BoW baseline in each measure and for all cut-off values. Claim and NP do not compare in performance with the two aforementioned. This suggests that the claim structures we model with R_{claim} are not enough for well performing recommendations on their own, but do capture important information that non-semantic models (BoW) miss. NPmarker, only present in Figure 10, gives particularly good results for lower cut-offs and performs especially well in the MRR metric. It performs the worst at high cut-offs measured by NDCG. Note that NPmarker is only evaluated for test set items where the model was

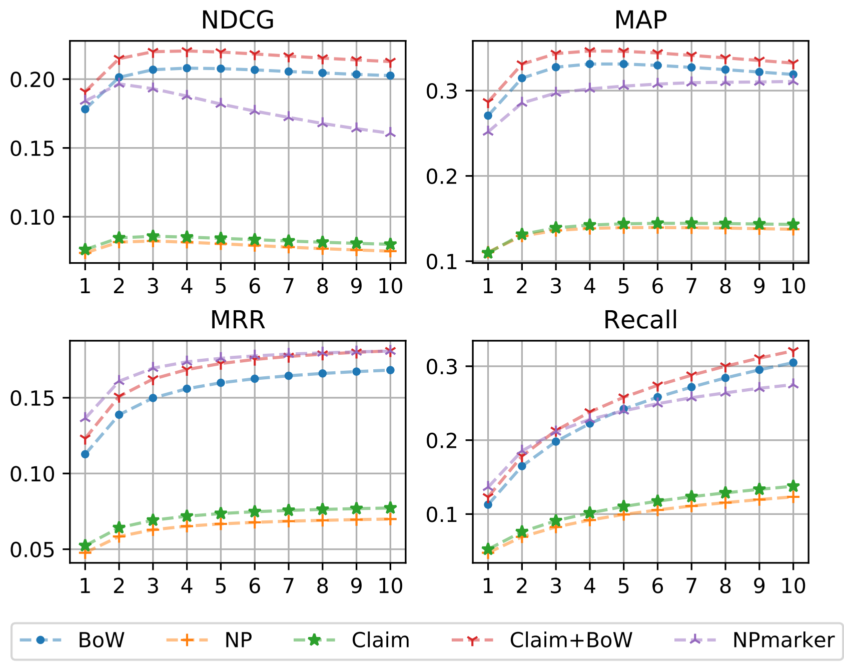


Figure 10: Evaluation using arXiv.

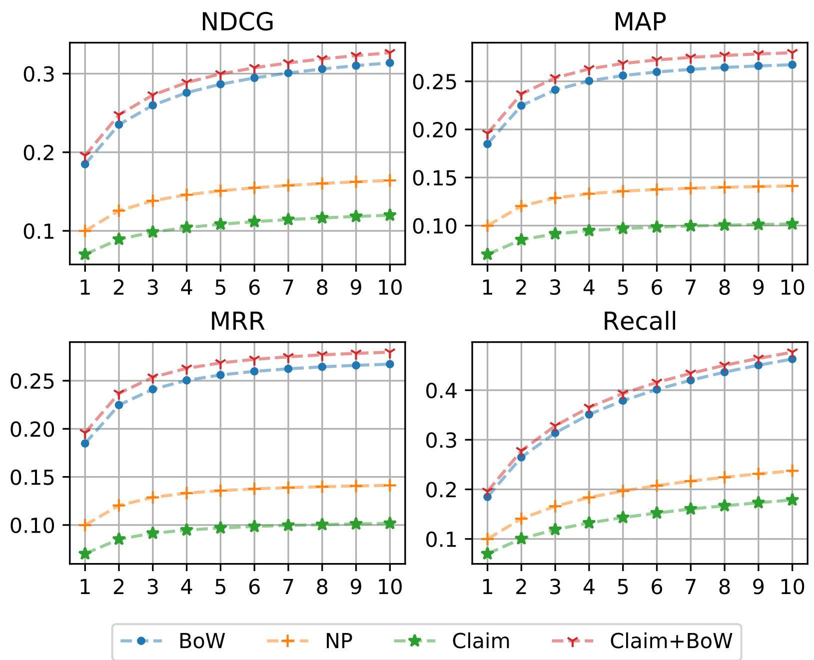


Figure 11: Evaluation using the MAG.

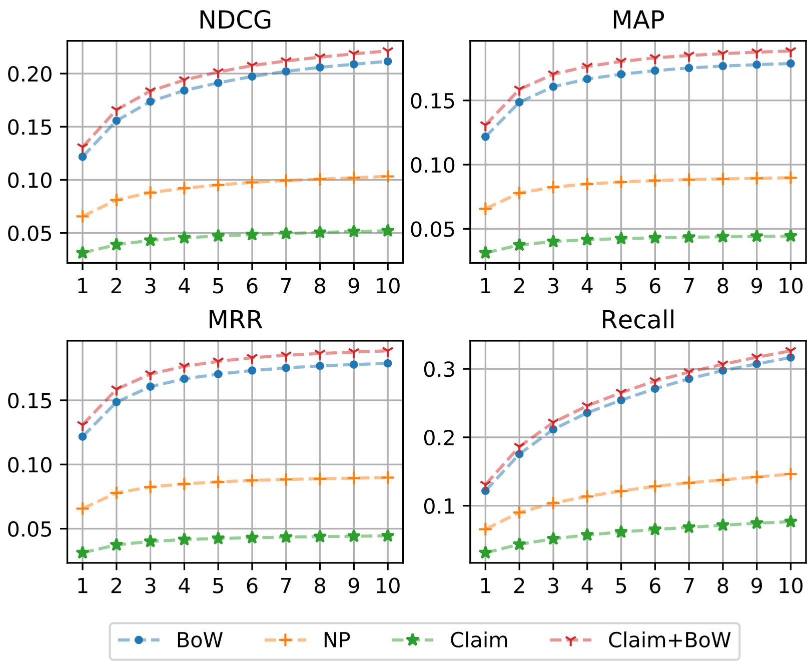


Figure 12: Evaluation using RefSeer.

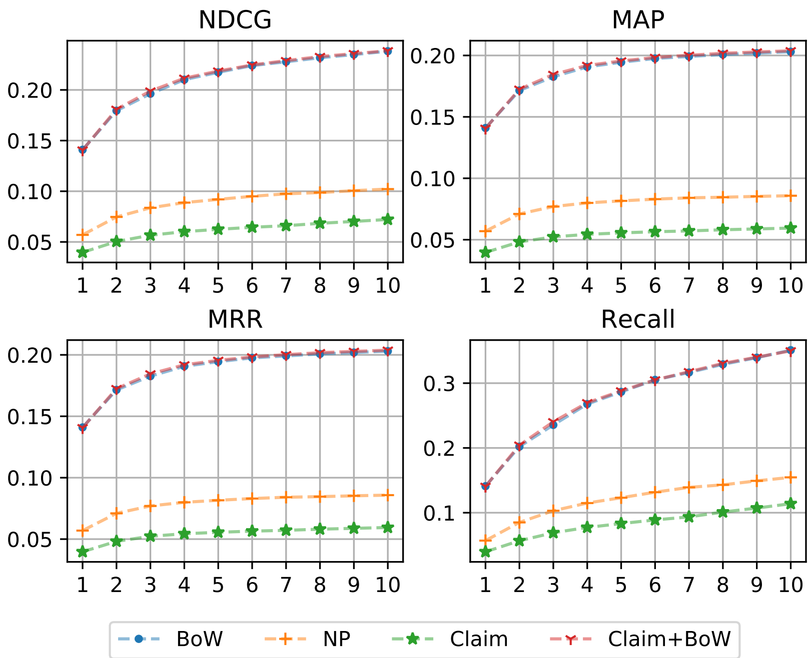


Figure 13: Evaluation using ACL-ARC.

applicable (i.e. where a noun phrase of minimum length 2 is directly preceding the citation marker; cf. Section 5.1.2). For our evaluation this was the case for 100,308 out of the 490,018 test set items (20.5%). The evaluation results for NPmarker indicate that it is comparatively well suited to recommend citations where there is one particularly fitting publication (e.g. a reference paper for a tool, method or data set) and less suited for exemplifications (cf. Table 5 in Chapter 5). Comparing values between the different data sets, we can confirm our expectations based on the values in Table 6. For the MAG data, where we ensured very well described recommendation candidates by setting a threshold, the models perform best. This shows us that differences in data set composition can make a large difference in evaluation results. As a consequence, comparison between approaches is only realistically possible when using the same data.

6.3 User study

To get more insight into the nature of our evaluation data, as well as a better understanding of our models, we perform a user study in which 2 human raters judge input-output pairs of our offline evaluation (i.e. citation contexts and the recommendations given for them). For this, we randomly choose 50 citation contexts from our arXiv data, where the citing paper belongs to the category "Computer Science - Information Retrieval". This is to ensure that raters are able to reasonably judge the relevance of the recommendations given. For each input context, we show the top 5 recommendations of the BoW, Claim+BoW and NPmarker models—i.e. 750 recommended items in total. We furthermore let the raters judge the type of citation (Claim, NE/concept, Exemplification, Other; cf. Table 5 in Chapter 5), whether or not the name of the cited document's author is present in the citation contexts and whether or not the citation marker has a grammatical function (i.e. syntactic/non-syntactic citation). Raters can also choose not to judge a context and give one of two reasons: (1) there's not enough information contained to reasonably make a decision, (2) the rater can not make a relevance judgement for some other reason ("pass"). Figure 14 shows the interface used to record the judgements. Likewise to the offline evaluation, the NPmarker model was

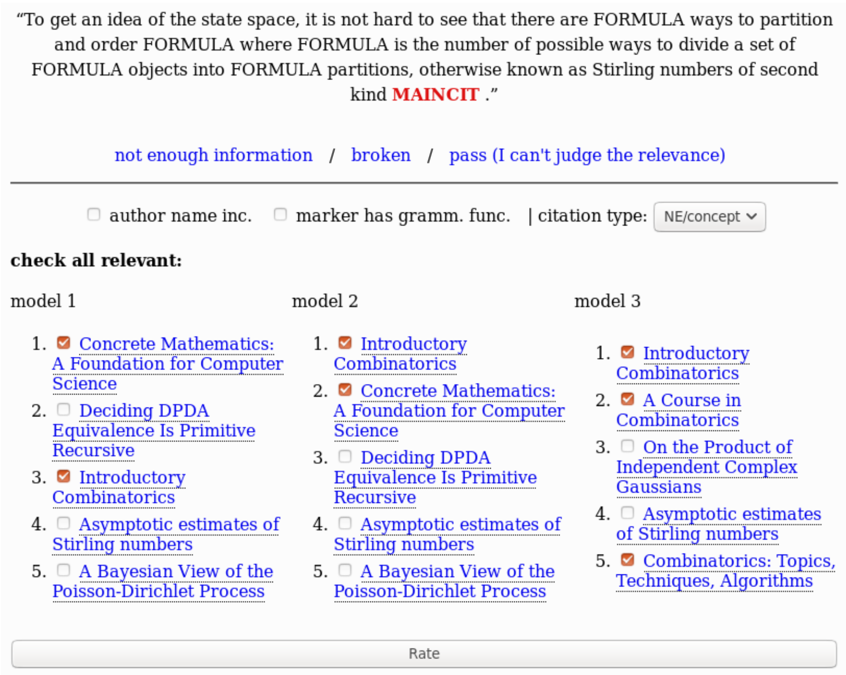


Figure 14: Interface used in the user study.

only evaluated when applicable (19 out of the 50 contexts). When no NPmarker based recommendations where available, the text "Model not applicable" was displayed in their stead.

Overall, we observe an inter-rater agreement of 87.8%. Judgements on the presence of an author's name and the citation being syntactic or not matched 100%, relevance judgements were 86.9% consistent and type categorization 73.2%. Of the 50 contexts, 8 contained an author's name and 2 were syntactic. This shows us that non-syntactic citations seem to be the norm and that author's names might be a confounding variable in recommendation that could be eliminated, by, for example, heuristically removing names from citation contexts before they are trained and evaluated on. Because we observe the lowest inter-rater agreement for citation types, we generate a confusion matrix, which is presented in Table 8[1]. Apparently, the distinction between the types is not always clear cut. This can be seen with a citation context like *"Convolutional neural networks have also been shown*

[1]Note that the values within the matrix do not add up to 50, because type information was not recorded when recommendation relevance was deemed not judgeable and the item therefore skipped.

Table 8: Confusion matrix of citation types.

		Rater 1: Claim	NE/concept	Exemplific.	Other
Rater 2	Claim	8	0	1	0
	NE/concept	1	13	2	0
	Exemplific.	2	1	8	1
	Other	2	0	0	2

Table 9: Evaluation scores at cut-off 5.

Model	Recall@5	MRR@5	MAP@5	NDCG@5
all contexts				
Claim+BoW	**0.50**	0.44	0.40	0.44
BoW	0.47	**0.45**	**0.42**	**0.46**
NPmarker	0.34	0.38	0.35	0.37
only contexts of type "claim"				
Claim+BoW	**0.61**	0.51	0.46	0.51
BoW	0.55	**0.55**	**0.53**	**0.56**
NPmarker	0.17	0.17	0.17	0.17
only contexts of type "NE/concept"				
Claim+BoW	0.42	0.38	0.35	0.38
BoW	0.42	0.38	0.34	0.38
NPmarker	**0.53**	**0.60**	**0.61**	**0.56**
only contexts of type "exemplification"				
Claim+BoW	**0.54**	0.44	0.38	0.45
BoW	0.49	**0.46**	**0.43**	**0.48**
NPmarker	0.24	0.29	0.29	0.29
only contexts of type "other"				
Claim+BoW	0.50	0.50	0.50	0.50
BoW	0.50	0.50	0.50	0.50
NPmarker	0.50	0.50	0.45	0.48

to be highly effective for natural language processing and have achieved excellent results in information retrieval [1], semantic parsing [2], sentence modeling [3] and other traditional natural language processing tasks [4].”, which can equally be interpreted as claims being backed or a list of examples.

Table 9 shows the results of the combined relevance judgements. Before discussing the results themselves, we need to point out a noteworthy difference to the offline evaluation. This is the handling of the number of relevant items. In a simple re-prediction setting (cf. Section 3.2.2) there is always only one relevant item. Furthermore including adjacent citations, as we did with our arXiv data, increases this number, but does certainly not guarantee an exhaustive list of possible citations. Setting the number of possibly relevant items to the cut-off value is also not an option, because for reference publications there often is just one relevant item (e.g. Neural ParsCit [53]). We approached this problem, by setting the number of relevant items for a given context to the maximum number of relevant (as judged by the raters) items identified by any of the models. To give an example, in the case shown in Figure 14 this would be 3, because the rightmost model recommended 3 relevant items, resulting in Recall values of 0.67, 0.67 and 1.

As for the resulting numbers (see Table 9), we present measurements for all contexts, as well as each of the citation classes on its own. We note that Claim+BoW and BoW are close, but in contrast to the offline evaluation, Claim+BoW only outperforms BoW in the Recall metric. In the case of NE/concept type citations, the NPmarker model performs better than the other two models in all metrics. Furthermore, we can see that both Claim+BoW and NPmarker achieve their best results for the type of citation they’re designed for—Claim and NE/concept respectively. This indicates that both models actually capture the kind of information they’re conceptualized for. Comparing to the offline evaluation, we measure higher numbers overall. While the user study is of considerably smaller scale and a direct comparison therefore not necessarily possible, the notably higher numbers indicate, that a re-prediction setting, even when adjacent citations are considered, involves a non-negligible number of false negatives (actually relevant recommendations counted as not relevant).

7 Conclusion

In the field of local citation recommendation, the explicit semantic modelling of citation contexts is not well explored yet. In order to investigate the merit of such approaches, we generated a new data set from arXiv LaTeX sources with accurate citation data as well as full text paper contents. Using this data set we developed semantic models of citation contexts based on entities as well as claim structures. We then evaluated our models on several data sets in a citation re-prediction setting and furthermore conducted a small user study. One of the entity based models, NPmarker, which captures noun phrases preceding the citation marker, performs best at low cut-offs and in the MRR metric. Low cut-offs and measuring the MRR can be interpreted as emulating citations for reference publications. This interpretation is also backed by the results of the user study, where NPmarker outperformed all other models when recommending for citation contexts, that referenced a NE or concept. We therefore conclude that NPmarker is well suited for recommending such types of citations. Our claim based model on its own does not compare in performance to a BoW baseline, but outperforms aforementioned when combined with it (Claim+BoW). We take this as an indication that the model encodes important information which the non-semantic BoW model is not able to capture. In the user study Claim+BoW performs best for citation contexts, in which a claim is backed by the target citation. This suggests that the model indeed captures information related to claim structures.

Overall, we note that when considering citations *in general*, it is far from trivial to outperform tried and tested information retrieval techniques like a TFIDF weighted BoW model. This may not be that surprising though, as citations come in many forms, and for each of them different aspects of the citation context are important. Significant improvements in citation recommendation might therefore require taking into account these differences, for example through the application of specialized models preceded by a classification step. The evaluation of our models indicates their suitability for recommending certain types of citations. The further development of these models and research on the classification of citations are therefore a promising direction of pursuit.

8 Future Work

Several parts of the work presented merit continued development, others lay bare questions yet to be answered. In closing, we want to briefly highlight these with regards to the newly created data set, the entity as well as claim based models, and finally semantic citation recommendation in general.

The flexible nature of our data set makes it suitable for the development and evaluation of various citation based tasks, not just the one presented here. Because arXiv is an established resource in active use, new publications are added to it every day. These new submissions can be a valuable addition to our data set. In parallel to the development and evaluation of our models, we already extended the data with all publications from 2018 [70]. Apart from the addition of new papers, the inclusion of mathematical formulas in some way, shape, or form (instead of replacing them with a substitution token) would presumably improve recommendation, as they can contain key information and are abundant in mathematics and physics papers. Such improvements can be guided by existing arXiv based efforts in mathematical information retrieval [71, 72].

Because the arXiv data set spans several disciplines, a comparative analysis of the performance of our models could reveal interdisciplinary differences. As shown in our offline evaluation, the size and composition of a data set can have a large impact on the results. A comparative analysis would therefore require the adjustment of the data set's math and physics subsets in terms of number of recommendation candidates and number of citing papers describing those.

Concerning our NPmarker model, an evaluation on more than just the arXiv data is needed. Because the model relies on the exact position of the citation marker, which is not given in the other data sets used, a heuristic for the marker's identification could be used. Another possibility would be the use of PMC-OAS data, as their JATS XML files also provide exact citation marker positions. Our Claim model in its current state uses the lemmatized form of predicates. These could be further generalized into a number of semantic relation types as done in [73]. Given the models current *predicate-argument* structure, it might be possible to implement a basic semantic citation search on top of it, allowing users to, for example, search for publications showing that something *"is:NP-hard"* or where the authors *"improve:local citation recommendation"*. A last point concerning the claim based model is the realization of grammatical citation marker awareness. This requires proper handling of non-syntactic citations, which, judging by our user study results, make up the vast majority of citations. The algorithm presented in [34] can be used as a starting point for identifying non-syntactic citations, but would have to be properly evaluated. The same holds true for their method of using the head of the nearest noun phrase as the new position of the marker.

Generally speaking, we think a more thorough and systematic examination of citation types—that is, structural types with respect to the grammatical and character-level location of information most descriptive of the cited document—should be a first step towards refining our existing models and creating new, citation type specific approaches.

In the long run, models revolving around claims could furthermore benefit from an assessment of credibility [74]; and beyond the level of claims, the modelling of argumentative structures, informed by existing work in the field of argumentation mining [75–77], could enable even more refined recommendation results.

Bibliography

[1] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Information extraction from scientific articles: a survey," *Scientometrics*, vol. 117, pp. 1931–1990, Dec. 2018.

[2] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, pp. 305–338, Nov. 2016.

[3] S. Buckingham Shum, E. Motta, and J. Domingue, "ScholOnto: an ontology-based digital library server for research documents and discourse," *International Journal on Digital Libraries*, vol. 3, pp. 237–248, Oct. 2000.

[4] S. Peroni and D. Shotton, "FaBiO and CiTO: Ontologies for describing bibliographic resources and citations," *Journal of Web Semantics*, vol. 17, pp. 33 – 43, 2012.

[5] S. Huh, "Journal Article Tag Suite 1.0: National Information Standards Organization standard of journal extensible markup language," *Science Editing*, vol. 1, no. 2, pp. 99–104, 2014.

[6] M. Y. Jaradeh, S. Auer, M. Prinz, V. Kovtun, G. Kismihók, and M. Stocker, "Open Research Knowledge Graph: Towards Machine Actionability in Scholarly Communication," 2019. arXiv:1901.10816 [cs].

[7] A. Abu-Jbara, J. Ezra, and D. Radev, "Purpose and Polarity of Citation: Towards NLP-based Bibliometrics," in *Proceedings of the 2013 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 596–606, Association for Computational Linguistics, 2013.

[8] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic Classification of Citation Function," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, (Stroudsburg, PA, USA), pp. 103–110, Association for Computational Linguistics, 2006.

[9] M. Berger, K. McDonough, and L. M. Seversky, "Cite2vec: Citation-Driven Document Exploration via Word Embeddings," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, pp. 1–1, Jan. 2016.

[10] K. D. Bollacker, S. Lawrence, and C. L. Giles, "CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications," in *Proceedings of the Second International Conference on Autonomous Agents*, AGENTS '98, (New York, NY, USA), pp. 116–123, ACM, 1998.

[11] Q. He, D. Kifer, J. Pei, P. Mitra, and C. L. Giles, "Citation Recommendation Without Author Supervision," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, (New York, NY, USA), pp. 755–764, ACM, 2011.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[13] F. Zarrinkalam and M. Kahani, "SemCiR: A citation recommendation system based on a novel semantic distance measure," *Program: Electronic Library and Information Systems*, vol. 47, pp. 92–112, 2013.

[14] S. E. Middleton, D. D. Roure, and N. Shadbolt, "Capturing knowledge of user preferences: ontologies in recommender systems," in *Proceedings of the 1st International Conference on Knowledge Capture*, pp. 100–107, 2001.

[15] M. Zhang, W. Wang, and X. Li, "A Paper Recommender for Scientific Literatures Based on Semantic Concept Similarity," in *Digital Libraries: Universal and Ubiquitous Access to Information*, (Berlin, Heidelberg), pp. 359–362, Springer Berlin Heidelberg, 2008.

[16] Y. Jiang, A. Jia, Y. Feng, and D. Zhao, "Recommending academic papers via users' reading purposes," *RecSys'12 - Proceedings of the 6th ACM Conference on Recommender Systems*, Sept. 2012.

[17] F. Zarrinkalam and M. Kahani, "A multi-criteria hybrid citation recommendation system based on linked data," in *2012 2nd International eConference on Computer and Knowledge Engineering (ICCKE)*, pp. 283–288, IEEE, 2012.

[18] C. Caragea, J. Wu, A. Ciobanu, K. Williams, J. Fernández-Ramírez, H.-H. Chen, Z. Wu, and L. Giles, "CiteSeer^{x}: A Scholarly Big Dataset," in *Advances in Information Retrieval*, (Cham), pp. 311–322, Springer International Publishing, 2014.

[19] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware Citation Recommendation," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), pp. 421–430, ACM, 2010.

[20] W. Huang, P. Mitra, and C. L. Giles, "RefSeer: A citation recommendation system," in *IEEE/ACM Joint Conference on Digital Libraries*, pp. 371–374, Sept. 2014.

[21] W. Huang, Z. Wu, C. Liang, P. Mitra, and C. L. Giles, "A Neural Probabilistic Model for Context Based Citation Recommendation," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2404–2410, AAAI Press, 2015.

[22] D. Duma and E. Klein, "Citation resolution: A method for evaluating context-based citation recommendation systems," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 358–363, 2014.

[23] D. Duma, E. Klein, M. Liakata, J. Ravenscroft, and A. Clare, "Rhetorical Classification of Anchor Text for Citation Recommendation," *D-Lib Magazine*, vol. 22, 2016.

[24] T. Ebesu and Y. Fang, "Neural Citation Network for Context-Aware Citation Recommendation," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, (New York, NY, USA), pp. 1093–1096, ACM, 2017.

[25] Y. Kobayashi, M. Shimbo, and Y. Matsumoto, "Citation Recommendation Using Distributed Representation of Discourse Facets in Scientific Articles," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, (New York, NY, USA), pp. 243–251, ACM, 2018.

[26] C. Jeong, S. Jang, H. Shin, E. Park, and S. Choi, "A Context-Aware Citation Recommendation Model with BERT and Graph Convolutional Networks," 2019. arXiv:1903.06464 [cs].

[27] C. R. Myers, "Journal citations and scientific eminence in contemporary psychology.," *American Psychologist*, vol. 25, no. 11, pp. 1041–1048, 1970.

[28] J. Swales, *Genre analysis: English in academic and research settings.* Cambridge University Press, 1990.

[29] K. Hyland, "Academic attribution: citation and the construction of disciplinary knowledge," *Applied Linguistics*, vol. 20, no. 3, pp. 341–367, 1999.

[30] P. Thompson, *A pedagogically-motivated corpus-based examination of PhD theses: Macrostructure, citation practices and uses of modal verbs.* PhD thesis, University of Reading, 2001.

[31] A. Okamura, "Citation forms in scientific texts: Similarities and differences in L1 and L2 professional writing," *Nordic Journal of English Studies*, vol. 7, no. 3, pp. 61–81, 2008.

[32] W. Lamers, N. J. v. Eck, L. Waltman, and H. Hoos, "Patterns in citation context: the case of the field of scientometrics," in *STI 2018 Conference proceedings*, pp. 1114–1122, Centre for Science and Technology Studies (CWTS), 2018.

[33] M. Whidby, D. Zajic, and B. Dorr, "Citation handling for improved summarization of scientific documents," tech. rep., 2011.

[34] A. Abu-Jbara and D. Radev, "Reference Scope Identification in Citing Sentences," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, (Stroudsburg, PA, USA), pp. 80–90, Association for Computational Linguistics, 2012.

[35] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticae Investigationes*, vol. 30, pp. 3–26, Jan. 2007.

[36] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: introduction and challenges," in *Recommender Systems Handbook*, pp. 1–34, Springer, 2015.

[37] A. Elkiss, S. Shen, A. Fader, G. Erkan, D. States, and D. Radev, "Blind men and elephants: What do citation summaries tell us about a research article?," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 1, pp. 51–62, 2008.

[38] C. C. Aggarwal, "Evaluating Recommender Systems," in *Recommender Systems: The Textbook*, pp. 225–254, Springer, 2016.

[39] D. Roy, K. Ray, and M. Mitra, "From a Scholarly Big Dataset to a Test Collection for Bibliographic Citation Recommendation," in *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[40] M. Färber, A. Thiemann, and A. Jatowt, "A High-Quality Gold Standard for Citation-based Tasks," in *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC, 2018.

[41] K. Sugiyama and M.-Y. Kan, "Exploiting Potential Citation Papers in Scholarly Paper Recommendation," in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, (New York, NY, USA), pp. 153–162, ACM, 2013.

[42] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, "The ACL anthology network corpus," *Language Resources and Evaluation*, vol. 47, pp. 919–944, Dec. 2013.

[43] S. Bird, R. Dale, B. J. Dorr, B. R. Gibson, M. T. Joseph, M. Kan, D. Lee, B. Powley, D. R. Radev, and Y. F. Tan, "The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics," in *Proceedings of the International Conference on Language Resources and Evaluation*, LREC, 2008.

[44] B. Gipp, N. Meuschke, and M. Lipinski, "CITREC : An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central," in *iConference 2015 Proceedings*, iSchools, 2015.

[45] L. Galke, F. Mai, I. Vagliano, and A. Scherp, "Multi-Modal Adversarial Autoencoders for Recommendations of Citations and Subject Labels," in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, (New York, NY, USA), pp. 197–205, ACM, 2018.

[46] C. Bhagavatula, S. Feldman, R. Power, and W. Ammar, "Content-Based Citation Recommendation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 238–251, Association for Computational Linguistics, June 2018.

[47] M. Bertin, I. Atanassova, C. R. Sugimoto, and V. Lariviere, "The linguistic patterns and rhetorical structure of citation context: an approach using n-grams," *Scientometrics*, vol. 109, pp. 1417–1434, Dec. 2016.

[48] M. Bertin and I. Atanassova, "InTeReC: In-text Reference Corpus for Applying Natural Language Processing to Bibliometrics," in *7th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2018) to be held as part of the 40th European Conference on Information Retrieval (ECIR)*, (Grenoble, France), Mar. 2018.

[49] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang, "An Overview of Microsoft Academic Service (MAS) and Applications," in *Proceedings of the 24th*

International Conference on World Wide Web, WWW '15 Companion, (New York, NY, USA), pp. 243–246, ACM, 2015.

[50] P. Ginsparg, "First Steps Towards Electronic Research Communication," *Computers in Physics*, vol. 8, pp. 390–396, July 1994.

[51] H. Bast and C. Korzen, "A Benchmark and Evaluation for Text Extraction from PDF," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 1–10, June 2017.

[52] H. Nanba, "Towards Multi-paper Summarization Using Reference Information," Master's thesis, Japan Advanced Institute of Science and Technology, Feb. 1998. (in Japanese).

[53] A. Prasad, M. Kaur, and M.-Y. Kan, "Neural ParsCit: A Deep Learning Based Reference String Parser," *International Journal on Digital Libraries*, vol. 19, pp. 323–337, 2018.

[54] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval Estimation for a Binomial Proportion," *Statistical Science*, vol. 16, no. 2, pp. 101–133, 2001.

[55] M. Hernández-Alvarez and J. M. Gomez, "Survey about citation context analysis: Tasks, techniques, and resources," *Natural Language Engineering*, vol. 22, no. 3, pp. 327–349, 2016.

[56] M. J. Moravcsik and P. Murugesan, "Some Results on the Function and Quality of Citations," *Social Studies of Science*, vol. 5, no. 1, pp. 86–92, 1975.

[57] T. Berners-Lee, J. Hendler, O. Lassila, *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[58] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "A Survey on Open Information Extraction," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3866–3878, Association for Computational Linguistics, 2018.

[59] B. Petrić, "Rhetorical functions of citations in high- and low-rated master's theses," *Journal of English for Academic Purposes*, vol. 6, no. 3, pp. 238 – 253, 2007.

[60] S. Zhang, R. Rudinger, and B. V. Durme, "An Evaluation of PredPatt and Open IE via Stage 1 Semantic Role Labeling," in *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*, 2017.

[61] A. S. White, D. Reisinger, K. Sakaguchi, T. Vieira, S. Zhang, R. Rudinger, K. Rawlins, and B. Van Durme, "Universal Decompositional Semantics on Universal Dependencies," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1713–1723, Association for Computational Linguistics, 2016.

[62] Mausam, "Open information extraction systems and downstream applications," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 4074–4077, AAAI Press, 2016.

[63] L. Del Corro and R. Gemulla, "ClausIE: clause-based open information extraction," in *Proceedings of the 22nd international conference on World Wide Web*, pp. 355–366, ACM, 2013.

[64] Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, "Open Language Learning for Information Extraction," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, (Stroudsburg, PA, USA), pp. 523–534, Association for Computational Linguistics, 2012.

[65] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman, "Universal dependencies v1: A multilingual treebank collection," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), 2016.

[66] J. Swales, “Citation Analysis and Discourse Analysis,” *Applied Linguistics*, vol. 7, no. 1, pp. 39–56, 1986.

[67] J. He and C. Chen, “Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications,” *Frontiers in Research Metrics and Analytics*, vol. 3, 2018.

[68] X. Wu, X. Collilieux, Z. Altamimi, B. L. A. Vermeersen, R. S. Gross, and I. Fukumori, “Accuracy of the international terrestrial reference frame origin and earth expansion,” *Geophysical Research Letters*, vol. 38, no. 13, 2011.

[69] W. G. Potter, “Lotka’s Law Revisited,” *Library Trends*, vol. 30, no. 1, pp. 21–39, 1981.

[70] T. Saier and M. Färber, “Bibliometric-Enhanced arXiv: A Data Set for Paper-Based and Citation-Based Tasks,” in *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 41st European Conference on Information Retrieval (ECIR 2019)*, pp. 14–26, Apr. 2019.

[71] A. Aizawa, M. Kohlhase, I. Ounis, and M. Schubotz, “NTCIR-11 Math-2 Task Overview,” in *Proceedings of the 11th NTCIR Conference*, vol. 11, pp. 88–98, Dec. 2014.

[72] R. Zanibbi, A. Aizawa, M. Kohlhase, I. Ounis, G. Topic, and K. Davila, “NTCIR-12 MathIR Task Overview,” in *Proceedings of the 12th NTCIR Conference*, vol. 12, June 2016.

[73] K. Gábor, D. Buscaldi, A.-K. Schumann, B. QasemiZadeh, H. Zargayouna, and T. Charnois, “SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers,” in *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 679–688, Association for Computational Linguistics, 2018.

[74] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, “Credibility Assessment of Textual Claims on the Web,” in *Proceedings of the 25th ACM International on Conference on*

Information and Knowledge Management, CIKM '16, (New York, NY, USA), pp. 2173–2178, ACM, 2016.

[75] C. Stab and I. Gurevych, “Parsing Argumentation Structures in Persuasive Essays,” *Computational Linguistics*, vol. 43, pp. 619–659, Sept. 2017.

[76] M. Lippi and P. Torroni, “Argumentation Mining: State of the Art and Emerging Trends,” *ACM Transactions on Internet Technology*, vol. 16, pp. 10:1–10:25, Mar. 2016.

[77] I. Habernal and I. Gurevych, “Argumentation Mining in User-generated Web Discourse,” *Computational Linguistics*, vol. 43, pp. 125–179, Apr. 2017.

[78] M. Färber, A. Thiemann, and A. Jatowt, “To Cite, or Not to Cite? Detecting Citation Contexts in Text,” in *Proceedings of the 40th European Conference on Information Retrieval*, ECIR, 2018.

Appendices

A Evaluation of reference resolution

The sample of 300 matched reference items was acquired from the reference data base as shown in Listing A.1. The table `bibitem` holds most of the information on reference items. The table `bibitemmagidmap` contains per row the UUID of a reference item and the MAG ID it was matched to.

Listing A.1: SQL query used to acquire the sample

```
select b.bibitem_string, m.mag_id
    from bibitem as b
        join
         (select *
            from bibitemmagidmap
                order by random()
                limit 300
         ) as m
        on b.uuid = m.uuid;
```

Table 10 shows the full evaluation. Wrongly matched items are highlited in red. Note that there are seven cases where the reference item refers to more than one publication. In such cases our method only captures the first one and is consequentially evaluated on the first one. Note further that in one case a reference item named a PhD thesis while the match was for the very same thesis published two years later. This was deemed a correct match. Lastly, there was one case where a book was cited with the date indicating its second edition while the matched record in the MAG has a date indicating the books third edition. This also was deemed a correct match.

Table 10: Evaluation

Reference item	MAG ID	
V. N. Senoguz and Q. Shafi, "Reheat temperature in supersymmetric hybrid inflation models," Phys. Rev. D 71, 043514 (2005) [hep-ph/0412102].	2075392245	✓
Keiding, N. and Nielsen, J.E. (1975) Branching processes with varying and random geometric offspring distributions. J. Appl. Prob. 12, 135–141.	2332540167	✓
H. Izeki and S. Nayatani, Combinatorial harmonic maps and discrete-group actions on Hadamard spaces, Geom. Dedicata 114 (2005), 147–188.	2017711173	✓
T. Adamo, M. Bullimore, L. Mason and D. Skinner, "Scattering Amplitudes and Wilson Loops in Twistor Space," J. Phys. A 44 (2011) 454008 [arXiv:1104.2890 [hep-th]].	2002091616	✓
Eren Mehmet Kıral and Matthew Young, The fifth moment of modular FORMULA -functions, arXiv preprint arXiv:1701.07507 (2017).	2582886839	✓
T. Baumgarte and S. Shapiro, Numerical Relativity: Solving Einstein's Equations on the Computer. Cambridge University Press, 2010. http://books.google.co.uk/books?id=dxU1OEinvRUC.	2566410267	✓
R.E. Renfordt, D. Schall, R. Bock, R. Brockmann, J.W. Harris, A. Sandoval, R. Stock, H. Ströbele, D. Bangert, W. Rauch, G. Odyniec, H.G. Pugh, and L.S. Schroeder, Phys. Rev. Lett. 53 (1984) 763.	2001418221	✓
T. A. Porter, I. V. Moskalenko, A. W. Strong, E. Orlando and L. Bouchet, arXiv:0804.1774 [astro-ph].	2129746122	✓
Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807, 2016.	2522104760	✓
R.M. Fernandes, L. H. VanBebber, S. Bhattacharya, P. Chandra, V. Keppens, D. Mandrus, M.A. McGuire, B.C. Sales, A.S. Sefat, and J. Schmalian, Phys. Rev. Lett. 105, 157003 (2010)	2143202785	✓
Fomin, S., Wei, P., Chugunov, V., 1995. Contact melting by a non-isothermal heating surface of arbitrary shape. Int. J. Heat Mass Transfer 38 (17), 3275–3284.	2030656707	✓
R.F. Lebed, arXiv:1507.05867v1 [hep-ph].	1844403609	✓
R. Billinton, R. Karki, Y. Gao, D. Huang, P. Hu, and W. Wangdee, "Adequacy Assessment Considerations in Wind Integrated Power Systems," IEEE Trans. Power Syst., vol. 27, no. 4, pp. 2297–2305, 2012.	2024825567	✓
C. Morningstar and M. Peardon, Phys. Rev. D 56, 4043 (1997).	2109255696	✓
I. B. S. Passi, M. Singh and M. K. Yadav, Automorphisms of abelian group extensions, J. Algebra 324 (2010), 820–830.	2051680489	✓
S. Dimopoulos, P. W. Graham, J. M. Hogan, M. A. Kasevich, and S. Rajendran, Phys. Rev. D 78, 122002 (2008).	2749889157	✓
P.Jaworski, Value at Risk in the presence of the power laws, Acta Physica Polonica B 36 (2005) 2575-2587.	2566822874	✓
L. L. Chau, M. L. Ge and Y. S. Wu, Phys. Rev. D 25, 1086 (1982); L. L. Chau and Wu Yong-Shi, Phys. Rev. D 26, 3581 (1982); L. L. Chau, M. L. Ge, A. Sinha and Y. S. Wu, Phys. Lett. B 121, 391 (1983).	2075757391	✓
E. C. Blomberg, M. a. Tanatar, R. M. Fernandes, I. I. Mazin, B. Shen, H.-H. Wen, M. D. Johannes, J. Schmalian, and R. Prozorov, Nat. Commun. 4, 1914 (2013).	2081425933	✓
A. Arenas, A. Díaz-Guilera, C. J. Pérez-Vicente, Synchronization reveals topological scales in complex networks, Phys. Rev. Lett. 96 (11) (2006) 114102.	2100240966	✓
M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan. Data Center TCP (DCTCP). ACM SIGCOMM, 2010.	2164740236	✓
M. Nishiyama, T. Okabe, Y. Sato, and I. Sato. Sensation-based photo cropping. In ACM Multimedia, pages 669–672, 2009.	2013339738	✓
L. Landau and E. Lifchitz, Classical theory of fields, Butterworth-Heinemann, Oxford, 1994, p. 87.	119088996	✓
Peregrine, D., 2003. Water-wave impact on walls. Annu. Rev. Fluid Mech. 35, 23–43.	2118259172	✓
M. S. Khalil, S. Gladchenko, M. J. A. Stoutimore, F. C. Wellstood, A. L. Burin, and K. D. Osborn, Phys. Rev. B 90, 100201 (2014).	1998393159	✓
F. Gabbiani, E. Gabrielli, A. Masiero and L. Silvestrini, Nucl. Phys. B 477, 321 (1996) [arXiv:hep-ph/9604387].	2133327165	✓
Abramowitz, M. and Stegun, I. A., Handbook of Mathematical Functions, (Dover, New York, 1965).	2120062331	✓
N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16(1):321–357, 2002.	2148143831	✓
B. H. Lee, W. Lee, R. MacKenzie, M. B. Paranjape, U. A. Yajnik and D. h. Yeom, Phys. Rev. D 88, 085031 (2013).	2074315988	✓
D. Nishiguchi, K. H. Nagai, H. Chaté, and M. Sano, Long-range nematic order and anomalous fluctuations in suspensions of swimming filamentous bacteria. arXiv preprint arXiv:1604.04247 (2016).	2336881770	✓
Huelga, S. F. et al., Improvement of frequency standards with quantum entanglement. Phys. Rev. Lett. 79, 3865 (1997).	2015876000	✓
M. Henneaux and C. Teitelboim, Asymptotically anti-de Sitter spaces, Comm. Math. Phys. 98, 391 (1985).	2134856726	✓
D. Boucher, W. Geiselmann, and F. Ulmer, Skew-cyclic codes, Applicable Algebra in Engineering, Communication and computing, (18) (4), (2007), 379-389.	2151359594	✓
R. C. Brower, H. Nastase, H. J. Schnitzer and C.-I. Tan, arXiv:0801.3891 [hep-th].	1995185450	✓

S. Sarkar, Big bang nucleosynthesis and physics beyond the standard model, Rept. Prog. Phys. 59 (1996) 1493–1610, [hep-ph/9602260].	2013442457	✓
M. Iskin and J. K. Freericks, Phys. Rev. A 80, 053623 (2009).	2076979008	✓
M. L. Skoge & T. W. Baumgarte, Phys. Rev. D 66 107501 (2002).	2016175541	✓
A. Lozano, A. M. Tulino, and S. Verdú, “Optimum power allocation for parallel Gaussian channels with arbitrary input distributions,” IEEE Trans. Inf. Theory, vol. 52, no. 7, pp. 3033–3051, Jul. 2006.	2097695636	✓
F. Paci, A. Gruppuso, F. Finelli, A. De Rosa, N. Mandolesi, and P. Natoli, MNRAS 434, 3071 (Oct. 2013), arXiv:1301.5195	1976439668	✓
Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 804–812. Association for Computational Linguistics.	2113786470	✓
R. Islam, C. Senko, W. C. Campbell, S. Korenblit, J. Smith, A. Lee, E. E. Edwards, C.-C. J. Wang, J. K. Freericks, and C. Monroe, Emergence and frustration of magnetism with variable-range interactions in a quantum simulator, Science 340, 583 (2013).	1606923443	✓
I. Ermolli, K. Matthes, T. Dudok de Wit, N.A. Krivova, K. Tourpali, M. Weber, Y.C. Unruh, L. Gray, U. Langematz, P. Pilewskie, E. Rozanov, W. Schmutz, A. Shapiro, S.K. Solanki, T.N. Woods, Recent variability of the solar spectral irradiance and its impact on climate modelling. Atmos. Chem. Phys. 13, 3945–3977 (2013). doi:10.5194/acp-13-3945-2013	2158950390	✓
Tauchen G. and Zhou, H. (2011), “Realized Jumps on Financial Markets and Predicting Credit Spreads,” Journal of Econometrics, 160, 102–118	2113380547	✓
S. Yoon, W. Ye, J. Heidemann, B. Littlefield, and C. Shahabi. Swats: Wireless sensor networks for steamflood and waterflood pipeline monitoring. Network, IEEE, 25(1):50–56, January 2011.	2054094122	✓
J.Y. Vaishnav and C.W. Clark, Phys. Rev. Lett. 100, 153002 (2008).	2752272568	✓
L. J. Hall, T. Moroi and H. Murayama, Phys. Lett. B 424, 305 (1998). [arXiv:hep-ph/9712515].	2078593377	✓
Movassaghi, Samaneh and Abolhasan, Mehran and Smith, David, Smart spectrum allocation for interference mitigation in Wireless Body Area Networks, IEEE International Conference on Communications (ICC), pages 5688-5693, 2014	2083602816	✓
M. Apollonio et al. [CHOOZ Collaboration], Phys. Lett. B 466, 415 (1999) [arXiv:hep-ex/9907037].	1571701324	✓
G. ’t Hooft, “Magnetic Monopoles in Unified Gauge Theories,” Nucl.Phys. B79 (1974) 276–284.	2027710569	✓
D. C. Cabra, M. D. Grynberg, P. C. W. Holdsworth, A. Honecker, P. Pujol, J. Richter, D. Schmalfß, and J. Schulenburg, Phys. Rev. B 71, 144420 (2005).	2032654705	✓
V.A. Dolgushev, Erratum to: ""A Proof of Tsygan’s Formality Conjecture for an Arbitrary Smooth Manifold"", arXiv:math/0703113.	132267651	✓
Eddy, J.A.: 1983, The maunder minimum - a reappraisal. Solar Phys. 89, 195. ADS.	2024069573	×
Bouliotis, G. and Billingham, L. (2011) Crossing survival curves: alternatives to the log-rank test. Trials, 12, 1.	2136058878	✓
I. Peschel and V. Eisler, Reduced density matrices and entanglement entropy in free lattice models, J. Phys. A 42 504003 (2009).	2141130841	✓
C. Gundlach, Critical phenomena in gravitational collapse, submitted to Adv. Theor. Math. Phys., preprint gr-qc/9712084.	2103342599	✓
A. Blumen, G. Zumofen, and J. Klafter, Target annihilation by random walkers, Phys. Rev. B 30, 5379 (1984).	2060897997	✓
M. Alidoust, K. Halterman, and J. Linder, Phys. Rev. B 89, 054508 (2014).	2012436162	✓
G. Ciavola, L. Celona, S. Gammino, M. Presti, L. Ando, S. Passarello, X.Zh. Zhang, F. Consoli, F. Chines, C. Percolla, V. Calzona and M. Winkler, A version of the Trasco Intense Proton Source optimized for accelerator driven system purposes. Rev. Sci. Instrum. 75 (2004) 1453–1456,.	2075428341	✓
D.Litim Phys. Rev. Lett. 92 (2004) 201301, hep-th/0312114	2000661521	✓
P.K. Kovtun, D.T. Son, and A.O. Starinets, Phys. Rev. Lett. 94, 111601 (2005).	2097909025	✓
Jones, J. A. et al. Magnetic Field Sensing Beyond the Standard Quantum Limit Using 10-Spin NOON States. Science 324, 1166–1168 (2009).	2022149792	✓
G.E.Brown and M.Rho, Phys.Rev.Lett. 66, (1991) 2720;	1971702030	✓
K. Banaszek and K. Wódkiewicz, “Operational theory of homodyne detection,” Phys. Rev. A 55, 3117 (1997).	1985850877	✓
Barbara Di Eugenio, Pamela W. Jordan, and Liina Pylkkänen. 1998. The COCONUT project: Dialogue annotation manual. Technical Report 98-1, University of Pittsburgh, Intelligent Systems Program. [www.isp.pitt.edu/intgen/coconut.html].	116082719	✓
Beirlant, J., Y. Goegebeur, J. Segers, and J. Teugels (2004). Statistics of extremes: Theory and Applications. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd.	1598342322	✓
Guillaumin, M., Mensink, T., Verbeek, J.J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009. (2009) 309–316	2536305071	✓
T. Faulkner et al., “Gravitation from entanglement in holographic CFTs,” (2013), arXiv:1312.7856v1.	2102970467	✓
René Thom, Quelques propriétés globales des variétés différentiables, Comment. Math. Helv. 28 (1954), 17–86.	1989427081	✓
A. Schikorra. Regularity of n/2-harmonic maps into spheres. PhD-Thesis, arXiv:1003.0646v1, 2010.	2029831933	✓
M. J. Simpson, J. A. Sharp, and R. E. Baker. Survival probability for a diffusive process on a growing domain. Phys. Rev. E 91, 042701 (2015).	1988447698	✓

D. Bergamini, N. Descoubes, C. Joubert & R. Mateescu (2005): Bisimulator: A Modular Tool for On-the-Fly Equivalence Checking. In: Proc. of TACAS'05, Lecture Notes in Computer Science 3440, Springer-Verlag, pp. 581–585.	1503429725	✓
J. A. Baldwin, O. Plamenevskaya, Khovanov homology, open books, and tight contact structures. math.GT/0808.2336	2054234174	✓
Eisenberger, P., et al., 1972, Phys. Rev. B 6, 3671.	1980522055	✓
O. Viehmann, C. Eltschka, and J. Siewert, Appl. Phys. B 106, 533 (2012).	2076107066	✓
A. Rényi. Representations for real numbers and their ergodic properties. Acta Math. Acad. Sci. Hungar 8 (1957), 477–493.	2089164015	✓
D. Chicharro and R. G. Andrzejak, Phys. Rev. E 80, 026217 (2009).	2078979753	✓
Kenward, M. G., Jones, B., 1992. Alternative approaches to the analysis of binary and categorical repeated measurements. Journal of Biopharmaceutical Statistics 2 (2), 137–170.	1967449535	✓
J. Nešetřil and V. Rödl. The partite construction and Ramsey set systems. Discrete Mathematics, 75(1-3):327–334, 1989.	2062861878	✓
E. Seiler, Gauge Theories as a Problem of Constructive Quantum Field Theory and Statistical Mechanics, Lecture Notes in Physics Vol. 159 (Springer, Berlin, 1982).	1608098855	✓
M. Gaye, Y. Chitour, and P. Mason. Properties of barabanov norms and extremal trajectories associated with continuous-time linear switched systems. In Proceedings of the 52nd IEEE Conference on Decision and Control, pages 716–721, Florence, Italie, 2013.	2090302562	✓
Kováčik R. and Ederer C., Phys. Rev. B 80 (2009) 140411; Kim M. et al., Phys. Rev. B 81 (2010) 100409.	1758648405	✓
Morandi, G., Ferrario, C., Lo Vecchio, G., Marmo, G. and Rubano, C. (1990). The inverse problem in the calculus of variations and the geometry of the tangent bundle, Phys. Rep. 188, 147.	2026992500	✓
G. Da Prato and J. Zabczyk, Ergodicity for infinite dimensional systems, London Mathematical Society Lecture Note Series, 229, Cambridge University Press, 1996.	1530927473	✓
R. Trotta, Bayes in the sky: Bayesian inference and model selection in cosmology, Contemp. Phys. 49 (2008) 71–104, [arXiv:0803.4089].	2021748112	✓
Christof, J., M. Gebhardt, A. E.-M. Clemen, J. Jaud, , and M. Rief, 2006. Myosin-V is a mechanical ratchet. Proc. Natl. Acad. Sci. U.S.A. 103:8680–8685.	2095633207	✓
G.M. Molchan: Burgers equation with self-similar Gaussian initial data: tail probabilities. J. of Stat. Phys. 88 (1997) 1139–1150.	2000957076	✓
Arata, I., Y. Ohno, F. Matsukura, and H. Ohno, 2001, "Temperature dependence of electroluminescence and I-V characteristics of ferromagnetic/non-magnetic semiconductor pn junctions," Physica E 10, 288–291.	1987870048	✓
I. Zlatev, L. Wang and P.J. Steinhardt, Phys. Rev. Lett. 82, 896 (1999); Phys. Rev. D 59, 123504 (1999).	2032901690	✓
A. Valentini, in: Bohmian Mechanics and Quantum Theory: an Appraisal, eds. J. T. Cushing et al. (Kluwer, Dordrecht, 1996).	207861407	✓
R. Killip, S. Kwon, S. Shao, and M. Visan. On the mass-critical generalized KdV equation. Discrete Contin. Dyn. Syst., 32(1):191–221, 2012.	1975802612	✓
Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770, 2015.	299440670	✓
D. Rossi and G. Rossini, "On sizing CCN content stores by exploiting topological information," in Proc. IEEE NOMEN, 2012.	1985355206	✓
F. Bezrukov, A. Magnin, M. Shaposhnikov and S. Sibiryakov, JHEP 1101 (2011) 016 [arXiv:1008.5157].	2064410211	✓
M. Horodecki, P. Horodecki, and R. Horodecki. Separability of mixed quantum states: linear contractions approach. preprint archiv quant-ph/0206008.	1668368460	✓
S. White, Phys. Rev. B 48, 10345 (1993).	2016407890	✓
A. Brandt et al., [UA8 Collaboration], Evidence for a Super-Hard Pomeron Structure, submitted to Phys. Lett. 1992.	1983143801	✓
Fan, T.-H. & Berger, J. O. (2000). Robust Bayesian displays for standard inferences concerning a normal mean. Computational Statistics & Data Analysis 33 381–399.	2030654698	✓
S. Ryu, J. E. Moore, and A. W. W. Ludwig, Phys. Rev. B 85, 045104 (2012), arXiv:1010.0936.	2083123179	✓
P. Balaz, V. K. Dugaev, and J. Barnaś Phys. Rev. B 85, 024416 (2012)	2322343165	✓
T. Giamarchi and A. Tsvelik, Phys. Rev. B 59, 11398 (1999).	2069018114	✓
M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. Combin. Probab. Comput., 7(3):295–305, (1998).	2129918926	✓
D. W. Sivers, Phys. Rev. D 41, 83 (1990); Phys. Rev. D 43, 261 (1991).	2174029682	✓
J. Boronat and J. Casulleras, Phys. Rev. B 49, 8920 (1994).	1982967539	✓
J. N. Bandyopadhyay and A. Lakshminarayan, Phys. Rev. E 69, 016201 (2004).	1977839735	✓
Priest ER, Forbes TG (2002) The magnetic nature of solar flares. 10:313–377, DOI 10.1007/s001590100013	2014209718	✓
W. Woerndl, C. Schueller, and R. Wojtech. A hybrid recommender system for context-aware recommendations of mobile applications. In Proceedings of ICDEW '07, pages 871–878, Washington, DC, USA, 2007. IEEE Computer Society.	2112166834	✓
[auto:STB|2013/06/05|13:45:01] Worsley, K. J.K. J., Liao, C. H.C. H., Aston, J.J., Petre, V.V., Duncan, G. H.G. H., Morales, F.F. Evans, A. C.A. C. (2002). A general statistical analysis for fMRI data. NeuroImage 15 1–15. imsref	1975938737	✓
L.H. Ford and N.F. Svaiter, Phys. Rev. D 58, 065007 (1998), quant-ph/9804056.	1963985219	✓
H. Häffner, S. Gulde, M. Riebe, G. Lancaster, C. Becher, J. Eschner, F. Schmidt-Kaler, R. Blatt, Precision measurement and compensation of optical Stark shifts for an ion-trap quantum processor, Phys. Rev. Lett. 90 (2003) 143602.	2129198554	✓

I. Jeon, K. Lee, J.-H. Park, and Y. Suh, Stringy Unification of Type IIA and IIB Supergravities under N=2 D=10 Supersymmetric Double Field Theory, Phys.Lett. B723 (2013) 245–250, [arXiv:1210.5078].	1967998606	✓
Z.A. Anastassi and T.E. Simos: A Trigonometrically-Fitted Runge-Kutta Method for the Numerical Solution of Orbital Problems, New Astronomy, 10, 301-309 (2005)	2024993485	✓
Fletcher, A. 2010, in Astronomical Society of the Pacific Conference Series, Vol. 438, The Dynamic Interstellar Medium: A Celebration of the Canadian Galactic Plane Survey, ed. R. Kothes, T. L. Landecker, & A. G. Willis, 197	1671679100	✓
V. A. Belinsky, I. M. Khalatnikov, and E. M. Lifshitz. Oscillatory approach to a singular point in the relativistic cosmology. Adv. Phys., 19:525–573, 1970.	2048737175	✓
Allen, D. A. et al., 1993. IRIS – an Infrared Imager and Spectrometer for the Anglo-Australian Telescope. Proceedings of the Astronomical Society of Australia 10, 298.	91162570	✓
D. Cooper, Automorphisms of free groups have finitely generated fixed point sets. J. Algebra, 111 (1987), no. 2 453–456	2076181847	✓
J.-L. Lehners and P. J. Steinhardt, “Intuitive understanding of non-gaussianity in ekpyrotic and cyclic models,” Phys.Rev. D78 (2008) 023506, arXiv:0804.1293 [hep-th].	2592352904	✓
Z.-B. Wu, Global transposable characteristics in the complete DNA sequence of the yeast. Physica A 389 (2010) 5698.	1548592618	✓
E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In Computer Vision–ECCV 2006, pages 490–503. Springer, 2006.	2171896402	✓
Benjamin C. Pierce. Types and programming languages: The next generation. LICS’03, 2003.	1951034176	✓
G. Tardos and G. Tóth. Multiple coverings of the plane with triangles. Discrete & Computational Geometry, 38(2):443–450, 2007.	2043718124	✓
T. Rivière, Analysis aspects of Willmore surfaces, Invent. Math., Vol. 174, (2008), 1–45.	1606077524	✓
H. Mabuchi, Phys. Rev. A 85, 015806 (2012).	1620088716	✓
I. Schienbein, J. Y. Yu, K. Kovarik, C. Keppel, J. G. Morfin, F. Olness and J. F. Owens, Phys. Rev. D 80, 094004 (2009)	2050053974	✓
D.Q. Goldin, S.A. Smolka, P.C. Attie, E.L. Sonderegger, Turing machines, transition systems and interaction, manuscript, 2003.	2048671682	✓
P. Collet and J.P. Eckmann, Iterated Maps on the Interval as Dynamical Systems, (Birkhäuser, Basel, 1980).	1573241742	✓
M. F. Maghrebi, R. L. Jaffe, and M. Kardar, Phys. Rev. Lett. 108, 230403 (2012).	2025571896	✓
A. Minami and A. Onuki, Phys. Rev. B 70, 184114 (2004); Acta Mater. 55, 2375 (2007).	1514590689	✓
Kfir Blum, Anson Hook, and Kohta Murase, “High energy neutrino telescopes as a probe of the neutrino mass mechanism,” (2014), arXiv:1408.3799 [hep-ph] .	1628049864	✓
O. E. Buryak, Phys. Rev. D 53 (1996) 1763 [gr-qc/9502032].	2069968655	✓
Kolb, E. W.; Turner, M. S. The Early Universe, AddisonWesley Publishing Company: California, USA, 1990.	2595419339	✓
G. Binasch, P. Grünberg, F. Saurenbach, and W. Zinn, Phys. Rev. B 39, 4828 (1989).	2043072234	✓
JC Baygents and DA Saville. The circulation produced in a drop by an electric field: a high field strength electrokinetic model. In AIP Conference Proceedings, volume 197, pages 7–17. AIP, 1990.	1612371284	✓
E. Altman and R. Vosk, Annual Review of Condensed Matter Physics 6, 383 (2015).	2166587989	✓
J.-M. Souriau, Structure des systèmes dynamiques (Dunod, 1970).	108534386	✓
F. Horn. Explicit Muller games are PTIME. In Proc. 28th Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS’08), LIPIcs 2, p. 235–243. Leibniz-Zentrum für Informatik, 2008.	2240543079	✓
A. G. Izergin and V. E. Korepin, “The Inverse Scattering Method Approach To The Quantum Shabat-Mikhailov Model,” Commun. Math. Phys. 79 (1981) 303.	1977168974	✓
Balzano, V. A. 1983 Star-burst Galactic Nuclei. ApJ 268, 602–627.	2059760395	✓
M. Ishak, A. Upadhye, D. N. Spergel, Phys. Rev. D 74, 043513 (2006) [arXiv:astro-ph/0507184].	2044422425	✓
M. Lynker and R. Schimmrigk, Landau–Ginzburg theories as orbifolds, Phys. Lett. B249 (1990) 237	2030907161	✓
S. T. Petcov, T. Shindou and Y. Takanishi, Nucl. Phys. B 738, 219 (2006) [arXiv:hep-ph/0508243].	2007551251	✓
J. E. Lye, L. Fallani, C. Fort, V. Guarrera, M. Modugno, D. S. Wiersma, and M. Inguscio, Phys. Rev. A 75, 061603(R) (2007).	2015993116	✓
J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. 77, 3865 (1996).	1981368803	✓
J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, Physical Review Letters 23, 880 (1969), URL http://doi.org/10.1103/PhysRevLett.23.880.	2028815089	✓
M. Horodecki and P. Horodecki, Phys. Rev. A 59, 4206 (1999).	2000407553	✓
B. Mazur, Rational isogenies of prime degree. Invent. Math. 44 (1978), 129–162.	2011844852	✓
Arnold, B.C, Balakrishnan, N., Nagaraja H.N., (1992), A First course in order statistics, Wiley and sons.	2318245334	✓
K. Bamba and S. D. Odintsov, JCAP 0804, 024 (2008) [arXiv:0801.0954 [astro-ph]].	2065428968	✓
Mujherjee N. P. and Bhattacharya, P., Fuzzy Groups Some Group-Theoretic Analogs, Information Science39,247-268 (1986).	1999360086	✓
Karen Suzanne Oberhauser, M. J. S. The monarch butterfly: biology and conservation. Cornell university press, 2004.	570971783	✓
A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. Advances in Economics and Econometrics: The 2010 World Congress of the Econometric Society, 3:245–295, 2013.	1720842782	✓
M. I. Aroyo, A. Kirov, C. Capillas, J. M. Perez-Mato, and H. Wondratschek, Acta Cryst. A 62, 115 (2006b).	1996820002	✓

V. Del Duca and C. R. Schmidt, Dijet Production At Large Rapidity Intervals, 4919944510 [9311290].	2003427747	✓
G. F. Giudice and A. Strumia, Nucl. Phys. B 858 (2012) 63 [arXiv:1108.6077 [hep-ph]].	2015208992	✓
C. D. Herrera, J. Kannala, P. Sturm, and J. Heikkila. A learned joint depth and intensity prior using markov random fields. In 3DTV-Conference, 2013 International Conference on, pages 17–24. IEEE, 2013.	1994295411	✓
J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. In Advances in Neural Information Processing Systems (NIPS), volume 16, pages 49–56, 2004.	2249237221	×
M. Beynon, B. Curry, and P. Morgan, ""The Dempster-Shafer theory of evidence:an alternative approach to multicriteria decision modelling"", Omega, vol. 28, no. 1, pp. 37–50, 2000.	2121042048	✓
S. Flach, M. V. Ivanchenko, and O. I. Kanakov, Phys. Rev. Lett. 95, 064102/1-4 (2005).	2004622729	✓
P. Carrasco, A. M. Cegarra, and A. R. Garzon. Nerves and classifying spaces for bicategories, 2010.	2108966476	✓
R. Sandhu, S. Dambreville, A. Tannenbaum, Particle Filtering for Registration of 2D and 3D Point Sets with Stochastic Dynamics. Pro. of IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8.	2146847221	✓
S. Ji, Y. Xue and L. Carin, “Bayesian compressive sensing,” IEEE Trans. Signal Process., vol. 56, no. 6, pp. 2346–2356, 2008.	2071284784	✓
Jun Li. A degeneration formula of GW-invariants. J. Differential Geom., 60(2):199–293, 2002.	1484479264	✓
J. Milnor, Dynamics in One Complex Variable, Vieweg, Göttingen, 2000.	1603977374	✓
Marcheselli, M., Baccini, A. and Barabesi, L. (2008). Parameter estimation for the discrete stable family. Communications in Statistics - Theory and Methods 37 815–830.	2035009392	✓
A. Chantasri, J. Dressel, and A. N. Jordan, Action principle for continuous quantum measurement, Phys. Rev. A 88, 042110 (2013).	2025224750	✓
A. Strominger, Black hole entropy from near horizon microstates, JHEP 9802 (1998) 009, [hep-th/9712251].	2058588165	✓
C. I. Lazaroiu, “On the structure of open-closed topological field theory in two dimensions,” Nucl. Phys. B 603, 497 (2001) arXiv:hep-th/0010269.	2033747261	✓
Liao, L. Z., Qi, H., & Qi, L. (2004). Neurodynamical optimization. Journal of Global Optimization, 28(2), 175-195.	2340554656	✓
J. L. Lions, Quelques méthodes de résolution des problèmes aux limites non linéaires, Dunod, Paris, 1969.	1519031678	✓
M. Scully and W. E. Lamb, Jr. Quantum theory of an optical maser. Phys. Rev. Lett., 16(19):853–855, 1966.	2015518403	✓
G. Lanckriet, M. Deng, N. Cristianini, M. Jordan, W. Noble, Kernel-based data fusion and its application to protein function prediction in yeast, in: Proceedings of the Pacific Symposium on Biocomputing, Vol. 9, 2004, pp. 300–311.	2013502943	✓
Aharonson, O., Hayes, A. G., Lunine, J. I., Lorenz, R. D., Allison, M. D., Elachi, C., Dec. 2009. An asymmetric distribution of lakes on Titan as a possible consequence of orbital forcing. Nature Geoscience 2, 851–854.	2083408367	✓
K. Ito and S. S. Ravindran. A reduced-order method for simulation and control of fluid flows. Journal of Computational Physics, 143(2):403–425, 1998.	2045627558	✓
P. You, Y. Sun, J. Pang, S. Low, and M. Chen, “Battery swapping assignment for electric vehicles: A bipartite matching approach,” SIGMETRICS Performance Evaluation Review, vol. 45, no. 2, pp. 85–87, 2017.	2762188191	✓
Panjer H. (1981). Recursive evaluation of a family of compound distributions. ASTIN Bulletin 12, 22-26.	2156812602	✓
M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. In Computer Vision and Pattern Recognition, 2016.	2259801182	✓
B. G. Saar, C. W. Freudiger, J. Reichman, C. M. Stanley, G. R. Holtom, and X. S. Xie, “Video-rate molecular imaging in vivo with stimulated raman scattering,” Science 330, 1368–1370 (2010).	2096138335	✓
J. L. Feng, C. F. Kolda and N. Polonsky, Nucl. Phys. B 546, 3 (1999) [arXiv:hep-ph/9810500]; J. Bagger, J. L. Feng and N. Polonsky, Nucl. Phys. B 563, 3 (1999) [arXiv:hep-ph/9905292]; J. A. Bagger, J. L. Feng, N. Polonsky and R. J. Zhang, Phys. Lett. B 473, 264 (2000) [arXiv:hep-ph/9911255]; H. Baer, C. Balazs, M. Brhlik, P. Mercadante, X. Tata and Y. Wang, Phys. Rev. D 64, 015002 (2001) [arXiv:hep-ph/0102156].	1997759872	✓
R. Fei, V. Tran, and L. Yang, Phys. Rev. B 91, 195319 (2015).	1590844150	✓
G. Thalhammer et al., Phys. Rev. Lett. 100, 210402 (2008)	1601643666	✓
D. S. Petrov, G. V. Shlyapnikov, and J. T. M. Walraven, Phys. Rev. Lett. 87, 050404 (2001).	2072707005	✓
Z. Bern, L. J. Dixon, D. C. Dunbar and D. A. Kosower, “Fusing gauge theory tree amplitudes into loop amplitudes,” Nucl. Phys. B 435, 59 (1995) [hep-ph/9409265].	2021404114	✓
D. Telnov and S.-I. Chu, Phys. Rev. A 79, 041401(R) (2009).	2046622440	✓
S. Gupta, Phys. Rev. D 64 (2001) 034507 [hep-lat/0010011].	2616732687	✓
T. Alazard, J.M. Delort, Global solutions and asymptotic behavior for two dimensional gravity water waves, Preprint, 2013.	1585883756	✓
M. Gastpar and M. Vetterli, “On the capacity of wireless networks: the relay case,” in Proc. IEEE Infocom, June 2002.	2097463269	✓
G. Vidal, “Efficient classical simulation of slightly entangled quantum computations,” Phys. Rev. Lett. 91 (2003).	2036604884	✓
I. Frank and J. Friedman. A statistical view of some chemometrics regression tools (with discussion). Technometrics, 35:109–148, 1993.	2079775628	✓
S. Rajagopalan and V. Vazirani. Primal-dual rnc approximation algorithms for set cover and covering integer programs. SIAM Journal of Computing, 28(2):525–540, 1998.	1988837529	✓
E. N. Parker: Cosmical Magnetic Fields: Their Origin and Their Activity, (Clarendon, Oxford 1979)	1661725509	✓

Li, J., Xin, Z.: Some uniform estimates and blowup behavior of global strong solutions to the Stokes approximation equations for two-dimensional compressible flows. J. Differ. Eqs. 221(2), 275-308 (2006).	2012319434	✓
L. Visinelli, Observational Constraints on Monomial Warm Inflation, JCAP 07 (2016) 054.	2403695403	✓
T. Kimura and V. Pestun, arXiv:1608.04651	2513775828	✓
S. Benvegnù, L. Dąbrowski: Relativistic point interaction, Lett. Math. Phys. 30 (1994), 159–167.	2053324072	✓
M. Abramowitz and I. A. Stegun, Handbook of Mathematical Functions (Dover, New York, 1970).	2120062331	✓
A. Korobeinikov, P. K. Maini, A Lyapunov function and global properties for SIR and SEIR epidemiological models with nonlinear incidence, Math. Biosci. Eng. 1 (1) (2004) 57–60.	2160057076	✓
S. J. Weidenschilling and F. Marzari. Gravitational scattering as a possible origin for giant planets at small stellar distances. , 384:619–621, December 1996.	2068108425	✓
I. Carusotto, D. Gerace, H. E. Tureci, S. De Liberato, C. Ciuti, and A. Imamoglu, Fermionized Photons in an Array of Driven Dissipative Nonlinear Cavities, Phys. Rev. Lett. 103, 033601 (2009).	2100065221	✓
I.Ya. Aref'eva, P.B. Medvedev, A.P. Zubarev, "New representation for string field solves the consistency problem for open superstring field theory,"" Nuclear Physics B, Volume 341, Issue 2.	2098143658	✓
H. Baer and X. Tata, Weak scale supersymmetry: From superfields to scattering events, . Cambridge, UK: Univ. Pr. (2006) 537 p.	1575963702	✓
Ron Kimmel and Nahum Kiryati. Finding shortest paths on surfaces by fast global approximation and precise local refinement. International Journal of Pattern Recognition and Artificial Intelligence, 10(6):643–656, 1996.	2019758632	✓
W.B. Kilgore, One-Loop Single-Real-Emission Contributions to FORMULA at Next-to-Next-to-Next-to-Leading Order, Phys. Rev. D89 (2014) 073008 [arXiv:1312.1296].	2057405276	✓
Kielpinski D Phys. Rev. A. 73 063407 (2006)	2039261148	✓
Gavrilov,L.A. FORMULA Gavrilova N.S (1991) , The Biology of life span: a quantitative approach, N.Y.:Harwood Academic Publisher.	1979363640	✓
M. Scadron, Phys. Rev. D 26, 239 (1982).	2141883609	✓
Y. Aoki, Z. Fodor, S. Katz, and K. Szabo, Phys.Lett. B643, 46 (2006), arXiv:hep-lat/0609068 [hep-lat] .	2020173052	✓
M. Glück, E. Reya, M. Stratmann, and W. Vogelsang, Phys. Rev. D 53 4775 (1996).	1550005211	✓
J.-Y. Courtois, G. Grynberg, B. Lounis and P. Verkerk, Phys. Rev. Lett. 72, 3017 (1994).	2074496196	✓
U. Leonhardt, ""Measuring the quantum state of light"", Cambridge University press, Cambridge, 1997.	1996720084	✓
M. T. Glossop and K. Ingersent, Phys. Rev. Lett. 95, 067202 (2005); Phys. Rev. B 75, 104410 (2007).	2013030742	✓
A. Giveon, A. Konechny, A. Pakman, and A. Sever, Type 0 strings in a 2-d black hole, JHEP 10 (2003) 025, [hep-th/0309056].	2070010107	✓
Hubbard, P. M., 1996. "Approximating polyhedra with spheres for time-critical collision detection". ACM Transac, 15(3), pp. 179–210.	2053212688	✓
T. Xu, L. Ma and G. Sternberg, "Practical interference alignment and cancellation for MIMO underlay cognitive radio networks with multiple secondary users," IEEE GLOBECOM, pp. 1031–1036, Dec. 2013.	2073404794	✓
T. S. Han and K. Kobayashi, "Exponential-type error probabilities for multiterminal hypothesis testing," IEEE Trans. Inform. Theory, vol. 35, no. 1, pp. 2–14, January 1989.	2071851353	✓
P. Fendley, F. Lesage, and H. Saleur, Journal of Statistical Physics 85, 211 (1996).	2007477100	✓
Akimasa Miyake and Miki Wadati. Geometric strategy for the optimal quantum search. Phys. Rev. A, 64:042317, Sep 2001.	2145249001	✓
Griffiths, G.A., Estimation of landform life expectancy, Geology, 21, 403-406, 1993.	1983389789	✓
J.-P. Kahane, Random coverings and multiplicative processes, In Fractal geometry and stochastics, II (Greifswald/Koserow, 1998), Progr. Probab. 46, 125–146, Birkhäuser, 2000.	2125142036	✓
D. C. Cox, "Antenna diversity performance in mitigating the effects of portable radiotelephone orientation and multipath propagation," IEEE Trans. Commun., vol. 31, pp. 620–628, May 1983.	2120721514	✓
Doron Zeilberger, The algebra of linear partial difference operators and its applications, SIAM J. Math. Anal. 11 (1980), no. 6, 919–932.	1991175354	✓
D. T. Limmer and D. Chandler. The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water. The Journal of Chemical Physics, 135(13):134503, 2011.	2599889364	×
M. G. Santos et al., "Cosmology with a SKA HI intensity mapping survey," PoS AASKA14 (2015) 019 [arXiv:1501.03989 [astro-ph.CO]].	1596200123	✓
L. Castillejo, R.H. Dalitz and F.J. Dyson, Low's Scattering Equation for the Charged and Neutral Scalar Theories, Phys. Rev. 101 (1956) 453-458.	2028108845	✓
S. Jennewein, M. Besbes, N. J. Schilder, S. D. Jenkins, C. Sauvan, J. Ruostekoski, J.-J. Greffet, Y. R. P. Sortais, and A. Browaeys, Phys. Rev. Lett. 116, 233601 (2016).	2345644887	✓
M. Chernicoff, J. A. Garcia, A. Guijosa and J. F. Pedraza, "Holographic Lessons for Quark Dynamics," J. Phys. G 39, 054002 (2012) [arXiv:1111.0872 [hep-th]].	2106350856	✓
C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics, 2(6):1152–1174, 1974.	1967687583	✓
R.D. Astumian : Symmetry relations for trajectories of a brownian motor, Phys. Rev. E 76, 020102 (2007).	2066644069	✓
Forest, S., Micromorphic approach for gradient elasticity, viscoplasticity, and damage. Journal of Engineering Mechanics, 135, 117-131 (2009).	2157979983	✓

L. Sironi and A. Spitkovsky, Astrophys. J. 726, 75 (2011) [arXiv:1009.0024 [astro-ph.HE]].	2032048508	✓
T. L. Barklow, arXiv:hep-ph/0312268.	2142208316	✓
M. Sharir and J. Solymosi, Distinct distances from three points, to appear in Combinatorics, Probability and Computing. Also in arXiv:1308.0814.	2100718844	✓
Ferrari, A. C., et al. Raman spectrum of graphene and graphene layers. Phys. Rev. Lett. 97, 187401 (2006).	2136334331	✓
M. Grützmann, T. Strobl, General Yang-Mills type gauge theories for p-form gauge fields: From physics-based ideas to a mathematical framework OR From Bianchi identities to twisted Courant algebroids, arXiv:1407.6759.	2107310943	✓
K. Maeda, M. Natsuume and T. Okamura, Vortex lattice for a holographic superconductor, Phys. Rev. D 81 (2010) 026002.	2111816020	✓
Y. Lai and T. Tél, Transient Chaos - Complex Dynamics on Finite-Time Scales (Springer, 2011).	654916661	✓
T. Gehrmann, Nucl. Phys. B 534, 21 (1998) [arXiv:hep-ph/9710508].	2073871748	✓
J. W. Fisher and S. Montgomery, Semiprime skew group rings, J. Algebra, 52, (1978), no. 1, 241-247	1977258644	✓
CMS Collaboration, "Search for new physics in events with same-sign dileptons and b-tagged jets in pp collisions at FORMULA TeV", JHEP 08 (2012) 110, doi:tt10.1007/JHEP08(2012)110, arXiv:1205.3933.	2790162885	✓
Oded Goldreich and Madhu Sudan. Locally testable codes and pcps of almost-linear length. J. ACM, 53(4):558–655, 2006.	2022381972	✓
J.-P. Serre, Algebraic groups and class fields, vol. 117 of Graduate Texts in Mathematics, Springer-Verlag, New York, 1988. Translated from the French.	1572771201	✓
H. Cheung and E. K. Riedel, "Energy spectrum and persistent current in one-dimensional rings," Physical Review B, vol. 40, no. 14, pp. 9498–9501, Nov. 1989.	2006367954	✓
G.-B. Huang, Q.-Y. Zhu, K. Mao, C.-K. Siew, P. Saratchandran, and N. Sundararajan, "Can threshold networks be trained directly?" IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 53, no. 3, pp. 187–191, 2006.	2161055889	✓
Hauser, Oliver P, Rand, David G, Peysakhovich, Alexander, and Nowak, Martin A. Cooperating with the future. Nature, 511(7508):220–223, 2014.	2072455842	✓
P. Arnoux, C. Mauduit, I. Shiokawa, and J. Tamura. Complexity of sequences defined by billiard in the cube. Bull. Soc. Math. France, 122(1):1–12, 1994.	2089093075	✓
C. Klix, F. Ebert, F. Weysser, M. Fuchs, G. Maret, and P. Keim, Phys. Rev. Lett. 109, 178301 (2012).	2163069494	✓
M. Srednicki, "Entropy and area," Phys. Rev. Lett. 71, 666 (1993) [hep-th/9303048].	2053387157	✓
Hughes, I.G. Velocity selection in a Doppler-broadened ensemble of atoms interacting with a monochromatic laser beam, J. Mod. Opt. 2017.	2617563275	✓
A. B. Aceves, J. V. Moloney , and A. C. Newell, Phys. Rev. A. 39 (1989) 1809.	2005701305	✓
H. Yabuki, "Feynman path integrals in the young double-slit experiment," International Journal of Theoretical Physics 25, 159–174 (1986).	2028663815	✓
L. F. Santos and M. Rigol, Phys. Rev. E 81, 036206 (2010).	2000628151	✓
W. H. Kleiner, L. M. Roth, and S. H. Autler, Phys. Rev. 133, A1226 (1964).	2047723053	✓
H.G. Bock, M. Diehl, E.A. Kostina, and J.P. Schlöder. Constrained optimal feedback control of systems governed by large differential algebraic equations. In L. Biegler, O. Ghattas, M. Heikenschloss, D. Keyes, and B. Bloemen Waanders, editors, Real-Time PDE-Constrained Optimization, pages 3–22. SIAM, 2007.	2477791619	✓
D. Larson et. al., Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Power Spectra and WMAP-Derived Parameters, Astrophys. J. Suppl. 192 (2011) 16, [arXiv:1001.4635].	2105687315	✓
G. Leon, Y. Leyva and J. Socorro, Quintom phase-space: beyond the exponential potential, Phys. Lett. B732 (2014) 285–297, [1208.0061].	2015773494	✓
Jennings E., Baugh C. M., Li B., Zhao G.-B., Koyama K., 2012, ArXiv:1205.2698 [astro-ph.CO]	1836372023	✓
P. Hu and D. Ramanan. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pages 5600–5609, 2016.	2346846221	✓
Whitall MWG, Gehring GA. Path integral approach to methyl group rotation. J Phys C 1987;20:1619-1639.	2084791981	✓
S.S. Gubser and I.R. Klebanov, "Absorption by Branes and Schwinger Terms in the World Volume Theory", Phys. Lett. B413 (1997) 41, hep-th/9708005.	2064448358	✓
R. Ferraro and F. Fiorini, Phys. Lett. B702, 75 (2011); R. Ferraro and F. Fiorini, arXiv:1106.6349 [gr-qc].	2052693432	✓
Gert Almkvist and George E. Andrews, A Hardy-Ramanujan formula for restricted partitions, Journal of Number Theory 38 (1991), no. 2, 135 – 144, Dedicated to the Memory of S. Ramanujan.	2016968046	✓
D. H. Lyth and D. Wands, Phys. Lett. B 524, 5 (2002) [arXiv:hep-ph/0110002].	1979147028	✓
Rapoport A, Chammah A M (1966) The game of chicken, American Behavioral Scientist 10:10-14	2050880379	✓
M. Proebster, M. Kaschub, T. Werthmann and S. Valentin, "Context-aware resource allocation for cellular wireless networks," EURASIP Journal on Wireless Communications and Networking, DOI: 10.1186/1687-1499-2012-216, Jul 2012.	2123032963	✓
Audinot, M., Pinchinat, S., Kordy, B.: Is my attack tree correct? In: ESORICS. LNCS, Springer (2017), (to appear)	2744888375	✓
A. Kuhn, M. Hennrich, and G. Rempe, Phys. Rev. Lett. 89, 067901 (2002).	2001597879	✓
I. F. Blake, Codes over certain rings, Inform. Control 20(1972), 396-404.	2016367126	✓

E. Braaten and H. W. Hammer, Phys. Rev. Lett. 91, 102002 (2003) [arXiv:nucl-th/0303038].	2157831165	✓
J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In NIPS, 2016.	2407521645	✓
Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints, abs/1605.02688.	2384495648	✓
C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, pages 424–438, 1969.	2178225550	✓
M. Eckstein and P. Werner, Phys. Rev. B 82, 115115 (2010).	2075658311	✓
H. Halpern, V. Kaftal and G. Weiss, The Relative Dixmier Property in Discrete Crossed Products, J. Functional Anal. 68(1986),	1993094219	✓
H. Li and F. D. M. Haldane, Phys. Rev. Lett. 101, 010504 (2008).	2082257352	✓
Bouchaud J.-P., Potters M., Meyer M. Apparent multifractality in financial time series Eur. Phys. J. B 13 (2000) 595-599	2110077303	✓
H Nakada, Phys Rev C 68, 014316 (2003)	2005421586	✓
P. D. B. Collins, An Introduction to Regge Theory and High-Energy Physics, . Cambridge 1977, 445p.	2054027276	✓
G. Schütz, S. Sandow, Non–Abelian symmetries of stochastic processes: Derivation of correlation functions for random–vertex models and disordered–interacting–particle systems, Phys. Rev. E 49, 2726 (1994) DOI:10.1103/PhysRevE.49.2726	2012762389	✓
Waxman, E. 2010, arXiv:1010.5007	2134348647	✓
W. Wang, S. G. Hanson, Y. Miyamoto, and M. Takeda, Phys. Rev. Lett. 94, 103902 (2005).	2066919669	✓
W. Schoutens. Lévy Processes in Finance: Pricing Financial Derivatives. Wiley series in probability and statistics, Wiley, Chichester, 2003.	2002530172	✓
K. K. Kwong, Some sharp Hodge Laplacian and Steklov eigenvalue estimates for differential forms, Calc. Var. Partial Differential Equations 55 (2016), no. 2, Art. 38, 14. MR 3478292	2309623648	✓
T. Iwashyna, J. Christie, J. Kahn, and D. Asch. Uncharted paths: hospital networks in critical care. CHEST, 135(3):827–833, 2009.	2101874017	✓
A. Demir, A. Mehrotra, and J. Roychowdhury, "Phase noise in oscillators: A unifying theory and numerical methods for characterization," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications 47, 655–674 (2000).	2111079473	✓
]csiszar78a Csiszár, I. and J. Körner, 1978, Broadcast Channels with Confidential Messages, IEEE Trans. Inf. Theory IT-24, 339.	2144007657	✓
F. Beaudoin, J. M. Gambetta, and A. Blais, Phys. Rev. A 84, 043832 (2011).	2074456944	✓
R. Ablamowicz, B. Fauser, On the transposition anti-involution in real Clifford algebras III: The automorphism group of the transposition scalar product on spinor spaces, Linear and Multilinear algebra, 60(6), 2012, 621-644.	2080980639	✓
H. Kanao, S. Tanaka, S. Oka, M. Hirata, S. Yoshida, K. Chayama, Narrow-band imaging magnification predicts the histology and invasion depth of colorectal tumors., Gastrointest Endosc 69 (3 Pt 2) (2009) 631–636.	2038420951	✓
S.Q. Murphy et al., Phys. Rev. Lett. 72, 728 (1994).	2088746433	✓
Y. Kubo, C. Grezes, A. Dewes, T. Umeda, J. Isoya, H. Sumiya, N. Morishita, H. Abe, S. Onoda, T. Ohshima, V. Jacques, A. Dréau, J. -F. Roch, I. Diniz, A. Auffeves, D. Vion, D. Esteve, and P. Bertet. Phys. Rev. Lett. 107, 220501 (2011).	2064756338	✓
Uriel Frisch, Turbulence: The Legacy of A.N. Kolmogorov. (Cambridge University Press, Cambridge, 1995).	1611318213	✓
Joe, H. (1997), Multivariate Models and Multivariate Dependence Concepts, Boca Raton, FL: CRC Press.	2507039649	✓
A. Hanke, F. Schlesener, E. Eisenriegler, and S. Dietrich, Phys. Rev. Lett. 81, 1885 (1998b).	2054116307	✓
O. Hohm, D. Lüst and B. Zwiebach, Fortsch. Phys. 61 (2013) 926 [arXiv:1309.2977 [hep-th]].	2139390505	✓
F. Karsch and E. Laermann, Phys. Rev. D 50, 6954 (1994) .	2132089715	✓
Lev Kapitanski. Global and unique weak solutions of nonlinear wave equations. Math. Res. Lett., 1(2):211–223, 1994.	2060831697	✓
S.-W. Wei snd Y.-X. Liu, Critical phenomena and thermodynamic geometry of charged Gauss-Bonnet AdS black holes, Phys. Rev. D 87 044014 (2013). arXiv:1209.1707 [gr-qc].	2095004813	✓
S. A. Haine and J. J. Hope, Phys. Rev. A 72, 033601 (2005).	2092020630	✓
T. Cochran, A. Gerges, and K. Orr, Dehn surgery equivalence relations on 3-manifolds, Math. Proc. Camb. Phil. Soc. 131 (2001) 97-127.	2163237620	✓
K. Chou and X.-J. Wang, The Lp-Minkowski problem and the Minkowski problem in centroaffine geometry, Adv. Math. 205 (2006), 33—83.	2082611711	✓
C. S. O'Hern, D. A. Egolf, and H. S. Greenside, Phys. Rev. E 53, 3374 (1996).	2002846857	✓
Vincent, O.; Szenicer, A.; Stroock, A. D. Capillarity-driven flows at the continuum limit. Soft Matter 2016, 12, 6656–6661.	2191910464	✓

B Evaluation of tools for claim extraction

Listing B.1 shows our small preliminary evaluation of Open IE 5.0, PredPatt, ClausIE and Ollie using 5 citation contexts of 3 sentences each. Because Open IE 5.0 and PredPatt showed the most promising output on a first look, their precision and recall values were determined, leading to the conclusion that PredPatt performed best.

Listing B.1: Manual evaluation of tools for claim extraction.

```
- Recall: count correct claims (✓) as 1 and kind of correct ones (Δ) as .5
          total is number of claims identified by hand
- Precision: count correct claims (✓)
             ignore kind of correct/trivial claim (Δ)
             total is sum of correct (✓) and incorrect (X) claims

- Open IE 5.0: 38/48 (see details below)
- PredPatt: 59/77 (see details below)
- ClausIE: has obvious problems with CIT tokens
- Ollie: often chooses very large chunks as predicates

### overall

(O=Open IE 5.0; P=PredPatt)

   Recall  Precision            Recall  Precision
O:  3.5/5        3/4        P:     2/5        2/2
O:   6/11       7/13        P:    7/11        7/9
O:    1/6        1/2        P:     5/6        5/7
O:  1.5/8        2/6        P:   2.5/8        3/3
O:    3/9        3/8        P:   6.5/9        3/5
∑   15/39      16/33        ∑    23/39      20/26
      38%        48%               59%        77%

##### 1 #####

One main contribution in this paper is the introduction of a projection operator which achieves the same
     effect. In CIT we already discussed this operator which has recently also been addressed under the name "
     reduced stabilization" in CIT , CIT , CIT , MAINCIT . In these papers the construction of the
     corresponding operator is only possible in two dimensions using special integration rules.
```

```
# by hand
OP
✓✓ [One main contribution] is [the introduction of a projection operator]
✓  [projection operator] achieves [the same effect]
✓✓ [we] discussed [this operator]
Δ  [this operator] has recently been addressed [under the name “reduced stabilization”]
   [construction of the corresponding operator] is only possible [in two dimensions using special integration
        rules]

   Recall Precision
O:   3.5/5      3/4
P:     2/5      2/2

# Open IE 5
Δ0.73 (this operator; has also been addressed; T:recently)
X0.90 (two dimensions; using; special integration rules)
✓0.93 Context(we discussed ,List([120, 140))):(One main contribution in this paper; is; the introduction of a
      projection operator)
✓0.65 (we; discussed; this operator which has recently also been addressed under the name; L:In CIT; T:already
      )
Δ0.97 (the construction of the corresponding operator; is; only possible; L:In these papers)
✓0.90 (a projection operator; achieves; the same effect)

# PredPatt
✓   ?a is the introduction of ?b
        ?a: One main contribution in this paper
        ?b: a projection operator
Δ   ?a achieves ?b
        ?a: which
        ?b: the same effect. In CIT we already discussed this operator
✓   ?a already discussed ?b
        ?a: we
        ?b: this operator
Δ   ?a is only possible
        ?a: the construction of the corresponding operator
Δ   using ?a
        ?a: special integration rules..

# ClausIE
|One main contribution in this paper| |is| |the introduction which achieves the same effect In CIT of a
      projection operator|
|One main contribution in this paper| |is| |the introduction which achieves the same effect In CIT|
|which| |achieves| |the same effect|
|we| |discussed| |this operator CIT already|
|we| |discussed| |this operator CIT|
|this operator| |has been addressed| |under the name reduced stabilization recently|
|this operator| |has been addressed| |under the name reduced stabilization also|
|this operator| |has been addressed| |under the name reduced stabilization in CIT CIT|
|this operator| |has been addressed| |under the name reduced stabilization|
|CIT CIT| |is| |CIT|
|CIT CIT| |is| |MAINCIT|
|the construction of the corresponding operator| |is| |possible using special integration rules MAINCIT In
      these papers|
|the construction of the corresponding operator| |is| |possible only using special integration rules MAINCIT|
|the construction of the corresponding operator| |is| |possible in two dimensions using special integration
      rules MAINCIT|
```

|the construction of the corresponding operator| |is| |possible using special integration rules MAINCIT|

Ollie
0.924: (One main contribution; is the introduction of; a projection operator which achieves the same effect)
0.843: (the construction of the corresponding operator; is only possible in; two dimensions using special integration rules)
0.778: (the construction of the corresponding operator; is only; possible)
0.767: (we; already discussed; this operator which has recently also been addressed under the name)
0.736: (One main contribution; is; the introduction of a projection operator)
0.693: (this operator which has recently also been addressed under the name; be already discussed in; these papers the construction of the corresponding operator is only possible in two dimensions)
0.564: (the same effect; be achieves by; a projection operator)
0.563: (we; already discussed this operator which has recently also been addressed under the name in; these papers the construction of the corresponding operator is only possible in two dimensions)
0.524: (this operator which has recently also been addressed under the name; be already discussed in; CIT)
0.441: (the introduction of a projection operator; be One main contribution in; this paper)
0.4: (we; already discussed this operator which has recently also been addressed under the name in; CIT)
0.399: (possible; be the construction of; the corresponding operator)
0.187: (CIT; be stabilization in; CIT)

2

In this paper we use the HDG approach to construct a simple discretization of the Navier-Stokes equations in which the approximate velocity field is FORMULA -conforming and pointwise divergence-free. To achieve this, we first note that unlike many other HDG methods for incompressible flows CIT , CIT , CIT , CIT , CIT , CIT , MAINCIT , the HDG methods of CIT and CIT contain also facet unknowns for the pressure. The element pressure unknowns play the role of cell-wise Lagrange multiplier to enforce the continuity, whereas the facet pressure unknowns play the role of Lagrange multipliers enforcing continuity of the normal component of the velocity across cell boundaries.

by hand
OP
✓✓ [we] use [the HDG approach]
Δ✓ [we] construct [a simple discretization of the Navier-Stokes equations]
✓Δ [the approximate velocity field] is [FORMULA -conforming]
[the approximate velocity field] is [pointwise divergence-free]
Δ [we] note [that ...]
✓Δ [the HDG methods of CIT and CIT] contain [facet unknowns for the pressure]
✓ [many other HDG methods for incompressible flows CIT , CIT , CIT , CIT , CIT , CIT , MAINCIT] do not contain [facet unknowns for the pressure]
✓✓ [The element pressure unknowns] play [the role of cell-wise Lagrange multiplier to enforce the continuity]
✓ [The element pressure unknowns/Lagrange multiplier] enforce [the continuity]
Δ✓ [facet pressure unknowns] play [the role of Lagrange multipliers enforcing continuity of the normal component
of the velocity]
[facet pressure unknowns/Lagrange multiplier] enforce [continuity of the normal component of the velocity]

	Recall	Precision
O:	6/11	7/13
P:	7/11	7/9

Open IE 5
X0.81 Context(we note the HDG methods of CIT contain The element pressure unknowns play,List([108, 332))):(the HDG methods of CIT; contain facet; unknowns for the pressure)
✓0.87 Context(we note The element pressure unknowns play,List([108, 332))):(the HDG methods of CIT; contain;

```
also facet unknowns for the pressure)
✓0.91 Context(we note,List([108, 121))):(The element pressure unknowns; play; the role of cell-wise Lagrange
      multiplier)
X0.13 Context(we use we note,List([14, 121))):(we; use the HDG approach To achieve; this)
X0.39 Context(we use we note,List([14, 121))):(we; use the HDG approach to construct; a simple pointwise
      divergence - free)
X0.58 Context(we note,List([108, 121))):(we; use; the HDG approach; to construct a simple pointwise divergence
      - free; L:In this paper)
X0.81 Context(we note the HDG methods of CIT contain The element pressure unknowns play,List([192, 416))):(the
      HDG methods of CIT; contain facet; unknowns for the pressure)
✓0.87 Context(we note The element pressure unknowns play,List([192, 416))):(the HDG methods of CIT; contain;
      also facet unknowns for the pressure)
Δ0.91 Context(we note,List([192, 205))):(The element pressure unknowns; play; the role of cell-wise Lagrange
      multiplier)
X0.13 Context(we use we note,List([14, 205))):(we; use the HDG approach To achieve; this)
✓0.43 Context(we use we note,List([14, 205))):(we; use the HDG approach to construct; a simple discretization
      of the Navier - Stokes equations)
✓0.69 Context(we note,List([192, 205))):(we; use; the HDG approach; to construct a simple discretization of
      the Navier - Stokes equations; L:In this paper)
Δ0.40 (we; note; that unlike many other HDG methods for incompressible flows CIT , CIT , CIT , CIT , CIT , CIT
      , MAINCIT , the HDG methods of CIT contain also facet unknowns for the pressure . The element pressure
      unknowns play the role of cell-wise Lagrange multiplier; T:first)
✓0.97 (the approximate velocity field; is; FORMULA - conforming; L:the Navier - Stokes equations)
✓0.93 (the facet pressure unknowns; play; the role of Lagrange multipliers)
Δ0.95 (Lagrange multipliers; enforcing; continuity of the normal component of the velocity across cell
      boundaries)

# PredPatt
✓✓  ?a use ?b to construct ?c
        ?a: we
        ?b: the HDG approach
        ?c: a simple discretization of the Navier-Stokes equations in which the approximate velocity field is
             FORMULA -conforming and pointwise divergence-free.
Δ   in ?a ?b is FORMULA -conforming
        ?a: which
        ?b: the approximate velocity field
X   ?a pointwise divergence-free.
        ?a: the approximate velocity field
Δ   ?a achieve ?b
        ?a: we
        ?b: this
X   In ?a , ?b first note ?c
        ?a: this paper we use the HDG approach to construct a simple discretization of the Navier-Stokes
             equations in which the approximate velocity field is FORMULA -conforming and pointwise divergence
             -free. To achieve this
        ?b: we
        ?c: SOMETHING := unlike many other HDG methods for incompressible flows CIT , CIT , CIT , CIT , CIT ,
             CIT , MAINCIT , the HDG methods of CIT and CIT contain also facet unknowns for the pressure. The
             element pressure unknowns play the role of cell-wise Lagrange multiplier to enforce the
             continuity , whereas the facet pressure unknowns play the role of Lagrange multipliers enforcing
             continuity of the normal component of the velocity across cell boundaries
    unlike ?a , ?b contain also ?c
✓✓      ?a: many other HDG methods for incompressible flows CIT , CIT , CIT , CIT , CIT , CIT , MAINCIT
        ?b: the HDG methods of CIT and CIT
        ?c: facet unknowns for the pressure. The element pressure unknowns play the role of cell-wise Lagrange
              multiplier to enforce the continuity
```

✓ ?a play ?b

?a: The element pressure unknowns

?b: the role of cell-wise Lagrange multiplier

✓ ?a enforce ?b

?a: The element pressure unknowns

?b: the continuity

✓ ?a play ?b

?a: the facet pressure unknowns

?b: the role of Lagrange multipliers enforcing continuity of the normal component of the velocity across cell boundaries

Δ enforcing ?a

?a: continuity of the normal component of the velocity across cell boundaries

ClausIE

|we| |use| |the HDG approach to construct a simple discretization of the Navier Stokes equations this paper|
|the HDG approach| |to construct| |a simple discretization of the Navier Stokes equations|
|the approximate velocity field| |is| |FORMULA conforming divergence-free To achieve this in a simple discretization of the Navier Stokes equations|
|the approximate velocity field| |is| |FORMULA pointwise divergence-free To achieve this in a simple discretization of the Navier Stokes equations|
|the approximate velocity field| |is| |FORMULA conforming divergence-free To achieve this|
|the approximate velocity field| |is| |FORMULA pointwise divergence-free To achieve this|
|we| |note| |In this paper|
|we| |note| |first|
|we| |note|
|facet unknowns for the pressure| |flows| |CIT MAINCIT unlike many other HDG methods for incompressible|
|facet unknowns for the pressure| |flows| |CIT MAINCIT|
|CIT MAINCIT| |is| |CIT|
|CIT MAINCIT| |is| |CIT|
|CIT MAINCIT| |is| |CIT|
|CIT MAINCIT| |is| |CIT|
|CIT MAINCIT| |is| |CIT|
|the HDG methods of CIT| |contain| |also CIT MAINCIT|
|the HDG methods of CIT| |contain| |also CIT MAINCIT|
|the HDG methods of CIT| |contain| |CIT MAINCIT|
|the HDG methods of CIT| |contain| |CIT MAINCIT|
|The element pressure unknowns| |play| |the role of cell-wise Lagrange multiplier to enforce the continuity whereas the facet pressure unknowns play the role of Lagrange multipliers enforcing continuity of the normal component of the velocity across cell boundaries|
|The element pressure unknowns| |play| |the role of cell-wise Lagrange multiplier to enforce the continuity|
|the role of cell-wise Lagrange multiplier| |to enforce| |the continuity|
|the facet pressure unknowns| |play| |the role of Lagrange multipliers enforcing continuity of the normal component of the velocity across cell boundaries|
|Lagrange multipliers| |be enforcing| |continuity of the normal component of the velocity across cell boundaries|

Ollie

0.756: (the role of cell-wise Lagrange; multiplier to enforce; the continuity)
0.691: (we; use; the HDG approach)
0.681: (the facet pressure unknowns; play; the role of Lagrange multipliers)
0.649: (the approximate velocity field; is; FORMULA -conforming and pointwise divergence-free)
0.447: (normal; be component of; the velocity)
0.362: (normal; be component of the velocity across; cell boundaries)
0.241: (we; use the HDG approach to construct a simple discretization of the Navier-Stokes equations in which the approximate velocity field is FORMULA -conforming and pointwise divergence-free in; this paper)

3

Their application to the incompressible Navier-Stokes problem has been considered in several works starting from the early 00's; a non exhaustive list of references includes CIT , CIT , CIT , CIT , CIT , CIT , CIT , CIT , CIT ; cf. also CIT for a pedagogical introduction. The Hybridizable discontinuous Galerkin (HDG) method of CIT , CIT has also been applied to the discretization of the incompressible Navier-Stokes equations in several recent works CIT , CIT , CIT , MAINCIT , CIT . Albeit this is not explicitly pointed out in all of the above references, also HDG methods often support general meshes as well as the possibility to increase the approximation order.

by hand
OP
✓ [Their application to the incompressible Navier-Stokes problem] has been considered [in several works]
✓✓ [a non exhaustive list of references] includes [CIT , CIT , CIT , CIT , CIT , CIT , CIT , CIT , CIT]
✓ [The Hybridizable discontinuous Galerkin (HDG) method of CIT , CIT] has been applied [to the discretization of the incompressible Navier-Stokes equations in several recent works CIT , CIT , CIT , MAINCIT , CIT .]
✓ [this] is not explicitly pointed out
Δ [HDG methods] support [general meshes]
Δ [HDG methods] support [the possibility to increase the approximation order]

	Recall	Precision
O:	1/6	1/2
P:	5/6	5/7

Open IE 5
Δ0.38 (this; is not explicitly pointed out; L:in all of the above references)
X0.88 (CIT; has been applied; to the discretization of the incompressible Navier-Stokes equations)
✓0.95 (a non exhaustive list of references; includes; CIT , CIT , CIT , CIT , CIT , CIT , CIT , CIT , CIT)
Δ0.90 (several works; starting; T:from the early 00's)

PredPatt
✓ ?a has been considered in ?b
?a: Their application to the incompressible Navier-Stokes problem
?b: several works
✓ ?a includes ?b
?a: a non exhaustive list of references
?b: CIT , CIT , CIT , CIT , CIT , CIT , CIT , CIT , CIT
X ?a also CIT for ?b
?a: cf.
?b: a pedagogical introduction. The Hybridizable discontinuous Galerkin -LRB- HDG -RRB- method of CIT , CIT has also been applied to the discretization of the incompressible Navier-Stokes equations in several recent works this is not explicitly pointed out in all of the above references
✓ ?a has also been applied to ?b
?a: The Hybridizable discontinuous Galerkin -LRB- HDG -RRB- method of CIT , CIT
?b: the discretization of the incompressible Navier-Stokes equations in several recent works this is not explicitly pointed out in all of the above references
✓ ?a is not explicitly pointed out in ?b
?a: this
?b: all of the above references
✓ also ?a often support ?b
?a: HDG methods
?b: general meshes as well as the possibility to increase the approximation order
X increase ?a
?a: the approximation order

ClausIE

|Their| |has| |application to the incompressible Navier Stokes problem|
|Their application to the incompressible Navier Stokes problem| |has been considered| |in several works CIT has also been applied to the discretization of the incompressible Navier Stokes equations in several recent works CIT MAINCIT also HDG methods often support general meshes to increase the approximation order|
|Their application to the incompressible Navier Stokes problem| |has been considered| |in several works CIT has also been applied to the discretization of the incompressible Navier Stokes equations in several recent works CIT MAINCIT also HDG methods often support the possibility to increase the approximation order|
|Their application to the incompressible Navier Stokes problem| |has been considered| |starting from the early 00 's CIT has also been applied to the discretization of the incompressible Navier Stokes equations in several recent works CIT MAINCIT also HDG methods often support general meshes to increase the approximation order|
|Their application to the incompressible Navier Stokes problem| |has been considered| |starting from the early 00 's CIT has also been applied to the discretization of the incompressible Navier Stokes equations in several recent works CIT MAINCIT also HDG methods often support the possibility to increase the approximation order|
|Their application to the incompressible Navier Stokes problem| |has been considered| |CIT has also been applied to the discretization of the incompressible Navier Stokes equations in several recent works CIT MAINCIT also HDG methods often support general meshes to increase the approximation order|
|Their application to the incompressible Navier Stokes problem| |has been considered| |CIT has also been applied to the discretization of the incompressible Navier Stokes equations in several recent works CIT MAINCIT also HDG methods often support the possibility to increase the approximation order|
|a non exhaustive list of references| |includes| |CIT CIT cf. also CIT for a pedagogical introduction The Hybridizable discontinuous Galerkin HDG method of CIT|
CIT CIT cf. also CIT for a pedagogical introduction The Hybridizable discontinuous Galerkin HDG method of CIT		is		CIT
CIT CIT cf. also CIT for a pedagogical introduction The Hybridizable discontinuous Galerkin HDG method of CIT		is		CIT
CIT CIT cf. also CIT for a pedagogical introduction The Hybridizable discontinuous Galerkin HDG method of CIT		is		CIT
CIT CIT cf. also CIT for a pedagogical introduction The Hybridizable discontinuous Galerkin HDG method of CIT		is		CIT
CIT CIT cf. also CIT for a pedagogical introduction The Hybridizable discontinuous Galerkin HDG method of CIT		is		CIT
CIT CIT cf. also CIT for a pedagogical introduction The Hybridizable discontinuous Galerkin HDG method of CIT		is		CIT
CIT CIT cf. also CIT for a pedagogical introduction The Hybridizable discontinuous Galerkin HDG method of CIT		is		CIT
cf. The Hybridizable discontinuous Galerkin HDG method of CIT		be CIT		also for a pedagogical introduction
cf. The Hybridizable discontinuous Galerkin HDG method of CIT		be CIT		also
CIT		has been applied		CIT MAINCIT also
CIT		has been applied		to the discretization of the incompressible Navier Stokes equations in several recent works CIT MAINCIT
CIT		has been applied		CIT MAINCIT
CIT MAINCIT		is		CIT
CIT MAINCIT		is		CIT
CIT MAINCIT		is		CIT Albeit
this		is not pointed out		CIT Albeit explicitly
this		is not pointed out		in all of the above references CIT Albeit
this		is not pointed out		CIT Albeit
HDG methods		support		general meshes to increase the approximation order also
HDG methods		support		the possibility to increase the approximation order also
HDG methods		support		general meshes to increase the approximation order often

|HDG methods| |support| |the possibility to increase the approximation order often|
|HDG methods| |support| |general meshes to increase the approximation order|
|HDG methods| |support| |the possibility to increase the approximation order|

Ollie
0.952: (Their application; has been considered in; several works starting from the early 00's)
0.759: ((HDG) method of CIT; has also been applied to; the discretization of the incompressible Navier-Stokes equations)
0.735: ((HDG) method of CIT; has also been applied in; several recent works CIT , CIT , CIT , MAINCIT , CIT . Albeit this is not explicitly pointed out in all of the above references , also HDG methods often support general meshes as well as the possibility)
0.606: (the above references; often support; general meshes)
0.571: (several works; starting from; the early 00's)
0.126: (cf; be a non exhaustive list of; references)

4

When polynomials of degree FORMULA are used, we show that both the energy-norm of the velocity and the FORMULA -norm of the pressure converge as FORMULA (FORMULA denotes here the meshsize). These convergence rates are similar to the ones recently derived in MAINCIT for a HDG method with pressure and velocity spaces chosen as in CIT , CIT . A major difference with respect to CIT is that we obtain them here using polynomials FORMULA instead of FORMULA inside mesh elements (this is precisely one of the major outcomes of the HHO technology identified in CIT).

by hand
OP
[we] show [that ...]
Δ [the energy-norm of the velocity] converges
[the FORMULA -norm of the pressure] converges
✓ [FORMULA] denotes [the meshsize]
✓ [These convergence rates] are similar to [the ones recently derived in MAINCIT for a HDG method with pressure and velocity spaces chosen as in CIT , CIT .]
✓ [A major difference with respect to CIT] is [that ...]
Δ [we] obtain [them here using polynomials FORMULA instead of FORMULA inside mesh elements]
[this] is [precisely one of the major outcomes of the HHO technology identified in CIT]

	Recall	Precision
O:	1.5/8	2/6
P:	2.5/8	3/3

Open IE 5
X0.93 Context(we show,List([46, 53))):(both the energy - norm of the FORMULA - norm of the pressure; converge; as FORMULA)
X0.51 (we; show; that both the energy - norm of the FORMULA - norm of the pressure converge as FORMULA; T:When polynomials of degree FORMULA are used)
X0.93 Context(we show,List([46, 53))):(both the energy - norm of the FORMULA - norm of the pressure; converge; as FORMULA (FORMULA denotes here the meshsize)
X0.51 (we; show; that both the energy - norm of the FORMULA - norm of the pressure converge as FORMULA (FORMULA denotes here the meshsize; T:When polynomials of degree FORMULA are used)
Δ0.91 (These convergence rates; are; similar to the ones)
✓0.81 (FORMULA; denotes; L:here; the meshsize)
Δ0.92 (velocity spaces; chosen; as in CIT)
Δ0.93 Context(we show,List([46, 53))):(both the energy - norm of the velocity; converge; as FORMULA)
Δ0.47 (we; show; that both the energy - norm of the velocity converge as FORMULA (FORMULA denotes here the meshsize); T:When polynomials of degree FORMULA are used)

✓0.79 (FORMULA; denotes; L:here; T:the meshsize)
Δ0.89 Context(we show, List([46, 53))):(These convergence rates; are; similar to the ones)
0.93 Context(we show, List([46, 53))):(both the energy - norm of the velocity; converge; as FORMULA (FORMULA denotes here the meshsize)
0.47 (we; show; that both the energy - norm of the velocity converge as FORMULA (FORMULA denotes here the meshsize; T:When polynomials of degree FORMULA are used)
0.94 (the ones; derived; in MAINCIT; T:recently)
0.63 (polynomials of degree FORMULA; are used;)

PredPatt
Δ When ?a are used
    ?a: polynomials of degree FORMULA
Δ ?a show ?b ?c
    ?a: we
    ?b: SOMETHING := These convergence rates are similar to the ones recently derived in MAINCIT for a HDG method with pressure and velocity spaces chosen as in CIT , CIT
    ?c: SOMETHING := A major difference with respect to CIT is that we obtain them here using polynomials FORMULA instead of FORMULA inside mesh elements
✓ ?a are similar to ?b
    ?a: These convergence rates
    ?b: the ones recently derived in MAINCIT for a HDG method with pressure and velocity spaces chosen as in CIT , CIT
✓ ?a is ?b
    ?a: A major difference with respect to CIT
    ?b: SOMETHING := we obtain them here using polynomials FORMULA instead of FORMULA inside mesh elements
✓ ?a obtain ?b here using ?c ?d
    ?a: we
    ?b: them
    ?c: polynomials
    ?d: FORMULA

ClausIE
|polynomials of degree FORMULA| |are used| |When|
|polynomials of degree FORMULA| |are used|
|we| |show| |that the energy-norm of the velocity When polynomials of degree FORMULA are used|
|we| |show| |that the FORMULA norm of the pressure converge as FORMULA FORMULA denotes here the meshsize When polynomials of degree FORMULA are used|
|we| |show| |that These convergence When polynomials of degree FORMULA are used|
|we| |show| |that A major difference with respect to CIT this is precisely one of the major outcomes of the HHO technology identified in CIT When polynomials of degree FORMULA are used|
|we| |show| |that the energy-norm of the velocity|
|we| |show| |that the FORMULA norm of the pressure converge as FORMULA FORMULA denotes here the meshsize|
|we| |show| |that These convergence|
|we| |show| |that A major difference with respect to CIT this is precisely one of the major outcomes of the HHO technology identified in CIT|
|FORMULA| |denotes| |here the meshsize|
|FORMULA| |denotes| |the meshsize|
|rates| |are| |similar to the ones recently derived in MAINCIT for a HDG method with pressure spaces chosen as in CIT CIT These convergence|
|rates| |are| |similar to the ones recently derived in MAINCIT for a HDG method with velocity spaces chosen as in CIT CIT These convergence|
|rates| |are| |similar These convergence|
|the ones| |be derived| |recently in MAINCIT for a HDG method|
|the ones| |be derived| |recently with pressure spaces chosen as in CIT CIT|
|the ones| |be derived| |recently with velocity spaces chosen as in CIT CIT|
|the ones| |be derived| |recently|

|pressure spaces| |be chosen| |as in CIT CIT|
|velocity spaces| |be chosen| |as in CIT CIT|
|we| |obtain| |them here using polynomials FORMULA instead of FORMULA inside mesh elements|
|them| |using| |polynomials FORMULA instead of FORMULA inside mesh elements here|
|them| |using| |polynomials FORMULA instead of FORMULA inside mesh elements|
|this| |is| |one precisely|
|this| |is| |one of the major outcomes of the HHO technology identified in CIT|
|this| |is| |one|
|the HHO technology| |be identified| |in CIT|

Ollie
0.787: (These convergence rates; are similar to; the ones)
0.616: (the HHO technology; be identified in; CIT)
0.615: (pressure and velocity spaces; be chosen in; CIT)
0.527: (polynomials FORMULA; be using instead of; FORMULA)
0.517: (we; obtain them here using polynomials FORMULA instead of; FORMULA)
0.423: (polynomials FORMULA; be using inside; mesh elements)
0.417: (we; obtain; them)
0.356: (we; obtain them here using; polynomials FORMULA)
0.339: (we; obtain them here using polynomials FORMULA inside; mesh elements)

5

These convergence rates are similar to the ones recently derived in CIT for a HDG method with pressure and velocity spaces chosen as in CIT , CIT . A major difference with respect to MAINCIT is that we obtain them here using polynomials FORMULA instead of FORMULA inside mesh elements (this is precisely one of the major outcomes of the HHO technology identified in CIT). Another difference with respect to CIT , CIT is that our trilinear form is expressed in terms of a discrete gradient reconstruction and designed so that it does not contribute to the kinetic energy balance, a feature which simplifies several arguments in the analysis; cf. Remark REF for further details.

by hand
OP
✓ [These convergence rates] are similar to [the ones recently derived in MAINCIT for a HDG method with pressure and velocity spaces chosen as in CIT , CIT .]
✓ [A major difference with respect to CIT] is [that ...]
Δ [we] obtain [them here using polynomials FORMULA instead of FORMULA inside mesh elements]
✓ [this] is [precisely one of the major outcomes of the HHO technology identified in CIT]
✓ [Another difference with respect to CIT , CIT] is [that ...]
✓ [our trilinear form] is [expressed in terms of a discrete gradient reconstruction]
[our trilinear form] is [designed so that ...]
✓✓ [it] does not contribute to [the kinetic energy balance]
✓✓ [a feature] simplifies [several arguments in the analysis]

	Recall	Precision
O:	3/9	3/8
P:	6.5/9	3/5

Open IE 5
Δ0.31 Context(CIT is,List([403, 409))):(our trilinear form; is expressed;)
Δ0.92 (velocity spaces; chosen; as in CIT)
✓0.45 (it; does not contribute; to the kinetic energy balance)
X0.69 (CIT; is designed; so that it does not contribute to the kinetic energy balance , a feature which simplifies several arguments in the analysis ; cf . Remark REF for further details)
Δ0.29 Context(These convergence rates are similar to the ones is we obtain,List([0, 120))):(we; obtain them

using; polynomials)
Δ0.26 Context(These convergence rates are similar to the ones is,List([0, 105))):(we; obtain; them; L:here)
X0.85 (These convergence rates are similar to the ones; is; that we obtain them here using polynomials FORMULA instead of FORMULA inside mesh elements (this is precisely one of the major outcomes of the HHO technology identified in CIT) . Another difference with respect to CIT)
✓0.89 (a feature; simplifies; several arguments in the analysis)
Δ0.31 Context(CIT is,List([328, 334))):(our trilinear form; is expressed;)
Δ0.26 Context(These convergence rates are similar to the ones is CIT is,List([0, 334))):(we; obtain; them; L: here)
X0.85 Context(CIT is,List([328, 334))):(These convergence rates are similar to the ones; is; that we obtain them here using polynomials FORMULA instead of FORMULA inside mesh elements (this is precisely one of the major outcomes of the HHO technology identified in CIT) . Another difference with respect to CIT)
X0.67 (CIT; is; that our trilinear form is expressed in terms of a discrete gradient reconstruction)
X0.94 (the HHO technology; identified; L:in CIT)
Δ0.94 (the ones; derived; in CIT; T:recently)
✓0.43 (this; is precisely; one of the major outcomes of the HHO technology)
Δ0.91 (These convergence rates; are; similar to the ones)

PredPatt
✓ ?a are similar to ?b
    ?a: These convergence rates
    ?b: the ones recently derived in CIT for a HDG method with pressure and velocity spaces
Δ ?a obtain ?b here using ?c ?d
    ?a: we
    ?b: them
    ?c: polynomials
    ?d: FORMULA
✓ ?a is ?b
    ?a: Another difference with respect to CIT , CIT
    ?b: SOMETHING := our trilinear form is expressed in terms of a discrete gradient reconstruction
✓ ?a is expressed in ?b
    ?a: our trilinear form
    ?b: terms of a discrete gradient reconstruction
X that ?a does not contribute to ?b
    ?a: it
    ?b: the kinetic energy balance , a feature which simplifies several arguments in the analysis
Δ ?a simplifies ?b in ?c
    ?a: which
    ?b: several arguments
    ?c: the analysis
X ?a cf. ?b for ?c
    ?a: These convergence rates
    ?b: Remark REF
    ?c: further details

ClausIE
|These convergence rates| |are| |similar to the ones recently derived in CIT for a HDG method with pressure spaces chosen as in CIT|
|These convergence rates| |are| |similar to the ones recently derived in CIT for a HDG method with velocity spaces chosen as in CIT|
|These convergence rates| |are| |similar|
|the ones| |be derived| |recently in CIT for a HDG method|
|the ones| |be derived| |recently with pressure spaces chosen as in CIT|
|the ones| |be derived| |recently with velocity spaces chosen as in CIT|
|the ones| |be derived| |recently|
|pressure spaces| |be chosen| |as in CIT|

|velocity spaces| |be chosen| |as in CIT|
|CIT| |is| |CIT this is precisely one of the major outcomes of the HHO technology identified in CIT Another difference with respect to CIT|
|A major difference with respect to MAINCIT| |is| |that we obtain them here using polynomials FORMULA instead of FORMULA inside mesh elements CIT this is precisely one of the major outcomes of the HHO technology identified in CIT Another difference with respect to CIT|
|we| |obtain| |them here using polynomials FORMULA instead of FORMULA inside mesh elements|
|them| |using| |polynomials FORMULA instead of FORMULA inside mesh elements here|
|them| |using| |polynomials FORMULA instead of FORMULA inside mesh elements|
|this| |is| |one precisely|
|this| |is| |one of the major outcomes of the HHO technology identified in CIT|
|this| |is| |one|
|the HHO technology| |be identified| |in CIT|
|CIT| |is| |These convergence rates are similar to the ones recently derived in CIT for a HDG method with pressure spaces chosen as in CIT that our trilinear form is expressed in terms of a discrete gradient reconstruction|
|CIT| |is| |These convergence rates are similar to the ones recently derived in CIT for a HDG method with pressure spaces chosen as in CIT that our trilinear form is designed so that it does not contribute to the kinetic energy balance|
|CIT| |is| |These convergence rates are similar to the ones recently derived in CIT for a HDG method with velocity spaces chosen as in CIT that our trilinear form is expressed in terms of a discrete gradient reconstruction|
|CIT| |is| |These convergence rates are similar to the ones recently derived in CIT for a HDG method with velocity spaces chosen as in CIT that our trilinear form is designed so that it does not contribute to the kinetic energy balance|
|our| |has| |trilinear form|
|our trilinear form| |is expressed| |in terms of a discrete gradient reconstruction|
|our trilinear form| |is designed| |so that it does not contribute to the kinetic energy balance|
|it| |does not contribute| |to the kinetic energy balance|
|the kinetic energy balance| |is| |a feature cf. Remark REF for further details|
|a feature cf. Remark REF for further details| |simplifies| |several arguments in the analysis|

Ollie
0.797: (These convergence rates; are similar to; the ones)
0.782: (a feature; simplifies several arguments in; the analysis)
0.629: (our trilinear form; is expressed; a feature which simplifies several arguments in the analysis)
0.616: (the HHO technology; be identified in; CIT)
0.615: (pressure and velocity spaces; be chosen in; CIT)
0.611: (REF; be Remark for; further details)
0.576: (our trilinear form; is expressed a feature which simplifies several arguments in the analysis ; cf . Remark REF for further details in; terms of a discrete gradient reconstruction)
0.564: (several arguments; be simplifies by; a feature)
0.527: (polynomials FORMULA; be using instead of; FORMULA)
0.517: (we; obtain them here using polynomials FORMULA instead of; FORMULA)
0.454: (it; so does not contribute to; the kinetic energy balance)
0.423: (polynomials FORMULA; be using inside; mesh elements)
0.417: (we; obtain; them)
0.356: (we; obtain them here using; polynomials FORMULA)
0.339: (we; obtain them here using polynomials FORMULA inside; mesh elements)

C Offline evaluation data filter criteria

Table 11 details how the citation contexts for our offline evaluation were chosen. The arXiv data used is the one described in Chapter 4, the MAG data used is a snapshot from February 2019, the RefSeer data set is from Huang et al. [21] available at `https://psu.app.box.com/v/refseer` and the ACL-ARC data is the same as used in Färber et al. [78] available at `http://citation-recommendation.org/publications/#To_Cite_or_Not_to_Cite`.

Table 11: Filter criteria for offline evaluation data.

<table>
<tr><th>Data set</th><th>Filter criteria and reasoning</th></tr>
<tr><td>arXiv</td><td>citing document
from the field of computer science
cited document
has at least 5 citing documents within the data set
Computer science is chosen so that the same data can be used in a user study where participants from the field computer science judge the quality of recommendations.
Filtering out recommendation candidates described by ≤4 contexts is done to ensure a minimum quality of candidate descriptions.</td></tr>
<tr><td>MAG</td><td>citing and cited document
from the field of computer science
in English
abstract in MAG not NULL
cited document
has at least 50 citing documents within the data set
Above criteria give us a number of recommendation candidates to rank that is close to that of the arXiv data while the quality threshold for those candidates is considerably higher.</td></tr>
<tr><td>RefSeer</td><td>citing and cited document
title in DB not NULL
venue in DB not NULL
venuetype in DB not NULL
abstract in DB not NULL
year in DB not NULL
Emulating the filtering criteria used in [24] to remove unclean data.</td></tr>
<tr><td>ACL-ARC</td><td>cited document
needs to have a globally consistent identifier (here a DBLP ID)
Requirement to be able to perform recommendation.</td></tr>
</table>