Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2231–2240
Marseille, 11–16 May 2020

Exploiting Citation Knowledge in Personalised Recommendation of Recent Scientific Publications

Anita Khadka[1], **Iván Cantador**[2], **Miriam Fernandez**[1]
[1]Knowledge Media Institute, The Open University, UK
[2]Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain
anita.khadka@open.ac.uk, ivan.cantador@uam.es, miriam.fernandez@open.ac.uk

Abstract

In this paper we address the problem of providing personalised recommendations of recent scientific publications to a particular user, and explore the use of citation knowledge to do so. For this purpose, we have generated a novel dataset that captures authors' publication history and is enriched with different forms of paper citation knowledge, namely citation graphs, citation positions, citation contexts, and citation types. Through a number of empirical experiments on such dataset, we show that the exploitation of the extracted knowledge, particularly the type of citation, is a promising approach for recommending recently published papers that may not be cited yet. The dataset, which we make publicly available, also represents a valuable resource for further investigation on academic information retrieval and filtering.

Keywords: Research publication dataset, citation context, citation types, recommender systems

1. Introduction

Keeping up with the most recent scientific literature is a challenge for many researchers giving the continuous and increasing growth of academic publications. A recent report[1] claims the existence of 33,100 active scholarly peer-reviewed scientific, technical and medical English-language journals in mid-2018 (plus a further 9,400 non-English-language journals), collectively publishing over 3 million papers a year. The report also states that the production of scientific publications is steadily increasing at a 4% yearly rate.

Aiming to address this information overload problem, academic search engines like Google Scholar[2] or PubMed,[3], as well as Recommender Systems (RS) (Beel et al., 2016), have emerged in the last years. RS in particular have focused on suggesting relevant papers for a given user (Middleton et al., 2001; Torres et al., 2004), but have also addressed other relevant tasks, including recommending relevant papers for a particular snapshot of content, such as title and abstract (Bethard and Jurafsky, 2010; Huang et al., 2015), recommending relevant papers for a given publication (Liang et al., 2011a; Khadka and Knoth, 2018), recommending relevant papers for a particular collection of publications (Ekstrand et al., 2010; Shimbo et al., 2007), and recommending relevant papers for an undergoing manuscript, i.e., a paper yet to be published (He et al., 2010; Strohman et al., 2007).

Most of academic paper recommender systems, however, do not take into consideration the time when papers were published, and do not address the real-world problem of **recommending recently published papers** (Ha et al., 2014). In fact, traditional collaborative filtering systems are ineffective to address that problem, since they are not able to recommend the latest, most recent papers, which have not been rated or cited yet.

On the other hand, while there are approaches that have explored the use of **citation knowledge** to enhance recommendations, e.g., citation graph (i.e., relations between papers based on citations), citation position (e.g., section of the paper where a citation appears), and citation context (i.e., text around a citation), the exploitation of *citation types*, i.e., categories of citations based on their purpose –background, criticism, etc.– to enhance the recommendation of scientific publications, and particularly the recommendation of recent scientific publications, is still under explored.

Addressing this gap, we investigate how citation knowledge could be extracted and exploited to support users towards the discovery of recent and relevant scientific publications. To capture such knowledge, access to the users' publications and the textual content of papers is needed. Existing datasets used for training and testing academic recommender systems do not provide either the entire authors' publication histories or the manuscript texts, but just their metadata, e.g., title, abstract and keywords (see Section 2.4.). As part of our work, we have built and made publicly available a **novel dataset, enriched with different forms of citation knowledge**, to enable the implementation and evaluation of RS for the particular task of recommending recently published papers to users. Hence, we provide the following contributions:

- An in-depth analysis of existing categorisations and taxonomies of citation types.
- A unique dataset that includes authors' publication history and entire textual contents of scientific publications.
- An enrichment of such dataset with four levels of citation knowledge: citation graph, citation position, citation context, and citation type
- An evaluation of how citation knowledge, and particu-

[1] `https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf`

[2] `https://scholar.google.co.uk/`

[3] `https://www.ncbi.nlm.nih.gov/pubmed/`

larly the type of citations, can help on the recommendation of recent scientific publications.

The remainder of this document is structured as follows: Section 2. surveys the state of the art on academic paper recommender systems, focusing on those works that have considered the time dimension and the exploitation of citation knowledge. Section 3. describes the generated dataset and its enrichment with citation knowledge. Next, Section 4. shows the conducted experiments on our dataset, assessing the usefulness of exploiting the types of citation to provide personalised recommendations of recent scientific publications. Finally, Section 5. ends with some discussions and conclusions.

2. State of the Art

Our work focuses on the task of recommending recent scientific publications to users, and exploring the use of citation knowledge to do so. Therefore, in this section we review related literature according to two aspects: publication time awareness and citation knowledge awareness. In addition, we provide a review of existing categorisations of citation types, as well as a review of existing datasets used in the field of academic recommender systems.

2.1. Publication Time Awareness

We address the real world problem where a user is interested in receiving notifications of recently published publications that are relevant to their work or research. Even though the time factor has been previously applied in the recommender systems literature to classify user preferences into short- or long-term interests (Sugiyama and Kan, 2010), or to suggest papers to new users for which there is no prior activity records (Cai et al., 2016; Hristakeva et al., 2017), to the best of our knowledge, only Ha et al. (2014) have considered the scenario of recommending new items, i.e., recent academic papers. The authors proposed a Belief Propagation approach where an undirected graph is built to capture relations between papers based on citations. The approach exploits the graph to compute probabilistic scores for an author based on their publication history. The top-k scored papers are then recommended. The authors experimented with a dataset of 6,241 papers published between 1971 and 2001 by 5,891 academics, in the fields of data mining. The used dataset is not available, and details on where and how the data could be collected are not provided.
Wang and Blei (2011) worked on the related problem of 'out-of-matrix prediction', i.e., recommending papers that have never been rated (although this does not necessarily mean that they are new or recent). The authors proposed a collaborative topic regression model (CTR) that combines collaborative filtering with topic modelling. Their evaluation was conducted on a selection of CiteULike data[4], but the dataset is not available.
From a survey of the literature, we observe that very few works have explored the recommendation of recent papers, and the datasets they used are no longer available. A key goal of our work is therefore the creation of a new dataset that can serve the RS community to further investigate how to design and develop systems capable of recommending the most recent and relevant scientific publications.

2.2. Citation Knowledge Awareness

A number of works have explored the use of citations knowledge to provide better recommendations, including:

- **Citation graph**, where nodes represent papers and edges represent relations between such papers based on their citations. Relations in the citation graph can be either directed –i.e., they capture the explicit source and target papers of the citations–, or undirected –i.e., they do not consider which paper is the one citing and which paper is the one being cited (Torres et al., 2004).

- **Citation position**, or the particular section of the paper where the citation appears, e.g., introduction, literature review, and conclusions (Chakraborty et al., 2016).

- **Citation context**, or text surrounding the citation. This text provides an indication of the semantic context in which the citation is mentioned (Sugiyama and Kan, 2013).

- **Citation type**, or category of citation, which mainly reflects the objective of the citation: background, criticise, support, etc. (Liang et al., 2011b)

Proposed approaches that extract and exploit citation knowledge have mainly focused on recommending a list of relevant papers for a given target paper (He et al., 2010; Khadka and Knoth, 2018; McNee et al., 2002; Chakraborty et al., 2016; Liang et al., 2011b), or on recommending relevant papers for a fragment of text (Bethard and Jurafsky, 2010; Huang et al., 2015). Fewer works, in contrast, have focused on recommending relevant papers for a given user (Sugiyama and Kan, 2013; Sugiyama and Kan, 2010; Torres et al., 2004), and to the best of our knowledge, none of them have addressed the particular use case of recommending recently published scientific papers.
Sugiyama and Kan (2010) explored the use of the *citation graph* to enrich user profiles, capturing users' research interests by considering not only their past publications, but also the citations of such publications, and the publications citing the users' work. In an extension of that work (Sugiyama and Kan, 2013), the authors also explored the use of *citation context*, or text around the citations, to enhance recommendations.
Citation positions, in combination with the citation graph, were explored by Chakraborty et al. (2016). They generated a directed graph linking papers based on citations and enriched with information about the positions of such citations, i.e., an edge representing a citation from paper p_1 to paper p_2 may be tagged with multiple labels, if p_1 cites p_2 in several sections. This graph is then used to provide recommendations (for a given query paper). Also in the context of recommending papers for a given paper, Liang et al. (2011b) proposed a method that explores citation graphs including *citation types*. Three types of citation were considered: *Based-on* –referring to when the citation highlights a paper that the current work is based

[4] http://www.citeulike.org/faq/data.a

on–, *Comparable* –referring to when the citation highlights a paper that is used to reflect on similarities or differences–, and *General* –referring to all citations that can not be categorised as Based-on or comparable, such as those that provide background knowledge, for example.

While these works show how citation knowledge can help enhancing recommendation accuracy, they do not consider the particular use case of recommending the latest scientific publications to users. Moreover, while most reported approaches have focused on the use of the citation graph, citation contexts, and citation positions, very few works, such as (Liang et al., 2011b), have explored the exploitation of citation types for recommendation.

A key goal of our work is therefore to provide language resources that (i) will enable the investigation of recommender systems for the particular real world scenario of recommending recent scientific publications to users, and (ii) will provide rich citation knowledge, in the form of citation graph, citation positions, citation contexts, and particularly citation types. Since citation types has been unexplored in the field of recommender systems, the next section aims to provide a comprehensive overview of taxonomies proposed in other fields, identifying those citation types that have been commonly used.

2.3. Citation Types

While citation types (categories) are, to the best of our knowledge, underexplored in the RS field, there has been a considerable amount of effort invested in the definition and exploitation of citation types in other areas, such as Scientometrics. For example, they have been used to study the evolution of scientific fields (Jurgens et al., 2018a), or to quantify the influence of research works (Valenzuela et al., 2015; Hassan et al., 2017). In this section, we discussed different citation categorisations and taxonomies existing in the literature. We aim to provide an in-deep analysis of the similarities and differences of such categorisations and to conduct an informed selection of citation types for our research. A summary of this analysis is provided in Table 1.

For instance, Nanba and Okumura (1999) proposed three classes: Type B (the citations to *base* on other researchers' theories), Type C (the citations to *compare* with related works or to point out their problems), and Type O (the citations *other* than types B and C). Jurgens et al. (2018a) considered six classes (*Background, Compare or Contrast, Motivation, Use, Extend* and *Future*) for measuring the evolution of scientific fields. Valenzuela and Hassan (Valenzuela et al., 2015; Hassan et al., 2018) referred to two classes (*important* –using and extending– and *not important* –related and comparison–) for quantifying the importance of cited references.

As can be observed in the Table 1, a significant amount of works have proposed different types of citations. However, on a close inspection, similarities emerge among existing taxonomies. For instance, (Nanba and Okumura, 1999; Nanba et al., 2000) used the types *Basis*, *Compare* and *Other*. These citation types are very similar in terms of semantic meaning to the ones proposed by (Liang et al., 2011b): *Based-on*, *Comparable* and *General*.

The Table 1 shows our summary and analysis on the different categorisations of citations identified in the literature, and proposes a high level categorisation, which includes the following types:

- **Background**: This type refers to citations providing general scientific knowledge towards a domain. This class has been named differently in previous works as background, organic, general, assumptive, read alert, and neutral, among others.

- **Discuss, compare and contrast**: This type refers to papers that are cited to discuss the proposed work or provide comparisons against it (i.e., to support it, dispute it, correct it, etc.). Two key subtypes in this category are also identified in the literature: **Criticise** and **Support**. These subtypes are also called as negative or negational (for citations that are used to express criticism), and positive or affirmational (for citations that are used to support the work).

- **Use**: This type refers to papers that are cited because the current work uses them. It has been also referred in the literature as technical basis, useful, or method.

- **Extension**: This type refers to papers that are cited because the current work extends them. It is also referred to as expand, evolutionary, developmental, and based on.

- **Motivation**: This type refers to cited papers that motivate or inspire the current work.

- **Future**: This type refers to cited papers that inspire future work.

Identifying the particular type of a citation is a complex task. While some works have focused on manually labelling citations based on proposed categories (Teufel, 1999; Moravcsik and Murugesan, 1975), others have attempted to develop methods to automatic categorisation. These methods are based on the use of cue phrases (Teufel et al., 2006b; Teufel et al., 2006a; Liang et al., 2011b), syntactic patterns (Meyers, 2013), rule-based heuristics (Abu-Jbara et al., 2013; Garzone and Mercer, 2000; Nanba et al., 2000), or Machine Learning (Hassan et al., 2018; Jurgens et al., 2018a; Cohan et al., 2019). Since the high-level citation types identified in our work (see Table 1) are very similar to the ones identified by (Jurgens et al., 2018a), we apply their classifier in our work, which was trained with computer science papers, an more particularly Natural Language Processing (NLP) papers.

2.4. Existing Datasets

Various datasets have been used in the past to investigate academic recommender systems. A comprehensive list of publicly available datasets is given in Table 2. For each dataset, the table shows a brief description of the type of data, the number of users, items and ratings in the dataset, and whether the full text PDF_{av} and publication history of the users UPH_{av} are available.

As we can observe, most of the existing datasets do not provide the full texts of publications, and hence, citation

		Discuss, Compare, Control						
Work	**Background**	**Description**	**Criticise**	**Support**	**Use**	**Extension**	**Motivation**	**Future**
(Jurgens et al.,2018)	Background		Compare or Contrast		Use	Extension	Motivation	Future
(Teufel, 1999), (Merity et al., 2009)		Other, Own, Textual	Contrast			Basis	Aim	
(Cohan et al., 2019)	Background Information		Result Comparison		Method			
(Dong and Schafer, 2011			Comparison		Technical Basis		Fundamental Idea	
(Chakraborty et al., 2016)	Background		Alternative Approaches, Comparison		Methods			
(Jochim andSchutze, 2012)	CONC-OP (Conceptual - Operational)	ORG, PERF (Organic - Perfunctory)	NEG, JUX (Negational, Juxtapositional)	CONF (Confirmative)		EVOL (Evolutionary)		
Teufel et al., 2006b), (Su et al., 2018)		Neutral	Compare and Contrast					
			Weakness	Positive				
(Li et al., 2013)	Standard	Neutral, Co-Citation	Corroboration, Contrast		Practical, Supply, Discover	Based on	Significant	
			Negative	Positive				
(Liang et al., 2011FRP)	General		Comparable			Based on		
(Garzone and Mercer, 2000)	Assumptive	Tentative, Reader Alert	Contrastive		Methodological, Use	Interpretational		Future Research
			Negational	Affirmational				
(Nanba and Okumura,1999), (Nanba et al., 2000)	Type O (Other)		Type C (Compare)			Type B (Basis)		
Meyers et al., 2013			Contrast	Corroborate		Expand		
(Bakhti et al., 2018)		Neutral	Mathematical		Useful			
			Contrast	Correct				
(AbuJbara et al.,2013)		Neutral	Comparison		Use	Basis		
			Criticising	Substantiating				
(Valenzuela et al., 2015), (Hassan et al., 2017), (Hassan et al., 2018)		Related Work	Comparison		Using	Extending		
(Pham et al., 2003)			Comparison			Basis		
			Limitation	Support				

Table 1: Types of citations proposed by prior works grouped into top-level classes

Dataset	**Description**	**Users**	**Items**	**Ratings**	PDF_{av}	UPH_{av}
AMiner (Ami, 2019)	AMiner contains a series of datasets capturing relations among citations, academic social networks, topics, etc. We report data here about the citations dataset V11	Not specified	4M	No	No	No
Open Citations (Ope, 2019)	Open repository of scholarly citation data	Not specified	7.5M	No	No	No
Open Academic Graph (OAG, 2019)	Large knowledge graph combining Microsoft Academic Graph and AMiner	253M	381M	No	No	No
ArXiv (ArX, 2019)	Open access e-prints publications in different fields such as physics, mathematics etc.	Not specified	1,5M	No	Yes	No
CORE (COR, 2019)	Dataset of open access research publications published up to 2018	No	9.8M	No	Yes	No
CiteULike (Cit, 2019)	Dataset of users' selected bookmarks to academic papers	5,551	16,980	No	No	No
Mendeley (Jack et al., 2010)	Dataset shared by Mendely for a recommender system challenge	50,000	4.8M	Yes	No	No
ACL anthology (ACL, 2019)	Corpus of scholarly publications about Computational Linguistics	Not specified	22,878	No	Yes	No
SPD 1 (SPD, 2019a)	ACL anthology based papers published between 2000-2006	28	597	Yes	Yes	No
SPD 2 (SPD, 2019b)	ACM proceedings based papers published between 2000-2010	50	100,531	Yes	No	No

Table 2: Publicly available datasets for academic recommender systems. PDF_{av} stands for PDF document availability, and UPH_{av} stands for authors' publication history availability

position, context and type cannot be extracted from them. Similarly, many datasets do not provide the authors' publication history, and thus knowledge about the users, particularly their preferences (i.e., publications and references), cannot be easily captured. To address these limitations we have built a new dataset that includes both the full textual content of papers, and the authors' publication history. The building process of our dataset is described next.

3. Dataset Building

We explain here the construction of the dataset, as well as how citation knowledge has been extracted from it. The dataset is accessible from here [5].

3.1. Collecting Data

As previously explained, we aimed to build a new dataset that (1) could serve to investigate the particular recommendation scenario of discovering the most recent academic papers relevant for a target user, and (2) provides the textual content of papers in addition to their metadata, so that fine grained citation knowledge could be extracted and exploited in information retrieval and filtering tasks.

Since we were interested in exploring the usage of citation knowledge for recommendation purposes, we needed to ensure that there are sufficient items (papers) cited by other items within the dataset. Following this requirement, we gathered the publication history of authors working on the same field (e.g., publishing in the same conference), since they are likely to cite each other's publications.

Specifically, we selected the ACM Conference Series on Recommender Systems (RecSys)[6] and collected data for

[5] https://doi.org/10.21954/ou.rd.10673132.v1

[6] https://recsys.acm.org/

1,931 authors who have published in the conference from the first RecSys conference in 2007 until the twelfth RecSys edition in 2018. The complete publication histories of these authors were collected from the well-known computer science bibliography data provider DBLP[7]. Note that the publication history of an author contains not only their RecSys papers, but also papers published in other venues (journals, conferences, etc.).

As shown in Figure 1, an author's publication history contains metadata of each of the author's papers, including their titles, abstracts, publication dates and venues. The metadata also includes a URL to the corresponding Google Scholar page of each paper, which we used to gather the PDF file of the paper.

Out of the 80,808 papers crawled from DBLP, we obtained textual content for 35,473 of them (about 44%). We note that, while initiatives like open access and pre-prints enabled full access to scientific publications, many of their papers are hidden behind pay-walls and thus are not publicly accessible[8]. Among the 35,473 papers for which we have textual content, we could obtain citation types for 21,924 of them using the approach proposed by (Jurgens et al., 2018a). The parser used as a part of their approach is ParsCit[9] which could only extract citation information for 61% of the collected papers, reducing the completeness of the data. Then, to ensure that we had sufficient historical data to capture user preferences, we discarded all the authors for which we obtained less than four publications, keeping a total of 1,102 authors.

Afterwards, we divided the dataset into training and test sets by observing the publication time distribution, and selected as breaking date the 1st of January 2018 (see Figure 2). All papers published before that date were considered part of the training set. All papers published after that date were considered part of the test set. Lastly, we kept those authors having at least 60% of the data in the training set, and at least 10% of the data in the test set. The final dataset consists of 446 authors and 9,399 academic papers, from which 7,786 belong to the training set and 1,613 represent the test set.

3.2. Modelling Authors, Papers and Citations

Figure 1 shows the different features that are captured for users (authors), publications and citations, as well as the different relations between them. For each user, we consider four different identifiers, including: the internal identifier within the dataset, the ORCID identifier[10], and both, the DBLP and the Google Scholar URLs. In addition, we capture name, last name, website and affiliation.

For each publication, we also extracted multiple identifiers, including the internal identifier within the dataset, the DBLP URL (from where metadata about the paper were extracted) and the Google Scholar URL (from where the PDF file of the paper was downloaded –if available–). Metadata about the publication includes title, abstract, publication date, and publication venue.

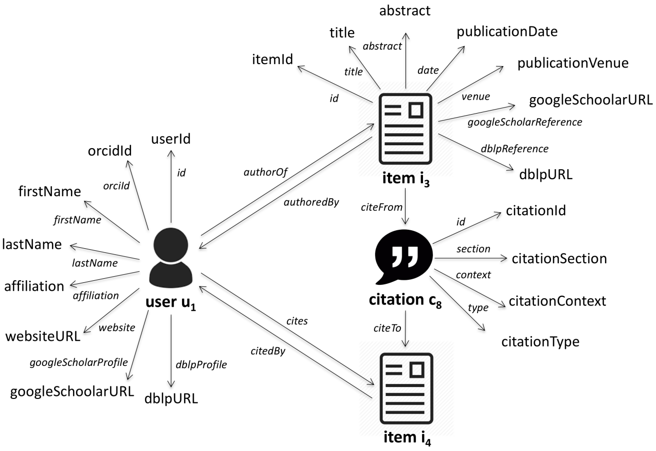


Figure 1: Capturing citation knowledge

To extract **citation knowledge**, we parsed the available PDF files using the GROBID parser[11] and the classifier of citation types provided by (Jurgens et al., 2018a).

3.2.1. Citation Graph

From the PDF of each publication, we extracted the reference list (i.e., all the papers that are cited within the publication). The reference lists are then matched against the 9,399 publications of the dataset to identify the citation-based relations and generate the citation graph. Paper titles, authors, and publication years were matched considering a series of heuristics to minimise errors including: applying lower case, matching at least one author, and computing the Levenshtein distance[12] between the title of publication and the title of the reference. An 85% minimum threshold was empirically selected for this distance. These heuristics were needed since references sometimes contained errors or incomplete information. We identified 1,071 distinct referenced publications within our dataset.

3.2.2. Citation Position

From the PDF of each publication, we extracted the citation position, i.e., the sections within the publication where those citations appear. We are considering four sections in this work: introduction, state of the art, conclusions, and others. The 1,071 distinct referenced publications within our dataset were cited 2,820 times in the introduction sections, 2,784 in the related work sections, 44 in the conclusion sections, and 8,113 in other sections.

3.2.3. Citation Context

From the PDF of each publication, we extracted the citation context (the text that surrounds the citation in the publication). We consider as citation context three sentences: the one where the citation appears, and the ones before and after, when available.

[7] https://dblp.uni-trier.de/

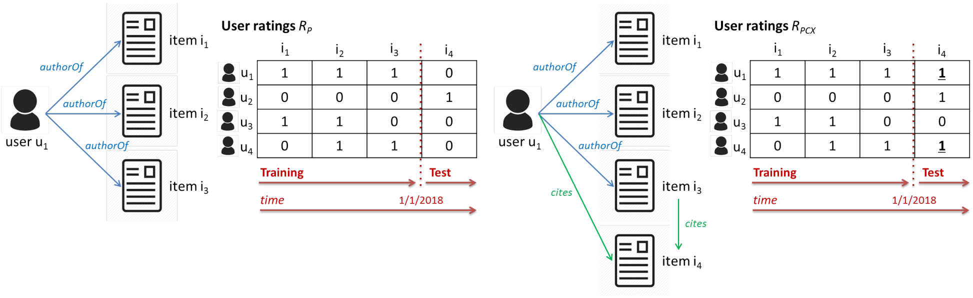
[8] https://www.theguardian.com/higher-education-network/2018/may/21/scientists-access-journals-researcher-article

[9] https://parscit.comp.nus.edu.sg/

[10] https://orcid.org/

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

[12] https://en.wikipedia.org/wiki/Levenshtein_distance



	i_1	i_2	i_3	i_4
u_1	1	1	1	0
u_2	0	0	0	1
u_3	1	1	0	0
u_4	0	1	1	0

	i_1	i_2	i_3	i_4
u_1	1	1	1	<u>**1**</u>
u_2	0	0	0	1
u_3	1	1	0	0
u_4	0	1	1	<u>**1**</u>

Figure 2: Modelling user preferences

3.2.4. Citation Type

As mentioned in Section 2.3., to assign citation types to the referenced papers in our dataset, we used the classifier provided by (Jurgens et al., 2018a), which categorises citations into six main types: background, compare and contrast, use, extension, motivation and future. This classifier was trained with papers from the ACL Anthology corpus[13], a corpus of scholarly publications about Computational Linguistics. To evaluate the performance of this classifier when identifying citation types from the RS field we conducted a manual assessment of 100 randomly selected citations. Based on 3 annotators (all of them with computer science background), displaying a moderate agreement (0.62 Fleiss kappa (Fleiss and Cohen, 1973)), we identified a 69% average correct classification, which is in line to the one reported by the authors (Jurgens et al., 2018a). The 1,071 distinct referenced publications were classified 9,933 times as *background*, 3,032 as *CompareOrContrast*, 225 times as *Extend*, 449 times as *Use* 83 times as *Motivation* and 39 times as *Future*

3.3. Modelling User Preferences

We distinguished between two main ways of capturing user preferences. On the one hand, we considered that a user has a preference (positive rating) for all the papers they have authored. Authored papers encapsulate the user's interest in terms of research areas, topics, methods, etc., and constitute a relevant source of information to build their profile. On the other hand, we considered that a user has a preference for all the papers they have authored as well as cited, since cited papers also encapsulate research that the user considers relevant in relation to their work. By doing so, we explore the use of the citation graph to enrich user profiles.

Figure 2 illustrates these two user preference models. In the left part of the figure, we show a rating matrix R_P relating authors (rows) and papers (columns) where a cell has a value 1 if the corresponding user authored the associated paper, and 0 otherwise. In the right part of the figure, we show a rating matrix R_{PC} where a cell has a value 1 if the user authored or cited the paper, and 0 otherwise. In addition to these rating matrices, we also considered two enriched versions of R_{PC}: (i) R_{PCX}, where X stands for context, and R_{PCXT}, where T stands for type. In case of R_{PCX} for every rating based on citations, it captures the position (section in the paper) and the text around a citation. In the case of R_{PCXT}, for every rating based on citations, it captures the citation type and the text around a citation. In the experiments, the above four matrices were split into training and test sets according to a target time, in particular, the 1st of January 2018. The final dataset splits are captured in Table 3:

Matrix	Details
$R_P^{training}$ R_P^{test}	446 users, 7, 786 items and 9, 348 ratings 446 users, 1, 613 items and 2, 033 ratings
$R_{PC}^{training}$ R_{PC}^{test}	446 users, 7, 786 items and 13, 104 ratings 446 users, 1, 613 items and 2, 126 ratings
$R_{PCX}^{training}$ R_{PCX}^{test}	446 users, 7, 786 items and 13, 104 ratings, and citation context and position information 446 users, 1, 613 items and 2, 126 ratings, and citation context and position information
$R_{PCXT}^{training}$ R_{PCXT}^{test}	446 users, 7, 786 items and 13, 104 ratings, and citation context and type information 446 users, 1, 613 items and 2, 126 ratings, and citation context and type information

Table 3: Capturing user preferences: training and the test sets

4. Evaluation

To assess the created dataset and investigate whether citation knowledge does indeed help enhancing the recommendation of recent scientific publications we present here a series of experiments comparing a novel hybrid recommendation method that explores the previously captured citation knowledge, against various baselines (standard recommendation methods). In this section, we describe the followed evaluation set-up and obtained results.

4.1. Evaluation Setting

Next we present our proposed recommendation method, the baselines considered for comparison, and the metrics used for evaluation. The evaluations were conducted through the RiVal framework (Said and Bellogín, 2014).

4.1.1. Baselines

We experimented with the following baselines: (i) content-based filtering (Lops et al., 2011), (ii) Matrix Factorisation (MF) (Koren et al., 2009), (iii) user-based collaborative filtering, (iv) item-based collaborative filtering (Ricci et al., 2010), (v) Factorisation Machine (FM) (Rendle, 2010), and (vi) the Random method from RankSys framework[14]. In addition, we evaluated ItemRank (Gori and Pucci, 2007), a popular graph-based recommendation method. It is important to note that certain baselines, such as collaborative filtering **cf**, are not able to provide recommendations in the

[13] `https://acl-arc.comp.nus.edu.sg/`

[14] `http://ranksys.org`

proposed scenario where new items (for which no ratings are available) are recommended to users.

Among the evaluated baselines, the best performing one was the content-based **cb** method. Hence, for simplicity, we report results against **cb** in Table 4. Content-based methods recommend items (academic papers) to a user that are 'similar' to those they positively rated, i.e., authored or cited. The similarity between users and items is computed based on profiles built from textual information. User preferences and item attributes correspond to text features; in our case, keywords extracted from the titles of the papers and text surrounding citations. Recommendations are generated by means of user and item similarities in the text feature space. More formally, an item i_n's profile consists of a vector $i_n = w_{n,1}, w_{n,2}, ..., w_{n,L} \in \mathbb{R}^L$ where $w_{n,l}$ denotes the relative relevance (weight) of feature f_l for i_n, and L is the number of existing features. To compute the weights $w_{m,l}$, we used TF-IDF (Jones, 1972).

Similarly, a user u_m's profile is represented as a vector $u_m = w_{m,1}, w_{m,2}, ..., w_{m,L} \in \mathbb{R}^L$, where $w_{m,l}$ denotes the relative relevance (weight) of feature f_l for u_m, and L is computed by aggregating the textual contents of all the papers that have a rating associated to the user, i.e., all the papers for which the user has expressed an interest. The recommendation score of an item i for a target user u is then computed as the cosine similarity $score(u, i) = cos(u, i)$. We refer to this method as **cb**.

The textual features utilised to build user profiles for **cb** varies according to the available citation knowledge (see Section 3.3.). For R_P, a user's profile is built by considering the titles of the papers the user authored. For R_{PC}, a user's profile is built by considering the titles of the papers they authored and cited. Lastly, for R_{PCX} and R_{PCXT}, a user's profile is built by considering the titles of the papers they authored, the tiles of the papers they cited, and the citation contexts.

4.1.2. Proposed Hybrid method

Next, we present a series of hybrid recommendation methods, **hyb**, that jointly exploit the content of the papers and the user-item ratings (see Table 3) to provide personalised recommendations. Hybrid methods (Burke, 2002) aim to mitigate the disadvantages of individual approaches by combining the strengths of various methods, in general, content-based and collaborative filtering. In our case, we aim to mitigate the ineffectiveness of **cf** when recommending the latest scientific publications by combining it with **cb** and exploiting the extracted citation knowledge.

The proposed approach, **hyb**, is based on the item-based **cf** nearest neighbour heuristic[15] where content-based features are used to compute item similarities. In item-based **cf** algorithms (Sarwar et al., 2001), similarities between items are used to estimate scores for a (user, item) pair. In our case, item (paper) profiles are generated based on textual features, which vary with respect to the available citation knowledge (see Section 3.3.): for R_P, item profiles are build from the title of the authored paper; for R_{PC}, item profiles are build from the title of the authored paper and the titles of the cited papers; and for R_{PCX} and R_{PCXT}, item profiles are build from the title of the authored paper, the tiles of the cited papers, and the citations context (i.e., texts around citations within the paper). We formulate our item-based hybrid method in Equation (1):

$$\hat{r}_{u,i} = \frac{\sum_{i \epsilon N(i')} Sim(i,i').r_{u,i'}}{\sum_{i \epsilon N(i')} |Sim(i,i')|} \tag{1}$$

where $\hat{r}_{u,i}$ is the preference score to be predicted for the target user u and item i, $Sim(i, i')$ is the similarity between the interacted item i' and an item i from the neighbourhood $N(i')$ of the item i'. Cosine similarity is used to measure the similarity between items. Finally, $r_{u,i'}$ is the preference (rating) given by user u to an item i'. We also use different sizes of neighbours, specifically 5, 10, 15 and 20.

To investigate the relevance of the different types of citations, we further modified our hybrid method by incorporating a weight, $w_{u,i'}$ in the eq. (1) heuristic, that reflects the strength of an item i' for a user u based on the different citation types, named as **hybType**. Equation (2) represent **hybType** hybrid method. We formulate this method **hybType** in Equations (2) and (3).

$$\hat{r}_{u,i} = \frac{\sum_{i \epsilon N(i')} Sim(i,i').r_{u,i'}.w_{u,i'}}{\sum_{i \epsilon N(i')} |Sim(i,i')|} \tag{2}$$

where the strength (weight) $w_{u,i'}$ is computed by considering all the instances where i' is cited by u; Note that an item i' may be cited by u in several publications, and with different citation types. Then, the weight is normalised by the total number of instances. Formally, the strength $w_{u,i'}$ of item i' for user u is calculated as:

$$w_{u,i'} = \frac{\sum_{j=1}^{n}(w_{j_{bkg}} + w_{j_{com}} + w_{j_{mot}} + w_{j_{use}} + w_{j_{ext}} + w_{j_{fut}})}{n_{u,i'}} \tag{3}$$

where $n_{u,i'}$ is the number of times i' is cited by u in their papers, and $w_{j_{bkg}}$, $w_{j_{com}}$, $w_{j_{mot}}$,$w_{j_{use}}$,$w_{j_{ext}}$ and $w_{j_{fut}}$ reflect the weight when i' is cited as background, compareOrContrast, motivation, use, extension, or future respectively

4.2. Evaluation Results

We summarise the results obtained from our experiment in Table 4. The foremost conclusion we draw from the experiment is that the incorporation of citation knowledge helps improving the performance of recommendation methods. Between the rating matrices R_P and R_{PC} (where R_{PC} reduces the sparsity because of the additional citation information added to it), **cb** and our **hyb** method show improvements with all the evaluation metrics at the cut-off points of 5 and 10. The improvement continues in all the sizes of neighbourhoods, i.e., 5, 10, 15 and 20. This shows that the decrease on the sparsity of the matrix, thanks to the incorporation of user preferences based on the citation graph, allows finding valuable relationships between items and users that are productively exploited.

When incorporating citation knowledge in the form of citation context, captured in the R_{PCX} matrix, we notice further increment in the performance of the hybrid methods, but a decrease in the performance of the **cb** method. The saturation in the textual features between R_{PC} and R_{PCX} may be the cause of this drop in performance.

[15] We also tested the user-based **cf** heuristic, but discarded it due to its non competitive performance results.

matrix	method	map@5	map@10	nDCG@5	nDCG@10	p@5	p@10	r@5	r@10	F1@5	F1@10
R_P	cb	0.044	0.052	0.08	0.091	0.053	0.043	0.065	0.105	0.058	0.061
	hyb5	0.039	0.045	0.073	0.08	0.05	0.037	0.063	0.094	0.056	0.053
	hyb10	0.039	0.045	0.073	0.081	0.05	0.039	0.063	0.099	0.056	0.056
	hyb15	0.039	0.045	0.072	0.081	0.049	0.039	0.061	0.099	0.054	0.056
	hyb20	0.037	0.043	0.069	0.079	0.048	0.039	0.057	0.096	0.052	0.055
R_{PC}	cb	0.047	0.052	0.083	0.095	0.056	0.046	0.071	0.112	0.063	0.065
	hyb5	0.042	0.047	0.075	0.081	0.05	0.035	0.065	0.095	0.057	0.051
	hyb10	0.046	0.052	0.082	0.09	0.055	0.041	0.074	0.106	0.063	0.059
	hyb15	0.048	0.053	0.085	0.092	0.057	0.042	0.077	0.108	0.066	0.06
	hyb20	0.047	0.052	0.083	0.09	0.056	0.041	0.077	0.108	0.065	0.059
R_{PCX}	cb	0.032	0.036	0.054	0.059	0.036	0.027	0.048	0.07	0.041	0.039
	hyb5	0.043	0.048	0.08	0.085	0.054	0.039	0.07	0.098	0.061	0.056
	hyb10	0.049	0.055	0.086	0.095	0.058	0.043	0.074	0.112	0.065	0.062
	hyb15	0.051	0.057	0.089	0.096	0.059	0.043	0.074	0.108	0.066	0.062
	hyb20	0.052	0.058	0.088	0.097	0.057	0.044	0.073	0.108	0.064	0.063
R_{PCXT}	hybType5 (0, 0, 0, 0, 0, 1)	0.066	0.078	0.099	**0.126**	**0.067**	**0.062**	**0.111**	**0.172**	**0.084**	**0.091**
	hybType5 (0, 0, 0, 0, 1, 0)	0.066	**0.079**	**0.101**	**0.126**	**0.067**	**0.062**	0.11	0.168	0.083	**0.091**
	hybType5 (0, 0, 0, 1, 0, 0)	0.064	0.075	0.096	0.121	0.06	0.058	0.106	0.161	0.077	0.085
	hybType5 (0, 0, 1, 0, 0, 0)	**0.067**	**0.079**	**0.101**	**0.126**	**0.067**	0.061	**0.111**	0.167	**0.084**	0.089
	hybType5 (0, 1, 0, 0, 0, 0)	0.046	0.055	0.069	0.089	0.04	0.039	0.074	0.125	0.052	0.059
	hybType5 (1, 0, 0, 0, 0, 0)	0.041	0.048	0.063	0.078	0.037	0.034	0.063	0.104	0.047	0.051

Table 4: Experiment results of the baselines and proposed hybrid recommendation methods. A gray scale is used to highlight better (dark gray) and worst (white) values for each ranking metric. For every metric, the best values are highlighted in bold.

Exploiting citation knowledge in the form of citation types (**hypType**), captured in the R_{PCXT} matrix, outperformed all other methods (i.e., **hyb** and **cb**) for all the evaluation metrics. This shows that citation type is a useful feature for recommending recent papers to users. In particular, the performance of our **hypType** method is higher when the citation belongs to a **Future** or **Extension** categories. This corroborates the intuition that citations that are related to future research and to work that is being extended or enhanced, represent relevant pointers for new directions in a research field. As can be seen in table 4, precision, recall and nDCG@5 values are higher when citations belong to *Future* category, while nDCG, precision and recall@5 values are higher when the citation type is *Extension*. In this context, while higher precision and recall values show the methods were able to find relevant and new items, the higher nDCG value shows that novel documents are appearing earlier in the recommendation lists. The **Motivation** type has also positive impact on the recommendation. Citations that motivate the work are helpful on reflecting problems that may be targeted not only in the present, but also in future scientific publications.

5. Discussion and Future Work

This work addresses the problem of recommending recent scientific publications to users, and investigated the extraction and exploitation of citation knowledge for such purpose. In doing so, we have generated a novel dataset capturing (i) users' preferences (in the form of publication and citation history) and, (ii) different notions of citation knowledge including the citation graph, citation positions, citation contexts, and citation types. To the best of our knowledge, this is the first available dataset capturing the above information, and hence a unique resource to enable the investigation of an important real-word problem, the personalised recommendation of recent scientific publications.

A particular interesting form of citation knowledge captured in this dataset is citation types, including *background, compare or contrast, use, extension, motivation* and *future*. Citation types are not only useful to enhance paper recommendation (as shown in our evaluation), but can also enable further refinement of the recommendation process based on the users' intention. For instance, users with an interest on how to use or apply certain algorithms or techniques, may receive recommendations for papers cited under the type 'use' or 'extension'. Refining recommendations based on the users' intent, in addition to the users' publication and citation history, is one of our future lines of work. We also aim to develop recommendation methods based on the combination of different notions of citation knowledge, particularly citation context, type and position, since these three elements capture fine-grained patterns of the intent with which papers are cited.

Despite the timeliness and potential of this dataset we also acknowledge several limitations:

Data sampling: This dataset is representative of the area of RS. Complementing it with papers from different fields will help to assess whether the obtained findings are specific to the RS research field or are representative of other areas.

Data annotation: Although we tested that the classifier provided by (Jurgens et al., 2018b) achieved comparable results when identifying citation types for our collected papers, the average obtained accuracy is 69%, indicating the presence of noise in the classification process. In addition, the PDF parser used as part of this classifier could only extract citation information for 61% of the collected papers, reducing the completeness of the data. Providing more effective citation type classification methods is an interesting research question for the analysis, search and recommendation of scientific publications.

Data filtering: In order to capture users' preferences we discarded from our dataset all authors for which we obtained less than four publications. Our dataset therefore does not capture the scenario of recommending recent papers to users for which no preferences have been gathered.

Despite the above limitations, our work provides a unique resource that will help the RS community to further investigate the problem of recommending recent scientific publications to users. We also hope that the use of this dataset will stimulate discussions across related disciplines (scientometrics, information retrieval, etc.) on the usefulness of citation knowledge to effectively target the problem of information overload in the scientific world.

6. References

Abu-Jbara, A., Ezra, J., and Radev, D. R. (2013). Purpose and polarity of citation: Towards nlp-based bibliometrics. In *HLT-NAACL*.

(2019). Acl anthology. `https://acl-arc.comp.nus.edu.sg/`.

(2019). Aminer. `https://www.aminer.cn/aminer_data`.

(2019). Arxiv. `https://arxiv.org/help/bulk_data`.

Beel, J., Gipp, B., Langer, S., and Breitinger, C. (2016). Paper recommender systems: A literature survey. *International Journal on Digital Libraries*, 17(4):305–338.

Bethard, S. and Jurafsky, D. (2010). Who should i cite: Learning literature search models from citation behavior. In *19th ACM International Conference on Information and Knowledge Management*, pages 609–618.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

Cai, T., Cheng, H., Luo, J., and Zhou, S. (2016). An efficient and simple graph model for scientific article cold start recommendation. In *International Conference on Conceptual Modeling*, pages 248–259. Springer.

Chakraborty, T., Krishna, A., Singh, M., Ganguly, N., Goyal, P., and Mukherjee, A. (2016). Ferosa: A faceted recommendation system for scientific articles. In *Proceedings, Part II, of the 20th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume 9652*, PAKDD 2016, pages 528–541, Berlin, Heidelberg. Springer-Verlag.

(2019). Citeulike. `https://old.datahub.io/dataset/citeulike`.

Cohan, A., Ammar, W., van Zuylen, M., and Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. *CoRR*, abs/1904.01608.

(2019). Core. `https://core.ac.uk/services/dataset/`.

Ekstrand, M. D., Kannan, P., Stemper, J. A., Butler, J. T., Konstan, J. A., and Riedl, J. T. (2010). Automatically building research reading lists. In *4th ACM Conference on Recommender Systems*, pages 159–166.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Garzone, M. and Mercer, R. E. (2000). Towards an automated citation classifier. In *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, AI '00, pages 337–346, Berlin, Heidelberg. Springer-Verlag.

Gori, M. and Pucci, A. (2007). Itemrank: A random-walk based scoring algorithm for recommender engines. In *20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2766–2771, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ha, J., Kwon, S.-H., Kim, S.-W., and Lee, D. (2014). Recommendation of newly published research papers using belief propagation. In *2014 Conference on Research in Adaptive and Convergent Systems*, pages 77–81.

Hassan, S.-U., Akram, A., and Haddawy, P. (2017). Identifying important citations using contextual information from full text. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '17, pages 41–48, Piscataway, NJ, USA. IEEE Press.

Hassan, S.-U., Imran, M., Iqbal, S., Aljohani, N. R., and Nawaz, R. (2018). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics*, 117:1645–1662.

He, Q., Pei, J., Daniel, K., Prasenjit, M., and Giles, L. (2010). Context-aware citation recommendation. pages 421–430.

Hristakeva, M., Kershaw, D., Rossetti, M., Knoth, P., Pettit, B., Vargas, S., and Jack, K. (2017). Building recommender systems for scholarly information. In *1st Workshop on Scholarly Web Mining*, pages 25–32.

Huang, W., Wu, Z., Liang, C., Mitra, P., and Giles, C. L. (2015). A neural probabilistic model for context based citation recommendation. In *29th AAAI Conference on Artificial Intelligence*, pages 2404–2410.

Jack, K., Hammerton, J., Harvey, D., Hoyt, J. J., Reichelt, J., and Henning, V. (2010). Mendeleys reply to the datatel challenge. *Procedia Computer Science*, 1(2):1–3.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., and Jurafsky, D. (2018a). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., and Jurafsky, D. (2018b). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Khadka, A. and Knoth, P. (2018). Using citation-context to reduce topic drifting on pure citation-based recommendation. In *12th ACM Conference on Recommender Systems*, pages 362–366.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August.

Liang, Y., Li, Q., and Qian, T. (2011a). Finding relevant papers based on citation relations. In *12th International Conference on Web-Age Information Management*, pages 403–414.

Liang, Y., Li, Q., and Qian, T. (2011b). Finding relevant papers based on citation relations. In *Proceedings of the 12th International Conference on Web-age Information Management*, WAIM'11, pages 403–414, Berlin, Heidelberg. Springer-Verlag.

Lops, P., de Gemmis, M., and Semeraro, G., (2011). *Content-based Recommender Systems: State of the Art and Trends*, pages 73–105. Springer US, Boston, MA.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., Konstan, J. A., and Riedl, J. (2002). On the recommending of citations for research papers. In *2002 ACM Conference on Computer Supported Cooperative Work*, pages 116–125.

Meyers, A. (2013). Contrasting and corroborating citations in journal articles. In *RANLP*.

Middleton, S. E., De Roure, D. C., and Shadbolt, N. R. (2001). Capturing knowledge of user preferences: Ontologies in recommender systems. In *1st International Conference on Knowledge Capture*, pages 100–107.

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

Nanba, H. and Okumura, M. (1999). Towards multi-paper summarization reference information. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, pages 926–931, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Nanba, H., Kando, N., and Okumura, M. (2000). Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online*, 11(1):117–134.

(2019). Open academic graph. https://www.openacademic.ai/oag/.

(2019). Open citations. https://opencitations.net/corpus.

Rendle, S. (2010). Factorization machines. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 995–1000, Washington, DC, USA. IEEE Computer Society.

Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B. (2010). *Recommender Systems Handbook*. Springer-Verlag, Berlin, Heidelberg, 1st edition.

Said, A. and Bellogín, A. (2014). Rival: A toolkit to foster reproducibility in recommender system evaluation. In *8th ACM Conference on Recommender Systems*, pages 371–372.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *10th International Conference on World Wide Web*, pages 285–295.

Shimbo, M., Takahiko, I., and Matsumoto, Y. (2007). Evaluation of kernel-based link analysis measures on research paper recommendation. In *7th ACM International Conference on Digital Libraries*, pages 354–355.

(2019a). Spd 1. https://acl-arc.comp.nus.edu.sg/.

(2019b). Spd 2. ttps://www.comp.nus.edu.sg/~sugiyama/Dataset2.html.

Strohman, T., Croft, W. B., and Jensen, D. (2007). Recommending citations for academic papers. In *30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 705–706.

Sugiyama, K. and Kan, M.-Y. (2010). Scholarly paper recommendation via user's recent research interests. In *10th Joint Conference on Digital Libraries*, pages 29–38.

Sugiyama, K. and Kan, M.-Y. (2013). Exploiting potential citation papers in scholarly paper recommendation. In *13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 153–162.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006a). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL '06, pages 80–87, Stroudsburg, PA, USA. Association for Computational Linguistics.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006b). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 103–110, Stroudsburg, PA, USA. Association for Computational Linguistics.

Teufel, S. (1999). Argumentative zoning information extraction from scientific text.

Torres, R., McNee, S. M., Abel, M., Konstan, J. A., and Riedl, J. (2004). Enhancing digital libraries with techlens+. In *4th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 228–236.

Valenzuela, M., Ha, V. A., and Etzioni, O. (2015). Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data*.

Wang, C. and Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM.