Social Network Analysis and Mining (2023) 13:81
https://doi.org/10.1007/s13278-023-01085-w

ORIGINAL ARTICLE

ACE 2.0: A Comprehensive tool for automatic extraction, analysis, and digital profiling of the researchers in Scientific Communities

Syed Tahseen Raza Rizvi[1,2] · **Sheraz Ahmed**[1] · **Andreas Dengel**[1,2]

Received: 24 January 2023 / Accepted: 14 April 2023 / Published online: 5 May 2023

Abstract
In the current digital era, it is remarkably convenient for researchers to share and collaborate on novel scientific ideas. Scientists aim to accomplish these endeavors through closely knitted scientific communities, depending on the domain. Technological advancements and their evolution overtime gave rise to a boom in the emergence of research communities with unique topics and focuses. Due to the enormous number and vastness of scientific communities, it is an intractable task to analyze scientific communities and administer them from a quantitative and qualitative perspective. Existing tools provide a limited and shallow glance into a scientific community. In this paper, we present a comprehensive system for the analysis of scientific communities called ACE 2.0 (Academic Community Explorer 2.0) which employs state-of-the-art models to automatically, efficiently, and smartly extract, and analyze bibliographic data. Moreover, it provides a range of insights from individual researchers to interactions between communities. These insights include different community-level aspects like collaboration patterns, citation patterns, influential persons with different roles, contributions from geographical locations, topics evolution, and many other fine-grained aspects within each scientific community. Our system considers scholarly publications as a primary source of information. However, it also employs several external resources to collect as much data as possible to correctly identify individual researchers and their contributions. Using all the collected data, ACE 2.0 performs an analysis of scientific communities and automatically performs detailed digital profiling of individual researchers. This analysis identifies trends in their citation, collaboration, contributions, popularity, and role in the community. Additionally, ACE 2.0 introduces a new Semantic index for researchers that takes into account both quantitative and qualitative aspects of the citations received by a researcher and quantifies their influence in the community. To conclude, ACE 2.0 enables us to analyze and oversee the scientific communities using trends and information gathered from different sources encompassing multiple aspects. Therefore, this work motivates us to discover endless new perspectives and opens it up to a wide range of applications in other domains. The demo of ACE 2.0 visualization engine is available at https://ace.opendfki.de/.

Keywords Semantic index · Scientific community · Social network analysis · Digital profiling · Community influence analysis

1 Introduction

Scientific research plays an important role in the development of a society. Researchers have always been fascinated by the scientific method and are compelled to ask questions themselves about the phenomena all around. Which led them to perform research and seek answers to their desired questions. Researchers present their research findings in front of a group of like-minded researchers called a scientific community, who also have similar interests. A scientific community can rather be small or large depending on the involvement and interest of the researchers. During this digital era, fast-paced development has resulted in the rapid evolution of

✉ Syed Tahseen Raza Rizvi
syed_tahseen_raza.rizvi@dfki.de

Sheraz Ahmed
sheraz.ahmed@dfki.de

Andreas Dengel
andreas.dengel@dfki.de

[1] Smart Data & Knowledge Services, DFKI, Kaiserslautern, Germany

[2] Department of Computer Science, RPTU Kaiserslautern-Landau, Kaiserslautern, Germany

the scientific community. With the ever-increasing number of researchers with diverse interests joining large scientific communities, the number of topics is also rapidly increasing within those communities. A scientific community has to promptly adapt to the popular demand and interests of society.

In order to properly adapt to the interests of society, a scientific community needs to have a thorough understanding and awareness of the public interests. For that purpose, an intelligent system is imminent for the analysis of the scientific communities which can efficiently handle and analyze the community data and identify trends that can help make smart decisions for the future of a scientific community. These useful trends are classified into different levels depending on the target and decisions. For example, to decide for the whole community, trends found on a community level will be considered. The analysis of scientific communities also involves identifying small tightly packed sub-groups where all involved researchers only cite one another from the same sub-group and identifying individuals who use self-citations to manipulate their author indexes.

In this paper, we propose a comprehensive system for the analysis of the scientific communities called ACE 2.0 (Academic Community Explorer 2.0) which takes scientific publications as its baseline information source and also collects related information about authors and publications from several external resources for digital profiling of the authors and their publications respectively. ACE 2.0 later performs an analysis of all the collected and extracted data in the context of a scientific community. All individual entities, i.e., authors are identified and separately profiled from each publication. Once the data analysis is completed, all the trends and data are visualized using a visualization engine. We visualize the whole scientific community in the form of a network where each individual is interacting with the other by citing the papers of the others. Our proposed system consists of several modules where each module is responsible for a specific task, i.e., bibliographic reference detection, Keyword detection, topic modeling, etc. For each model, we employed state-of-the-art models for the respective tasks. Additionally, this paper introduces a novel Semantic index that quantifies the contribution and influence of an author in a scientific community. Instead of relying merely on the raw citation count like traditional indexes, the Semantic index considers both qualitative and quantitative aspects of the citations to quantify the influence.

The contributions of this publication are as follows:

- We present a comprehensive system for the analysis of scientific communities which automatically extracts, consolidates, analyzes data and visualizes the extracted trends using a visualization engine.
- We present a tool for automatic digital profiling of the authors and publications from a given scientific community.
- We propose a novel Semantic index that takes both quantitative and qualitative aspects of a citation for determining the influence of a researcher in the scientific community.

The rest of the paper is structured as follows: Section 2 discusses the literature related to several modules of our proposed system. Section 3 discusses the details of the proposed system where each system module is described in detail along with their respective evaluations. It also discusses the feature highlights of our system where features from different levels are elaborated. And lastly, Sect. 4 includes the overall discussion of the proposed system and presents the concluding remarks for this paper.

2 Related works

There are several modules integrated into our proposed pipeline. Each module is responsible for performing a specific task, i.e., Bibliographic Reference Detection, Citation Sentiment Analysis, Keywords Detection, etc. In this section, we will discuss relevant literature for each of the modules of our proposed pipeline.

2.1 Bibliographic reference detection

Reference detection from scientific publications is a popular task in Scientometrics and is mainly an area of interest for library cataloging. Two kinds of solutions approach the task of bibliographic reference detection from different perspectives. The first one is text-based solutions, while the other is layout-based solutions.

Text-based reference detection is the most common solution for reference detection. Several approaches have been proposed that use/consider textual features for identifying references from a given publication. The popularity of text-based approaches is due to the existence of a finite number of referencing styles adopted by the scientific community. The most primitive approaches employed well-defined heuristics[1,2] to identify components of a reference string, i.e., author name, publication title, etc. These components then contribute toward the identification of a reference occurring in a text corpus. On the similar lines, *RefParse*Sautter and Böhm (2012) and *BibPro* (Chen et al. 2012) proposed using heuristics on the component similarity between reference

[1] Citaion Parser. https://github.com/manishbisht/Citation-Parser.

[2] AnyStyle. https://anystyle.io/.

strings to identify referencing style used in that publication. In addition to standalone tools, libraries[3] in Perl were presented to parse and extract reference string data from a given text corpus.

PDFSSA4MET[4] suggested converting the PDF of a given publication to intermediate form like XML and then analyzing the structure of the XML for identification of the section containing references. Some approaches like (Ahmed and Afzal 2020) employed a wide range of features like lexical properties, font type, location, neighbors distance, etc. These features are then used to identify and extract reference strings and their metadata.

A novel method of calculating conditional random fields was proposed by Lafferty et al. (2001). Using CRF, sequence data can be systematically labeled by a probabilistic approach. A reference string is identified by identifying the specific parts of the string, such as the authors, the title of the publication, the year, the name of the conference or journal, etc. Labeling such components facilitates the identification of a reference string based on the components that are tagged.

A CRF-based system for extracting and mining bibliographic metadata from references in born-digital PDF scientific articles was proposed by Tkaczyk et al. (2015) as Content ExtRactor and MINEr (*CERMINE*). There are also different tools[5,6] (Matsuoka et al. 2016) which are based on CRF for identifying and extracting metadata from reference strings. In a work published by Councill et al. (2008), a CRF-based package called *ParsCit* was presented for its application in relation to the reference metadata tagging problem. It is claimed by Councill et al. (2008) that *ParsCit* is among the best known and most widely used open-source systems based on heuristics and CRF for reference detection, string parsing, and metadata tagging. As part of the project, Tkaczyk et al. (2018) also proposed a recommendation system based on reference metadata, which integrated 10 of the most popular open-source citation parser tools into one system. A combination of simple heuristic-based and machine learning-based tools was selected as solutions.

To identify references, the literature discussed so far relies exclusively on textual features. The text-based approach overlooks an imperative aspect of layout due to its inability to take advantage of layout features. The potential for detecting bibliographic references using layout information has been explored in very few approaches.

Using layout information, Bhardwaj et al. (2017) detected references in scanned documents. In order to accomplish this, a Fully Convolutional Neural Network (FCN) was used (Long et al. 2015) to segment the references and then post-processed in order to identify individual references. A layout-based citation detection method was incorporated into Lauscher et al. (2018) to build an open database of citations for libraries for indexing purposes. To detect bibliographic references in scientific publications, Rizvi et al. (2019) evaluated four state-of-the-art object detection models based on layout information.

2.2 Sentiment classification

Several publications have been published to address the problem of sentiment classification because of its wide range of applications. Using sentiment-specific word embeddings, Tang et al. (2014) classified tweets based on their sentiment. As a result, highly specialized embeddings can be used to improve sentiment classification performance. Thongtan and Phienthrakul (2019) employed document embeddings trained with cosine similarity to perform sentiment classification on a movie review dataset. A sentiment classifier based on an ensemble of CNN and LSTM models was proposed by Cliche (2017) which was fine-tuned on a large database of unlabeled tweets.

In recent years, BERT (Devlin et al. 2018) has gained widespread recognition for its ability to perform a wide variety of Natural Language Processing (NLP) tasks. Several large volumes of unlabeled data were used in training the BERT model. In order to improve the performance of sentiment analysis, recent literature has utilized the BERT model. By combining pre-processing, attention modules, and structural features with a pre-trained BERT model, the authors maximize the performance of the model. In Munikar et al. (2019), Zhou et al. (2016), Xied et al. (2019), the authors adapt a pre-trained BERT model for sentiment classification by combining transfer learning with pre-processing, attention modules, and structural features.

The majority of the literature has been devoted to classifications of sentiments in tweets and movie reviews. The text in scientific publications has a more formal tone in comparison with the text in reviews. Thus, citation sentiment classification differs considerably from review sentiment classification. Esuli and Sebastiani (2006) have compared sentiment classifications to opinion mining and subjectivity mining. Citations can also be subject to subjectivity due to author preferences and writing styles. This is due to the fact that an author has the option of intentionally making a citation sound positive or negative. Athar (2011) has carried out experiments to simulate experimental outcomes that involve a wide range of features for sentiment classification in scientific papers, including science lexicon, contextual polarity,

[3] Biblio. https://metacpan.org/release/MJEWELL/Biblio-Citation-Parser-1.10.

[4] Kunnas E PDFSSA4MET. https://github.com/eliask/pdfssa4met.

[5] Goldberg M free_cite. https://github.com/miriam/free_cite.

[6] Science Parse. https://github.com/allenai/science-parse.

dependencies, negation, sentence splitting, and word-level features. As a result of the use of textual features such as n-grams, sentiment lexicons, and structure information, Xu et al. (2015) conducted a sentiment analysis of citations in clinical trial papers. Considering the lack of datasets suitable for scientific sentiment classification, sentiment classification is imperative in the field of scientific citation analysis. It is the result of the shallow definition of sentiment in this domain that has led to this situation. It is much more challenging to determine a sentiment in an analytical and objective text in contrast to finding a sentiment in high subjective texts, such as Twitter data. Recently, Mercier et al. (2021, 2022) released a clean dataset for citation sentiment analysis and proposed the use of transformers to achieve state-of-the-art performance for citation sentiment classification on different datasets.

2.3 Keywords detection

It is estimated that approximately millions of scientific articles are published in journals around the world every year (Ware and Mabe 2015). A manual search to link large volumes of scholarly publications with appropriate representative keywords would certainly prove to be impractical. Consequently, it is anticipated that a system will be developed that will be capable of automatically analyzing and indexing scientific articles. It is well known that automated keyword detection has been studied extensively; however, most approaches concentrate on social media such as tweets (Beliga 2014; Biswas et al. 2018; Boudin 2018; Carpena et al. 2009; Carretero-Campos et al. 2013; Duari and Bhatnagar 2019; Florescu and Caragea 2017; Mahata et al. 2018; Nikolentzos et al. 2017; Ohsawa et al. 1998; Pay and Lucci 2017; Rabby et al. 2018).

In order to detect keywords in a text, an undirected graph $G = (N, E)$ is commonly used, in which the nodes represent the individual terms within the text and the edges represent their relationship. In general, the term co-occurrence is the most common type of relationship, which adds an edge to the graph between nodes n_1 and n_2 if both corresponding terms appear within the same sliding window. Depending on the approach selected, the recommended window size is often between 2 and 10 (Litvak et al. 2011; Mihalcea and Tarau 2004; Rousseau and Vazirgiannis 2015). A study by Duari and Bhatnagar (2019) indicates that the size of the window w can have a significant impact on the properties of the resulting graph. When w increases, the density increases as well, resulting in a decrease in the average path length between any two nodes.

It is assumed that the words appearing closer together are related (Rousseau and Vazirgiannis 2015). It is possible to change the sliding window in several ways, for instance, allowing it to slide over individual sentences rather than the entire text, stopping at certain punctuation marks (Litvak et al. 2011). A novel concept referred to as Connectivity Aware Graph (CAG) has been proposed by Duari and Bhatnagar (2019). They use a dynamic window size instead of a fixed window size that spans over two sentences irrespective of their length. In their view, consecutive sentences are considered to be related. They demonstrated that using CAGs instead of graphs constructed using traditional window sizes generally enhances the performance of approaches.

2.4 Authors index

A study by Bollen et al. (2009) discusses science as a gift economy. An author's value can be defined as the extent to which he contributes to knowledge and the extent to which he influences the ideas of other scientists. As Hirsch (2005) pointed out, it is needed to have the possibility to quantify this kind of value, among others, for recruitment decisions of universities and the award of grants, especially in a world of limited resources. The increasing costs of research and the shortage of available economic resources lead to a high and increasing interest in scientific author assessment (Costas and Bordons 2008). Additionally, the usefulness of evaluating scientific author impact and author ranking when doing research, in general, should not be underestimated. It offers the possibility for every researcher to easily spot authors heavily contributing to a research field and to discover their publications which might be worth reading when executing research in a specific field. For achieving such an author impact assessment, different indicators are commonly used by many author assessment approaches. On the one hand, there are production indicators which are, for example, the total number of published papers and on the other hand, there are impact indicators which are usually based on the citations received by an author (Alonso et al. 2010). Hirsch consequently states that the large amount of useful information, which is given by the publication record of an individual, can be evaluated with various criteria by several researchers (Hirsch 2005). This leads to the emergence of different author assessment approaches. Each of these approaches can be considered as an attempt to highlight a specific aspect of an author's publication record that might be of interest when evaluating the author's importance and contribution to science (Cai et al. 2019). There are huge debates about which of them are the best for assessing the importance and contribution of a scientific author. However, it is widely accepted as a good approach to simply use multiple quantitative measures to support an expert judgment for improving objectivity and fairness in the evaluation process.

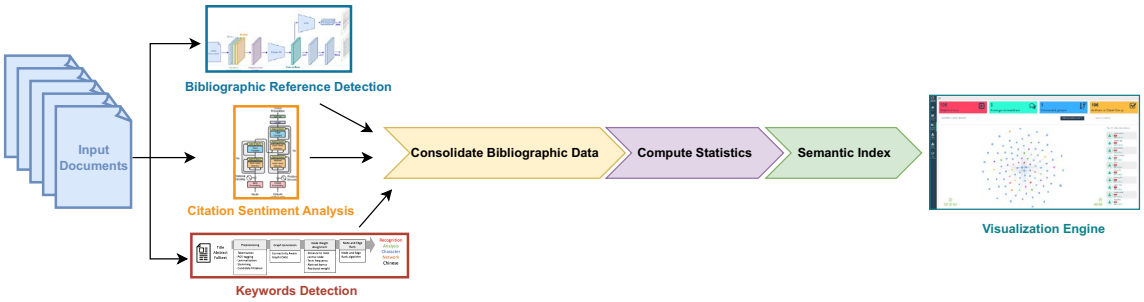


Fig. 1 Overview of ACE 2.0 system

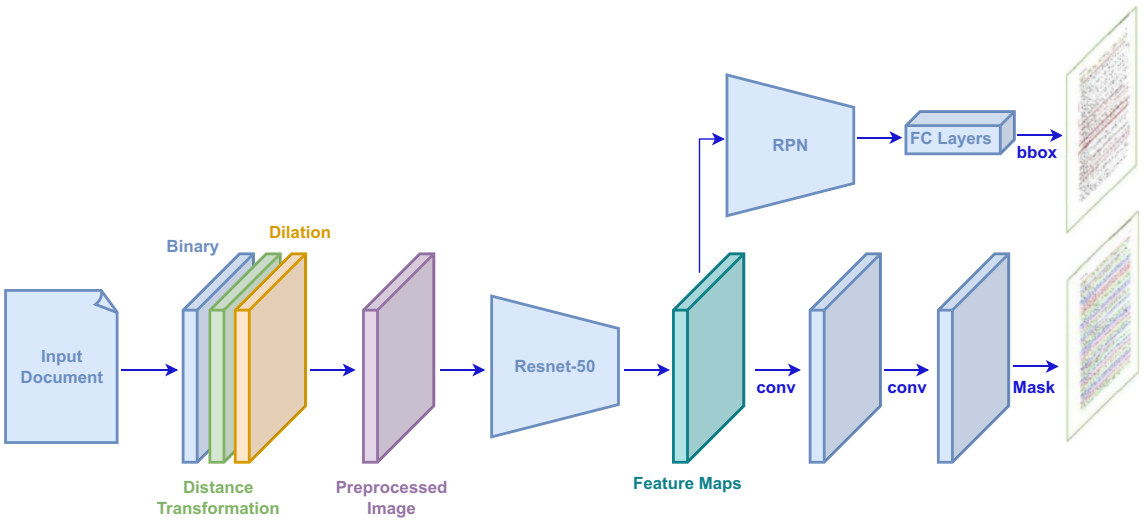


Fig. 2 DeepBiRD (Rizvi et al. 2020) pipeline overview

3 Proposed system

In this section, we will discuss the details of our proposed system called ACE 2.0. Figure 1 shows the overview of the proposed system. It consists of several modules where each module specializes in a specific task. Each of the modules will be discussed in detail in the following subsections:

3.1 Bibliographic reference extraction

This is the first module of our proposed system and is responsible for the extraction of bibliographic references from given scientific publications. We employed our previously proposed state-of-the-art model called DeepBiRD (Rizvi et al. 2020) for the task of bibliographic reference detection. Details of the selected model are as follows:

3.1.1 Methodology

DeepBiRD (Rizvi et al. 2020) is a layout-based reference detection approach. Typical reference detection approaches use textual features for bibliographic reference detection. However, DeepBiRD makes use of layout features to identify references in a given scanned document image. For this purpose, DeepBiRD compiles a hybrid representation using the input image which is later provided to the network to identify bibliographic references from the given image. Figure 2 shows the overview of the DeepBiRD pipeline.

The first step in the DeepBiRD pipeline is to prepare a hybrid representation from the input image. The idea behind compiling a hybrid representation is to make decisive layout features like line and word spacing more prominent. Such layout features play an important role in identifying bibliographic references. Hybrid representation consists of three different components.

The first component of the hybrid representation is the Distance transform in which we consider the distance of

Table 1 Evaluation results of DeepBiRD (Rizvi et al. 2020) on two datasets compared with other approaches

Dataset	Model	AP50 (%)
BibX (Bhardwaj et al. 2018)	DeepBiRD (Rizvi et al. 2020)	97.59
	Faster R-CNN (Ren et al. 2015; Rizvi et al. 2019)	84.50
	Deformable FPN (Dai et al. 2017; Lin et al. 2017; Rizvi et al. 2019)	84.17
	Deformable Faster R-CNN (Dai et al. 2017; Ren et al. 2015; Rizvi et al. 2019)	82.83
	Deformable RFCN (Dai et al. 2017, 2016; Rizvi et al. 2019)	82.37
	DeepBIBX Bhardwaj et al. (2017)	54.22
BibLy (Erhard et al. 2019)	DeepBiRD (Rizvi et al. 2020)	98.56
	DeepBIBX (Bhardwaj et al. 2017)	56.77

each pixel from its nearest input foreground pixel. Using distance transformation of a given image we can introduce additional information highlighting spaces between words, lines, and characters. In our experimental setup, we initially inverted the input image, followed by binarization using OTSU thresholding and inversion of the resultant image. Later, the distance transformation of this image is performed using Euclidean distance as the distance measure with a mask of 3×3.

The second component of the hybrid representation is a dilated version of the input image. The process of dilation highlights the characters and their vicinity. In order to get a dilated image, we performed OTSU thresholding on the input image with a horizontal kernel size of 1×5. The purpose of using a horizontal kernel is to highlight the characters and their vicinity in each line.

The third component of the hybrid representation is the binarized input image. Once all three components of the hybrid representation are ready, we use these components to fill three channels of the image characterizing a hybrid representation. The order of the components in the channels of the hybrid representation is set as transform image, binarized image, and dilated image. Text and their surroundings are represented in red color however the spaces between characters and lines are represented in blue color. The resultant hybrid representation is categorically better than the original input image as it has more visual features than its original counterpart.

The second step in the DeepBiRD pipeline involves detecting bibliographic references from a given hybrid representation. For this purpose, DeepBiRD is equipped with a state-of-the-art object detection model called Mask-RCNN which differentiates between detected reference instances in addition to identifying bibliographic references. Mask-RCNN generates a mask and a bounding box for individual detected reference. For training the network, we followed all the settings mentioned in the original paper (Rizvi et al. 2020).

3.1.2 Evaluation and discussion

The performance of *DeepBiRD* (Rizvi et al. 2020) was evaluated on two datasets named *BibX* and *BibLy* dataset. Both datasets are publicly available and are annotated for the task of bibliographic reference detection from scanned scientific publications. The performance of *DeepBiRD* model is compared with other existing object detection approaches in Table 1. It can be observed that *DeepBiRD* model outperforms other approaches by a significant margin on both datasets. In *BibX* dataset, our selected model outperforms the other approaches by an average of 13.38%. Similarly, on *BibLy*, our selected model also outperforms the previous state-of-the-art on *BibX* dataset by a huge margin. The reason behind the superior performance of *DeepBiRD* model is that it uses hybrid representations to assist the detection model in identifying bibliographic references using additional highlighted information. Additionally, the inherent property of the selected model which employs ROIAlign operation that also plays an important role in detecting references to a very fine detailed level.

3.2 Sentiment analysis

In this section, we will discuss the component which is responsible for the sentiment analysis of the citations. We will briefly discuss the architectural details of the selected model and its performance evaluation.

3.2.1 Methodology

For the task of sentiment analysis of citations, we employed another approach we proposed earlier which is an XLNet-based approach called *ImpactCite*. XLNet (Yang et al. 2019) is an auto-regressive language model which captures the context in both directions using a bi-directional attention mechanism. Such a mechanism helps comprehend the context of a sentence from right to left and left to right and is very effective in longer sentences. The most highlighting feature of XLNet is that in addition to the Transformer-XL (Dai

Table 2 Evaluation results of ImpactCite (Mercier et al. 2021) on CSC-C dataset (Mercier et al. 2021) compared with other approaches

Topography	Architecture	Class-based Accuracy			Macro-f1
		Positive	Negative	Neutral	
CNN	Standard	40.2	24.9	95.0	43.37
LSTM	Standard	34.8	19.0	92.1	46.13
RNN	Standard	20.7	17.9	86.0	41.53
BERT (Devlin et al. 2018)	Standard	72.8	80.2	70.3	74.4
ALBERT (Lan et al. 2019)	Standard	71.1	72.5	67.6	70.4
ImpactCite (Mercier et al. 2021)	Standard	64.6	86.6	82.0	**77.73**
SVM (Athar 2011)	*	*	*	*	76.4

The results of the best performing approach in every category are mentioned as bold

et al. 2019) as a backbone, it uses a permutation generalization approach which helps the model to achieve generalization and state-of-the-art performance in many Natural Language Processing problems. *ImpactCite* model is suitable for our problem of citation sentiment analysis as the sentences in publications can be longer and their context heavily depends on the preceding and proceeding sentences.

XLNet is generally used in a combination of many different settings with variations in the number of layers and units per layer. In this work, we selected the XLNet-Large model which is very suitable for problems with long sentences and their respective context. We used an XLNet-Large model with 24-layers, 1024 hidden units, and 16 heads. We used a standard pre-trained XLNet-Large model and later fine-tune it for the task of citation sentiment classification. In the experiments, we initially employed a warm-up phase with a fixed learning rate and later we employed learning rate decay in the training phase. Transfer learning played an important role as the dataset available for the citation sentiment analysis was a relatively small dataset. Therefore, the pre-trained network adapts to the new domain and task using scarcely available data.

3.2.2 Evaluation and discussion

Evaluation results of ImpactCite are shown in Table 2. As there was no official dataset split mentioned in Athar (2011), therefore 10-fold cross-validation was performed for the selected models. The results show that CNN, RNN, and LSTM were unable to handle highly imbalanced data and could only learn representations from class with most samples, i.e., Neutral class. However, their performance in classes with low representation in the dataset was significantly worse.

On the other hand, sophisticated language models like BERT, ALBERT & ImpactCite were able to handle class imbalance and were able to learn representations from all classes. The reason for the performance improvement is that all these models were pre-trained and were later fine-tuned for the task of citation sentiment classification. For the sake of completeness, we also included results reported by Athar (2011) on the CSC dataset.

To conclude, ImpactCite outperformed all other approaches by achieving the highest macro-F1. Macro-F1 treats all classes equally irrespective of their number of samples. Contrary to BERT (Devlin et al. 2018) and ALBERT (Lan et al. 2019), ImpactCite was able to deal with the instances in long sentences and their relationships with each other in a given context. Therefore, ImpactCite sets a new state-of-the-art for citation sentiment analysis.

3.3 Keywords detection and topic modeling

This module is responsible for the task of detecting keywords from scientific publications and later using those keywords to identify topics. For this purpose, we employed an approach called Collective Connectivity-Aware Node Weight (CoCoNoW) (Beck et al. 2020). The overview of the (CoCoNoW) (Beck et al. 2020) pipeline is shown in Fig. 1. It consists of two stages where keywords are detected in the first stage followed by using those detected keywords in combination with the Computer Science Ontology (Salatino et al. 2018) to perform topic modeling. Details of CoCoNoW (Beck et al. 2020) pipeline are as follows:

3.3.1 Methodology

CoCoNow (Beck et al. 2020) pipeline consists of several phases, where each phase is responsible for a set of tasks. CoCoNow (Beck et al. 2020) takes the title, abstract and full text of a publication as input and performs the standard NLP pre-processing steps. These steps involve, tokenization, parts of part of speech tagging, lemmatization, stemming and filtering the least probable keywords candidates.

Once the pre-processing is complete, a graph is generated using the pre-processed tokens. In the graph, the tokens appearing in two consecutive sentences are connected with an edge between them with their normalized co-occurrence serving as the edge weight.

Table 3 Performance comparison of CoCoNoW (Beck et al. 2020) on the different datasets

Approach	k	Hulth2003 (Hulth 2003)			SemEval2010 (Kim et al. 2010)			NLM500 (Aronson et al. 2000)		
		P (%)	*R* (%)	*F*1 (%)	*P* (%)	*R* (%)	*F*1 (%)	*P* (%)	*R* (%)	*F*1 (%)
TF-IDF (Liu et al. 2009)	5	33.3	17.3	24.2	–	–	–	–	–	–
Topic Clustering (Liu et al. 2009)		35.4	18.3	24.3	–	–	–	–	–	–
Key2Vec (Mahata et al. 2018)		68.8	**25.7**	36.2	41.0	14.4	21.3	–	–	–
CoCoNoW (Beck et al. 2020)		**84.0**	**25.7**	**37.3**	**84.1**	**17.5**	**28.7**	48.8	11.4	17.9
TextRank (Mihalcea and Tarau 2004)	10	45.4	47.1	39.8	–	–	–	–	–	–
Word Embeddings (Wang et al. 2015)		38.7	**52.8**	44.7	–	–	–	–	–	–
Key2Vec (Mahata et al. 2018)		57.6	42.0	48.6	35.3	24.7	29.0	–	–	–
CoCoNoW (Beck et al. 2020)		**73.3**	41.9	**50.0**	**72.3**	**29.8**	**41.6**	43.3	19.8	26.3
Supervised approach (Hulth 2003)	16	25.2	51.7	33.9	–	–	–	–	–	–
TextRank (Mihalcea and Tarau 2004)	14	31.2	43.1	36.2	–	–	–	–	–	–
TF-IDF (Kim et al. 2010)	15	–	–	–	11.6	14.5	12.9	–	–	–
HUMB (Lopez and Romary 2010)	15	–	–	–	27.2	27.8	27.5	–	–	–
Key2Vec (Mahata et al. 2018)	15	55.9	50.0	52.9	34.4	32.5	33.4	–	–	–
CoCoNoW (Beck et al. 2020)	15	**63.5**	**52.9**	**54.2**	**62.2**	**39.2**	**46.5**	37.11	25.2	29.0
TextRank (Mihalcea and Tarau 2004)	25	–	–	18.4	–	–	–	–	–	–
DegExt (Litvak et al. 2011)		–	–	18.2	–	–	–	–	–	–
k-core (Rousseau and Vazirgiannis 2015)		–	–	43.4	–	–	–	–	–	–
PositionRank (Florescu and Caragea 2017)		45.7	64.5	50.4	–	–	–	–	–	–
sCAKE (Duari and Bhatnagar 2019)		45.4	**66.8**	51.1	–	–	–	–	–	–
CoCoNoW (Beck et al. 2020)		**54.8**	66.2	**56.8**	47.3	47.8	46.8	29.3	32.6	29.9
TextRank (Mihalcea and Tarau 2004)	30	–	–	–	–	–	13.7	–	–	10.7
DegExt (Litvak et al. 2011)		–	–	–	–	–	14.6	–	–	10.9
k-core (Rousseau and Vazirgiannis 2015)		–	–	–	–	–	29.3	–	–	20.2
PositionRank (Florescu and Caragea 2017)		–	–	–	25.3	31.3	27.5	19.7	26.6	21.9
sCAKE (Duari and Bhatnagar 2019)		–	–	–	35.8	47.4	40.1	24.5	35.0	28.3
CoCoNoW (Beck et al. 2020)		52.5	70.1	57.2	**42.6**	**51.5**	**45.8**	**26.7**	**35.3**	**29.5**

The results of the best performing approach in every category are mentioned as bold

The node weight is assigned using four criteria which include distance to the most central node, term frequency, occurrence in abstract, and occurrence location in the document. The scores from these criteria are summed together to assign the final node weight. Lastly, the Node and edge rank algorithm is applied on the weighted graph to acquire a list of keywords sorted in descending order of their importance. Eventually, a set of top K tokens is selected as keywords for the input document.

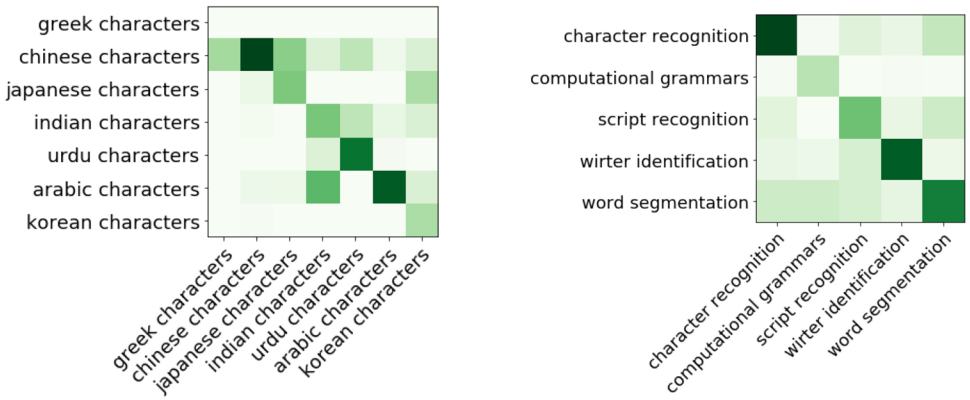
The second stage is responsible for performing topic modeling on the extracted keywords from the first stage. To perform topic modeling, we employed the Computer Science Ontology (CSO)[7] consisting of 23,800 nodes and 162,121 edges. The topic scores are assigned using the Levenshtein distance between each token or its synonym with a topic in the ontology. Lastly, the topic scores are normalized between 0 and 1.

3.3.2 Evaluation and discussion

This section discusses the evaluation of both stages of the selected approach. The first stage was responsible for keyword detection. It was evaluated on three datasets. Each approach in the literature uses a different value of *k* which represented the number of returned keywords. CoCoNow was also evaluated on different values of *k*. Table 3 shows the comparative evaluation of CoCoNoW with other approaches, it can be seen that CoCoNoW consistently outperforms all the approaches on all three datasets for a different number of *k* keywords.

The second stage of CoCoNoW was evaluated on ICDAR proceedings due to the lack of available ground truth. A hypothesis was developed for the evaluation of topic modeling which states that topics with similar topics more often

[7] https://cso.kmi.open.ac.uk.



(a) *Character Recognition* (b) *Pattern Recognition*

Fig. 3 Citation count for different super topics (Beck et al. 2020)

tend to cite each other. Figure 3a, b show the heatmaps representing citations among two sample topics, i.e., character recognition and pattern recognition respectively. It can be observed that the darkest colors on the diagonal axis support the evaluation hypothesis, therefore, suggesting that the topic assignment is correct.

3.4 Semantic index

In this section, we will introduce the formulation of our proposed novel Semantic index scores to assess the influence of the researchers in a scientific community. The purpose of the Semantic index is to assign a representative score to individual authors which depicts the extent of their contribution and its acceptance in the scientific community.

Generally, author indexes take into account the quantitative aspect of citations, i.e., number of citations received by publications. We propose a novel Semantic index that considers the nature of individual citations in addition to citation count, therefore, enabling us to integrate the qualitative aspect of citations in the Semantic index. In this work, we consider two qualitative aspects of citations namely citation sentiment and self-citation. The motivation behind this selection is fairly intuitive as we propose that not all the citations are equal, the first factor which sets apart one citation from the other is whether a paper is cited positively or negatively, i.e., appreciating and using the proposed approach or highlighting shortcomings of a research work respectively. In Semantic index, we only consider the citations which have a positive sentiment as those citations represent the appreciation and support of the scientific community for research work. For this purpose, we estimate the citation sentiment by using the approach mentioned in Sect. 3.2. The second factor which affects the quality of citation is whether an author is citing their own papers which is synonymous with a famous English idiom "Self-praise is no recommendation". Therefore, any citation which is an instance of self-citation is not considered during the estimation of the Semantic index for an author. The resultant number of citations will be referred to as $N_{positive}$ in this paper.

In addition to the above-mentioned qualitative aspects, we also consider the multi-faceted community interactions of an author to effectively evaluate their position in a scientific community. These multi-faceted interactions are represented by different centrality measures. In graph theory, a centrality measure is used to rank nodes based on their position in the graph. To estimate these centrality measures, we use two types of graph networks one is the author citation network and the other is the author collaboration network.

For this purpose, we construct an author citation network by representing all citations extracted from the publications in the form of a directed graph where each node represents an author of a publication, and the relation between two nodes highlights citations pointing in the direction of cited author. On the other hand, the author's collaboration network depicts the collaboration among the authors in a scientific community. It is constructed using the information extracted from the header of a publication, where each author is represented by a node and a non-directed relation between two nodes represents collaboration in a publication. Once both author citation and collaboration networks are ready, we can now estimate the value of different centrality indicators for the Semantic index of an author. The description of the selected centrality measures is as follows:

1. *Degree centrality* It represents the extent to which a node is connected in a network.
2. *Eigenvector centrality* It quantifies the transitive influence of a node on its neighboring nodes.
3. *Betweenness centrality* It measures the influence and control of a node on the flow of information.
4. *Closeness centrality* It estimates the extent to which a node can efficiently spread information.
5. *Indegree centrality* It represents the number of incoming connections, i.e., citations from other nodes.

Table 4 Aspects covering DORA guidelines for evaluating research impact

Aspects	h-index	g-index	i10-index	Eigenfactor	Impact factor	Semantic index
Quality	−−	−−	−−	−−	−−	++
Circumstances	−−	−	−−	−	−	++
Content oriented	−−	−−	−−	−−	−−	++
Manipulation	−−	−−	−−	−	−−	+
Reliability	−	−	−	−−	−−	++
Transparency	+	+	+	+	−	++

Each of these centrality measures refers to a specific role in the scientific community which would be discussed in Sect. 3.10.2. After computing all the centrality measures for every node in the network, we prepare the weighted centralities by taking a product of non-self-cited positive citations $N_{positive}$ with the sum of all centrality scores to finally compute the Semantic index value for each node in the network. The proposed Semantic index can be formulated as follows:

$$\mathrm{SIndex_n} = \log((c_{\mathrm{deg}} + c_{\mathrm{eig}} + c_{\mathrm{bet}} + c_{\mathrm{clo}} + c_{\mathrm{ind}}) \times N_{\mathrm{positive}}) \tag{1}$$

where N_{positive} represents the total number of positive citations received by an author. It is to be noted that N_{positive} does not include any self-citation. c_{deg}, c_{eig}, c_{bet}, c_{clo}, and c_{ind} represent the Degree, Eigenvector, Betweenness, closeness, and indegree centralities respectively. By default, all the centrality measures contribute equally toward estimating the value of the Semantic index. Additionally, the effect of increasing or decreasing the contribution of each centrality measure can be later visualized in the Sect. 3.10.2.

3.4.1 Evaluation and discussion

We evaluated the Semantic index by inspecting its compliance with the Declaration on Research Assessment (DORA) guidelines which were developed in 2012 as a result of the Annual Meeting of the American Society for Cell Biology in San Francisco. These guidelines were further developed and refined over a period of time and are currently been actively maintained. This initiative provides instructions and best practices to all researchers, organizations, funding agencies, and scientific communities to assess the quality of scholarly research. So far, there are 21,729 participants and organizations from 158 countries who have already signed the DORA declaration. For our comparative evaluation, we selected the 5 most widely adopted author indexes, i.e., h-index, g-index, i10-index, Eigenfactor, and Impact factor.

Now, we will discuss the core aspects of the DORA guidelines and their compliance in the case of the Semantic index and some most common author indexes. Table 4 provides a quick overview of this comparative compliance with DORA guidelines while their details are discussed as follows:

Suitability for quality evaluation This aspect represents the overall purpose of an author index, which is to evaluate the quality of the research work performed by a researcher. However, existing author indexes have certain limitations along with their advantages. The limitations of the h-index, g-index, i10-index, and Eigenfactor deem them not suitable for evaluating the quality of research work as they heavily rely on a subset of publications for assessment. For instance, any publication with citations less than the h-index of an author will be discarded. Moreover, the impact factor was initially introduced to help librarians to facilitate in deciding which potential journal volume they should buy for their libraries, and now the scientific community seems to measure the quality of a journal using the impact factor. Such affairs make the existing indexes unsuitable for accessing the quality of research work. However, our proposed Semantic index considers the qualitative aspect of a publication, i.e., citation sentiment to justify its suitability to serve as a tool for assessing the quality of research work.

Circumstances The second aspect that DORA deems important for evaluating the impact of research is the consideration of individual circumstances. For instance, a researcher who joined recently would have less time to get citations as compared to the long publishing old researchers. All h-index, g-index, i10-index, Eigenfactor, and Impact factor do not take into account such individual circumstances of an author and usually take a very long time to gradually increase the score of these indexes. On the other hand, the Semantic index considers all publications irrespective of their number of citations and hence provides a consistent increase in score upon receiving new citations.

Content oriented Another aspect of impact assessment of research work is to consider the content of the publications while performing the assessment. As already mentioned, the h-index, g-index, i10-index, Eigenfactor, and Impact factor only consider raw citation count to estimate the impact of research work and all these indexes do not consider the content of the publications. Citation count is a superficial feature, as it does not covey any information about the quality of a citation, i.e., if a publication is cited positively, negatively, or neutrally. Contrary to this, the Semantic index takes into account the content of the publications to identify

the sentiment behind a citation so that it can be given credit accordingly.

Transparency and reliability Transparency and Reliability of indexes are the key aspects of analyzing the quality of research work. Existing indexes are somewhat transparent as we know that based on what data it was estimated. However, they have limited reliability as the publications with a low number of citations are completely overlooked which might be crucial in some domains, i.e., Medicine. In contrast, for estimating the Semantic index, data are collected directly from the publications, therefore it is much more transparent and reliable. Additionally, it is generalizable as it considers all publications with any number of citations thus making it suitable for any domain.

3.4.2 Limitations of existing indexes

In this section, we will compare the limitations of existing indexes with the Semantic index. Similar to the previous section, we have selected the most widely used indexes for comparative analysis, i.e., h-index, g-index, i10-index, Eigenfactor, and Impact factor. Following are the key limitations of all the existing indexes:

Self citations Self-citation is the most common challenge when assessing the impact of a research profile. Indexes like h-index, g-index, i10-index, and Impact factor do not handle self-citations and continue to consider them during the estimation of their index values. However, unlike other indexes, Eigenfactor and the Semantic index discard the self-citations as they are not considered to be a part of the actual impact of a research publication.

Lack of fairness Indexes like the h-index and g-index favor the old researchers who have been publishing for a while to sustain their index score as their publications received many citations over years or decades. For the new researchers, they have to wait for years or decades to accumulate a high number of citations and eventually reach the same level as old researchers. On the other hand, the i10-index, Eigenfactor, and Semantic index consider the citations from all papers, and their values start increasing with the growing number of citations. Therefore, it does not require a lot of time to build up, hence supporting new and existing researchers relatively fairly and equally.

Quantity versus quality In the context of publishing, the scientific community has two informal popular schools of thought. One focuses more on the number of publications and the other one emphasizes more on the quality of the publications. This results in the cases where researchers have either high volume and low quality or low volume and high quality respectively. The selected indexes are ineptly not able to handle such cases as either of these cases affect the index score of the h-index, g-index, i10-index, Eigenfactor, and Impact factor. Contrary to popular indexes, the Semantic index can well handle the delicate balance between Quantity and Quality. Since the Semantic index favors all publications equally and therefore can handle both of these cases.

Coverage Some indexes, i.e., h-index and g-index employ components like h-core and g-core which results in partial coverage of publications including citations. These limited components restrict the scope of insights provided by the index scores and hence provide an incomplete picture of a researcher in the community. On the other hand, Eigenfactor, Impact factor, and Semantic index are independent of typical h-core or g-core components to estimate the index score for a given researcher. Therefore, the limitations associated with h-core or g-core are irrelevant for these indexes.

Meaningfulness Author indexes like h-index, g-index, i10-index, and Eigenfactor are meaningful to some extent as they attempt to highlight the importance of notable authors in the scientific community. However, they severely lack evaluation of the quality of the work due to the several limitations mentioned above. This results in partial meaningfulness of these indexes. On the contrary, the Semantic index not only analyzes all available citations but also the quality of each citation by analyzing the sentiment of each citation. The Semantic index has the same granularity as most popular indices like the h-index. However, the Semantic index is more comprehensive and is, therefore, more meaningful.

3.5 Consolidation of bibliographic data

Data extracted by the Data Extraction modules in Stage 1 is consolidated into a common data storage. For this purpose, ACE employs MongoDB as the central storage where all the data are collected and secured. One of the challenges faced during the consolidation of data was to precisely identify each author and correctly assign the respective publications to the right author. This phenomenon is known as Author name disambiguation. One of the reasons which gave rise to this challenge is the use of abbreviated names in the reference section of the publications. For instance, there are two persons with names Anthony Davidson and Andrew Davidson. Both persons write their short name as A. Davidson. In a scenario where we only see the shortened name, it is very challenging to identify which specific person is being referred to in this name. Another challenge could appear if there is an error in the extracted text. The text from detected bibliographic references is extracted by performing Optical Character Recognition (OCR). There is a possibility for the introduction of an OCR error in the extracted text. Especially, in the case of name initials, any misclassification can lead to an entirely different name for a person.

To tackle the challenges in author name disambiguation and ensure the quality of data, we employed a set of external resources, i.e., Crossref and Semantic Scholar to validate the

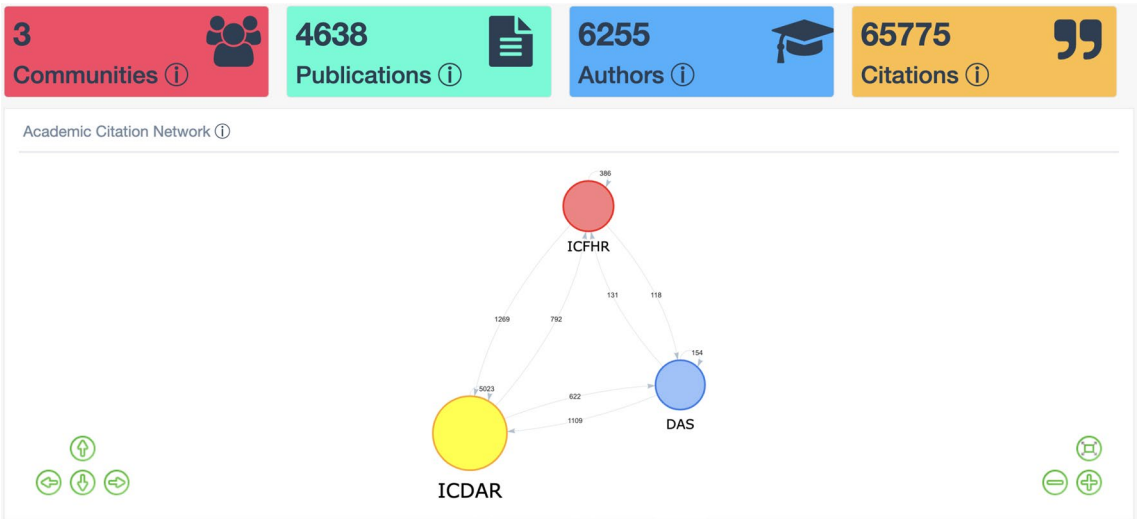


Fig. 4 Community interactions in the domain of document analysis

accuracy of extracted data. Both of these external resources have a huge collection of bibliographic data and have specified interfaces through which we can query data about a specific publication. Any error or disambiguation in names that arise during the data extraction phase is eradicated by verifying it with author names collected from the external resources. It is worth mentioning that the author name disambiguation is not performed solely on the data received from external resources. The external resources provide some metadata attributes in addition to those extracted from the scientific publication. The attribute values from external resources, i.e., affiliation, aliases, etc., are used to validate the identity of an author. Once the author name disambiguation is complete, all the attributes from different resources are consolidated. The data are now ready for further processing.

3.6 Computation of statistics

This module deals with analyzing the consolidated data for estimation of certain author level and community-level statistics. These statistics are crucial in identifying the author and community-level trends in different aspects.

Although, there are numerous fine-grained statistical figures extracted from the consolidated data, i.e., number of publications or citations for an author, etc. However, we will only discuss the most prominent statistical figures estimated in this module. One of such figures is the generation of co-authors graph. We represented all authors as nodes such that all co-authors have a unidirectional link with one another. Once the co-authorship graph for a whole community is ready, then we apply Girwan Newmann clustering (Girvan and Newman 2002) on the co-authorship graph. It results in clusters of co-authors.

On the other hand, we also generate a community network graph. Given the consolidated data, we take authors in a community and represent them as nodes in a community network graph. Additionally, we employed the citation data to draw links between network nodes. So the resultant graph is a citation network graph. To incorporate more information in the citation network graph we color-coded the nodes based on their co-authorship cluster.

3.7 Visualization engine

Once all the statistics have been successfully estimated, the final data are delivered to the visualization engine, which uses the given data and visualizes in more than one way to highlight different trends. There are different visualizations with a granularity that spans over three levels. The highest level contains the visualization representing the domain-level insights, followed by less detailed visualization representing the community-level insights, and finally the author-level insights. The visualizations for each of these granularity levels are described below.

3.8 Domain-level insights

As the name suggests, these visualizations represent insights from the domain level. For the proof of concept, we selected the domain of Document Analysis as our sample domain. Figure 4 shows overall statistics of the Document Analysis domain. The visualization shows three communities in the selected domain which have a total of 6255 authors who contributed 4638 publications with 65775 incoming

(a) *Top Cited Authors* (b) *Top Cited Papers* (c) *Top Influential Authors*

Fig. 5 Different interaction patterns of an author in a scientific community

Fig. 6 Topic popularity in the domain of document analysis

(a) *Top Topics by Citations* (b) *Top Topics by Contributions*

and outgoing citations. However, the graph shown below the statistics depicts the interaction between communities. Each node in the visualization represents a community in the domain and the size of each node represents the number of citations it received. If the publications in a community receive relatively more citations, then their size will be bigger than other communities. The links between nodes represent the citation relation between two communities. The direction of the link represents the citation direction and the number on the link shows the number of times the publications of the target community were cited by the other. Self-citations are also shown in the graph which is a key indicator to understanding the popularity of the publications within the community as well as in the other communities.

The next visualization in Fig. 5 shows the distribution of top publications and authors among all three communities. Figure 5a, b show the distribution of top 100 authors in all communities ranked in the order of the number of citations received and publications respectively. However, Fig. 5c shows the distribution of the top 100 authors with the highest Semantic index score in all communities. Lastly, Fig. 6a, b show the top topics with most citations and contributions among all communities. The size of the topic refers to the number of citations or contributions received by that topic.

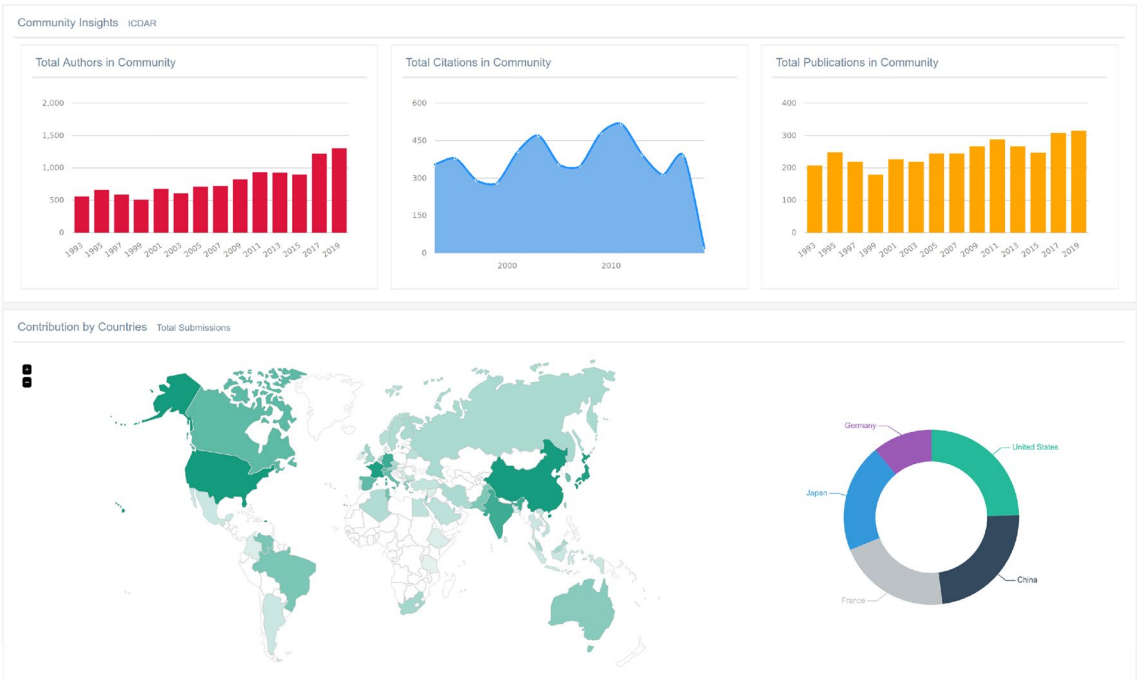


Fig. 7 Example of overall community highlights

3.9 Community-level insights

This section describes the community-level insights which are more abstract than author-level highlights, however, they are more precise about the overall community. Such highlights play an important role in policing and paving the future path of a scientific community.

3.9.1 Community highlights

Overall community highlights give us a quick insight into community-level statistics. Figure 7 shows an example

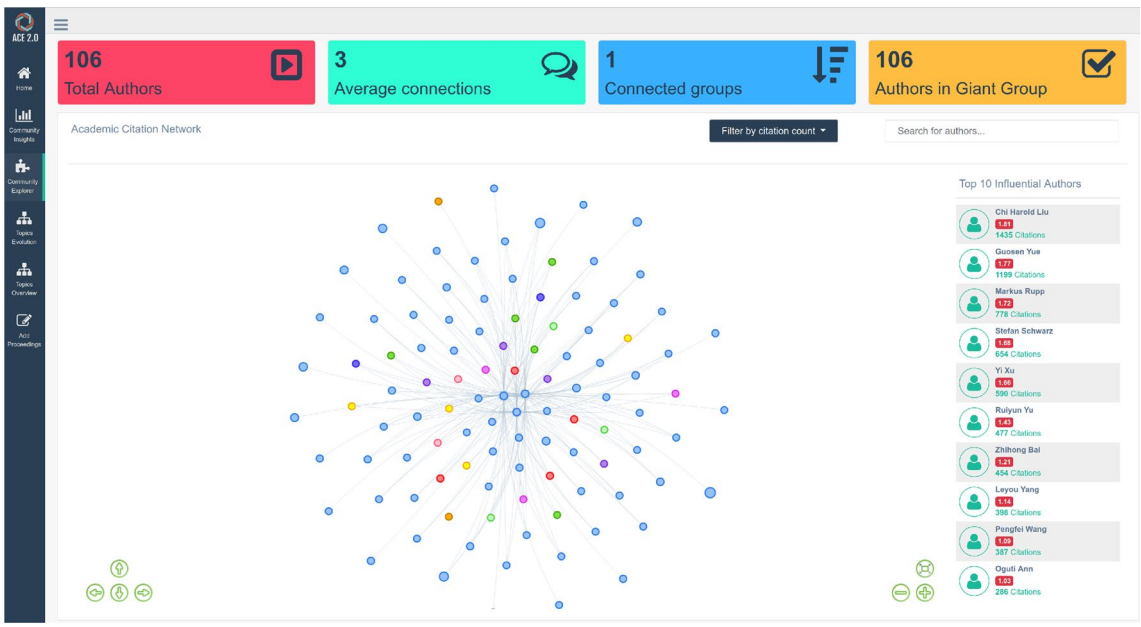


Fig. 8 Community as a network

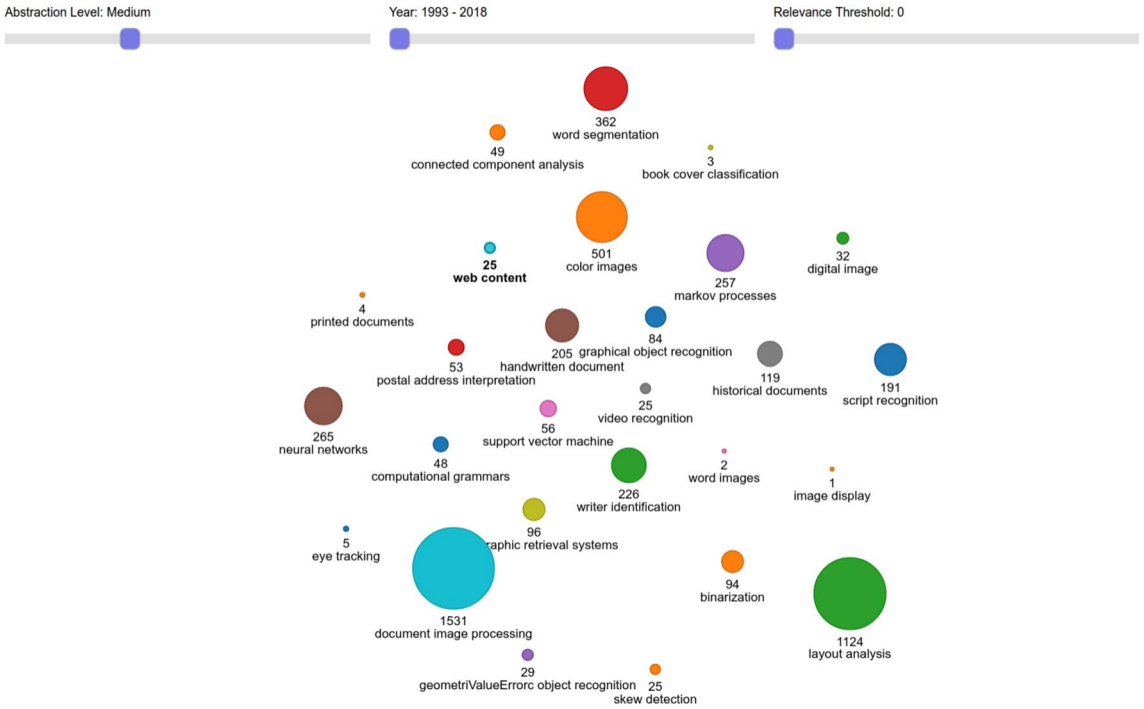


Fig. 9 Dynamic representation of topic evolution

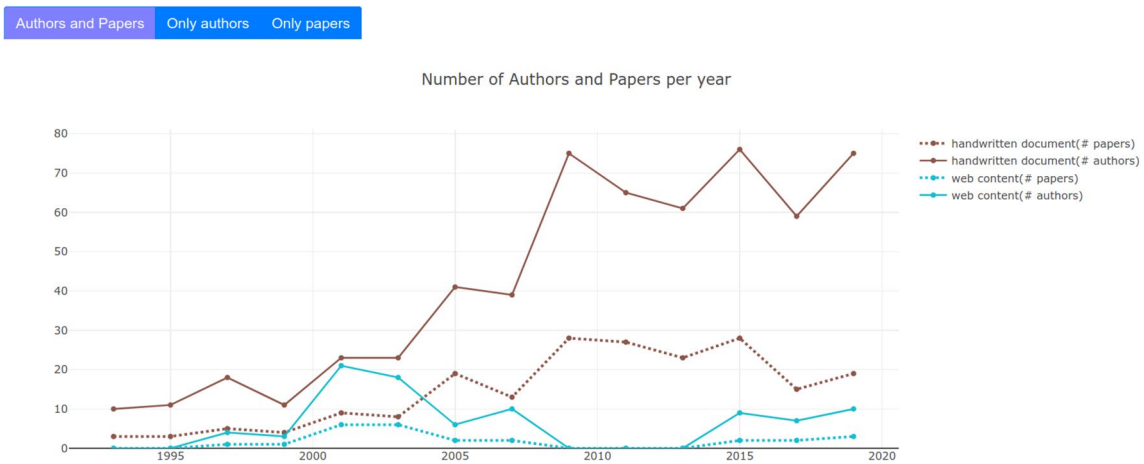


Fig. 10 Visualizing topic evolution of selected topics

of different visualizations related to a community. A red bar chart shows the number of authors who participated in a proceeding year. It allows us to gauge if the number of participants is increasing or decreasing over time. On the other hand, the number of submissions received each year is represented in the orange bar chart. We also visualize the number of citations received by a publication by each proceeding year with the blue chart. It provides an immediate insight into which years are more popular within the community. From the example given, it is clear that the proceedings of the year 2011 are the most popular and have received the highest number of citations. Furthermore, we also visualize participation by country to see which countries are contributing the most to this

Authors Overview

Show 10 entries

Search:

Author Name	Publications	Citations	Semantic Index Score	Collaborator Score	Idea Generator Score	Community Connector Score	Opinion Leader Score	Community Expert Score
Andreas Dengel	19	54	2.76	0.1075	0.1652	0.0125	0.1223	0.0750
C. V. Jawahar	15	54	3.15	0.1775	0.0752	0.0524	0.1285	0.0875
Basilios Gatos	15	53	2.12	0.0975	0.1153	0.0386	0.1541	0.0775
Seiichi Uchida	14	7	0.00	0.0350	0.0101	0.0006	0.0050	0.0075
Chew Lim Tan	14	47	1.68	0.0700	0.1120	0.0360	0.1403	0.0425
M. Blumenstein	12	21	1.44	0.1025	0.1840	0.0149	0.1053	0.0450
J. Ogier	12	13	0.00	0.0425	0.1386	0.0000	0.0000	0.0000
Koichi Kise	12	27	1.45	0.1000	0.0696	0.0042	0.1398	0.0775
U. Pal	12	23	1.12	0.0975	0.2671	0.0146	0.1249	0.0375
Faisal Shafait	12	25	0.00	0.0275	0.0125	0.0011	0.0075	0.0100

Showing 1 to 10 of 300 entries

Previous 1 2 3 4 5 ... 30 Next

Fig. 11 Authors overview

academic community. It can be observed that the United States has the most contributions among all countries in the given scientific community.

3.9.2 Community as a network

This section describes an important community-level feature highlight where the whole community is represented in the form of a community network graph. Figure 8 shows the example of an academic community, where each node represents an author. Authors are connected with links among them. Each link between two nodes represents a citation relationship. The color of the nodes represents the collaboration groups in the scientific community. Author nodes with the same color tend to collaborate in their research work. The network graph can be filtered using the citation count threshold. It will filter the graph and only shows the authors having the specified number of overall citations or more.

3.9.3 Topic evolution and its trends

This section describes the visualization related to topic evolution. Figure 9 shows a dynamic visualization of the evolution of the topics over a while. It represents topics in the form of bubbles. The year slider on the top middle of the chart can be moved to see the effect on individual topic bubbles. When you move the slider, a topic bubble might become larger or smaller, representing its popularity in the selected year. Figure 10 shows an example of two selected topics along with the number of contributions for

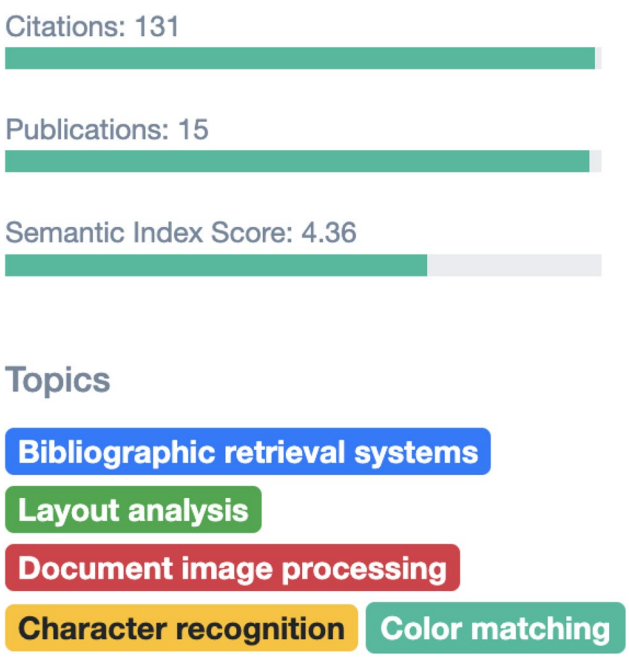


Fig. 12 Author statistics overview

these topics over the years. The topic "Handwritten Document" and "Web Content" are represented in brown and blue colors respectively. Solid lines represent the number of scientific publications submitted on a specific topic. On the other hand, dotted lines represent the number of authors who contributed to a specific topic. It is quite

evident that both topics started with minimal interest at the start. With time, the topic of 'Handwritten Documents' became increasingly popular within the scientific community. Contrary to this, the topic of "Web Content" gradually increased in popularity. However, after the year 2007 scientific community lost interest in this topic. Such trends help us understand the community's interests and make decisions regarding the future direction of the community.

3.9.4 Authors overview

The table shown in Fig. 11 shows different statistics for all authors in a scientific community. These statistics range from simple citation or publication count to complex Semantic index values. The authors can be sorted with respect to any measure in the table by clicking the title of the measure of interest. It provides a quick quantitative comparison between different authors in a single glimpse.

3.10 Author-level insights

This section discusses the visualizations designed for displaying the author-level statistics. Some key author-level visualizations are as follows:

3.10.1 Author statistics

Figure 12 shows how the basic statistics of a specific author are displayed in the author's profile. The bars on the top show the citation count, the number of publications, and the Semantic index score of a given author. The extent to which the bars are filled represents their percentile. On the bottom, we can see the top 5 topics on which this author is continuing their research.

3.10.2 Community roles

This section discusses the roles of an author in an academic community. In this work, we consider each author for five different roles. These roles are represented by different centrality measures discussed in Sect. 3.4. Different roles and their description are as follows:

- *Collaborator* Collaborates more often with members of the community. It is represented by the Degree centrality of the author.
- *Idea generator* Highly influential individual, who brings new ideas which are widely accepted by the community. It is represented by the Eigenvector centrality of the author.

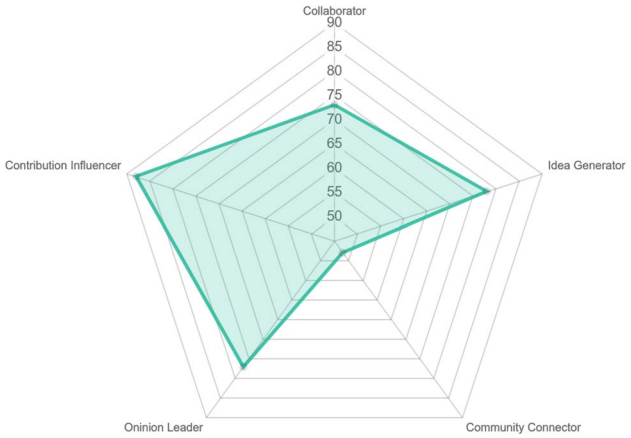


Fig. 13 Visualization of an author's roles in a community

- *Community connector* Diversely publishes with different cliques in the community. It is represented by the Betweenness centrality of the author.
- *Opinion leader* Holds a strong network and is capable of influencing an opinion about a trend in the community. It is represented by the Closeness centrality of the author.
- *Contribution influencer* Dominates the community with their important scientific contributions. It is represented by the Indegree Centrality of the author.

Figure 13 represents a visualization example of an author's role in the community. It can be noticed that in this specific example the person in the discussion is more of a contribution influencer as compared to any other role in the community. Figure 14 shows a set of sliders for each role that can slide to increase or decrease the extent to which they contribute toward estimating the Semantic index.

3.10.3 Citation sentiment analysis

In this section, we will discuss the system's features related to the citation sentiment of an author in a scientific community. Figure 15 shows an example visualization for the overall citation sentiment of an author. A doughnut chart is used for this visualization. Positive, negative, and neutral citation sentiments are represented in green, red, and gray colors. It can be seen that this specific author has mostly received neutral citations. However, positive citations also have a fair share in total citations which shows that the scientific contributions by the author in the discussion have been fairly accepted and appreciated in the academic community. Figure 16 shows a list of publications of a given author. The last column on the right shows the citation sentiment of all the citations received by each publication.

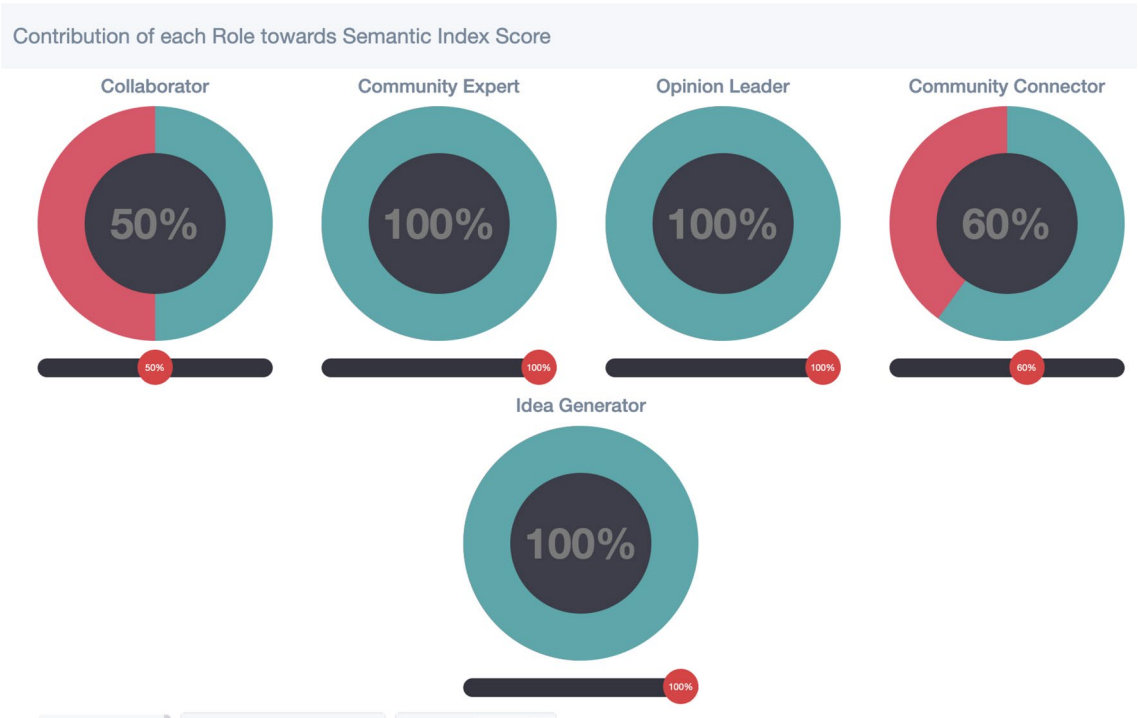


Fig. 14 Roles contribution sliders

3.10.4 Author citation patterns

This section discusses the citation pattern of individual authors. Figure 17a shows a visualization example of a radar chart representing the citation pattern of an author. The visualization shows the top 10 authors who were cited by the author in the discussion and the number of times they were cited is represented in green color. However, the data points in red color represent how many times those authors cited back the author in the discussion. With this visualization, we can instantly realize the citation interaction of an author with the other community members. In the given example we can see that the author in the discussion cited himself more than any other else in the community.

3.10.5 Author collaboration

This section presents the collaboration pattern of an author. Figure 17b shows an example collaboration visualization of an author. Each data point shows the number of times this author collaborated with other researchers in the academic community and is represented in blue color. This visualization in combination with citation pattern visualization can uncover even more patterns. For example, in the given examples, we can see that the current author tends not to cite his second most common collaborator very often.

3.10.6 Customized author network

Our proposed tool also includes a customized network graph on each author's profile page. Figure 18 shows an example of a custom network of an author. Where the author in the discussion is at the center of the network graph. Each node represents other authors in the academic community who either cited or were cited by the current author. With this

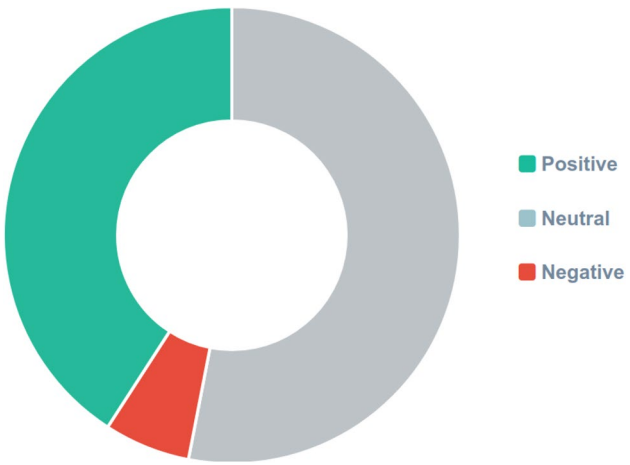


Fig. 15 Visualization of citation sentiment all the author's citations in a community

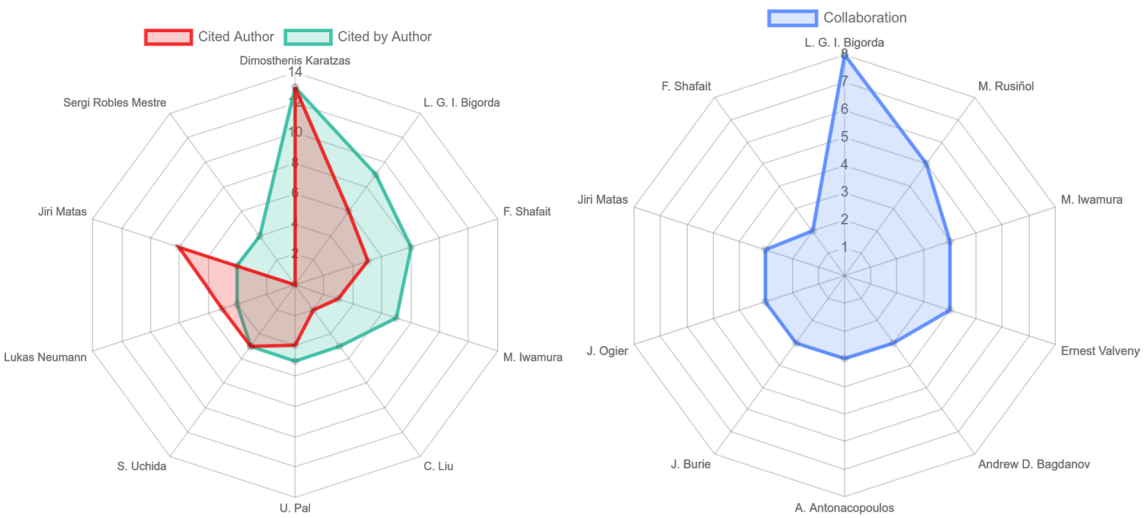
#	Publication Title	Co-Authors	Citation Count
1	A full English sentence database for off-line handwriting recognition		33 Citations
2	IAM-OnDB - an on-line English sentence database acquired from handwritten text on a whiteboard		24 Citations
3	Writer identification using text line based features		16 Citations
4	On-Line Handwritten Text Line Detection Using Dynamic Programming		13 Citations
5	Improving HMM-Based Keyword Spotting with Character Language Models		10 Citations
6	Text line segmentation and word recognition in a system for general writer independent handwriting recognition		10 Citations
7	Generation of synthetic training data for an HMM-based handwriting recognition system		9 Citations

Fig. 16 Publications list with sentiment

visualization, we can get a quick insight into the extent of networking of an author in the academic community.

4 Conclusion

This paper proposes ACE 2.0, a comprehensive tool for analyzing and obtaining instant insights about scientific communities. An integrated pipeline is characterized by modules that are each responsible for performing a specific task. Each module of the system utilized state-of-the-art models to detect bibliographic references, analyze sentiment, and identify keywords. Furthermore, we proposed a Semantic index that quantifies the influence of a researcher in a scientific community. To accomplish this goal, the Semantic index has taken into consideration both quantitative and qualitative aspects of citations, as well as the impact of various roles played by researchers within the scientific community. Its generic nature and robustness make it an ideal index to mitigate the challenges faced by other indexes. Additionally, ACE 2.0 is capable of automatic digital profiling of all the researchers in the



(a) *Citation Pattern* (b) *Collaboration Patter*

Fig. 17 Different interaction patterns of an author in a scientific community

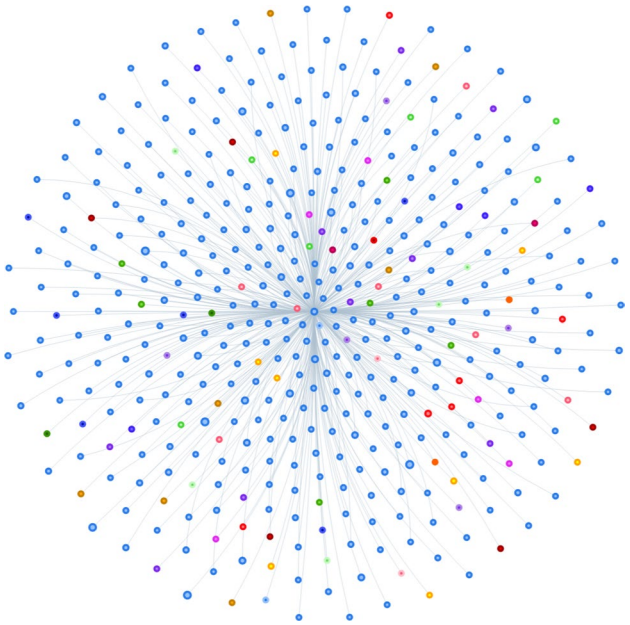


Fig. 18 Visualization of author's custom network in a community

community and is equipped with a diverse visualization engine. With the help of different examples discussed in this paper, we got instant community insights. Using these insights, we were able to identify citation patterns of authors, and interests of the scientific community, which can help us in policing and planning the future direction of the community.

Author contributions STRR implemented the tool, performed analysis & evaluations, and wrote the manuscript text. SA provided feedback regarding the usability of the frontend. AD proposed the overall idea for the tool, inspired & refined the idea of semantic index and reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ahmed MW, Afzal MT (2020) FLAG-PDFe: features oriented metadata extraction framework for scientific publications. IEEE Access 8:99458–99469

Alonso S, Cabrerizo F, Herrera-Viedma E, Herrera F (2010) hg-index: a new index to characterize the scientific output of researchers based on the h-and g-indices. Scientometrics 82(2):391–400

Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ (2000) The NLM indexing initiative. In: Proceedings of the AMIA symposium. American Medical Informatics Association, p 17

Athar A (2011) Sentiment analysis of citations using sentence structure-based features. In: Proceedings of the ACL 2011 student session, pp 81–87. Association for Computational Linguistics, Portland, OR, USA . https://www.aclweb.org/anthology/P11-3015

Beck M, Rizvi STR, Dengel A, Ahmed S (2020) From automatic keyword detection to ontology-based topic modeling. In: Bai X, Karatzas D, Lopresti D (eds) Document analysis systems. Springer, Cham, pp 451–465

Beliga S (2014) Keyword extraction: a review of methods and approaches. University of Rijeka, Department of Informatics, 1–9

Bhardwaj A, Mercier D, Dengel A, Ahmed S (2017) Deepbibx: deep learning for image based bibliographic data extraction. In: Liu D, Xie S, Li Y, Zhao D, El-Alfy E-SM (eds) Neural information processing. Springer, Cham, pp 286–293

Bhardwaj A, Erhard L, Klein A, Zander S, Zumstein P (2018) ICONIP dataset: labeled reference data from the linked open citation database (LOC-DB) project. https://madata.bib.uni-mannheim.de/id/eprint/268 . https://doi.org/10.7801/268

Biswas SK, Bordoloi M, Shreya J (2018) A graph based keyword extraction model using collective node weight. Expert Syst Appl 97:51–59

Bollen J, Van de Sompel H, Hagberg A, Chute R (2009) A principal component analysis of 39 scientific impact measures. PLoS ONE 4(6):1–11. https://doi.org/10.1371/journal.pone.0006022

Boudin F (2018) Unsupervised keyphrase extraction with multipartite graphs. arXiv preprint arXiv:1803.08721

Cai L, Tian J, Liu J, Bai X, Lee I, Kong X, Xia F (2019) Scholarly impact assessment: a survey of citation weighting solutions. Scientometrics 118(2):453–478

Carpena P, Bernaola-Galván P, Hackenberg M, Coronado A, Oliver J (2009) Level statistics of words: finding keywords in literary texts and symbolic sequences. Phys Rev E 79(3):035102

Carretero-Campos C, Bernaola-Galván P, Coronado A, Carpena P (2013) Improving statistical keyword detection in short texts: entropic and clustering approaches. Physica A 392(6):1481–1492

Chen C, Yang K, Chen C, Ho J (2012) BibPro: a citation parser based on sequence alignment. IEEE Trans Knowl Data Eng 24(2):236–250. https://doi.org/10.1109/TKDE.2010.231

Cliche M (2017) BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pp 573–580. https://doi.org/10.18653/v1/S17-2094

Costas R, Bordons M (2008) Is g-index better than h-index? An exploratory study at the individual level. Scientometrics 77(2):267–288

Councill IG, Giles CL, Kan M (2008) ParsCit: an open-source CRF reference string parsing package. In: Proceedings of the international conference on language resources and evaluation, LREC 2008, 26 May–1 June 2008, Marrakech, Morocco . http://www.lrec-conf.org/proceedings/lrec2008/summaries/166.html

Dai J, Li Y, He K, Sun J (2016) R-FCN: object detection via region-based fully convolutional networks. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural

information processing systems, vol 29. The MIT Press, Cambridge, pp 379–387
Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. In: 2017 IEEE international conference on computer vision (ICCV), pp 764–773. https://doi.org/10.1109/ICCV.2017.89
Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R (2019) Transformer-XL: attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860
Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
Duari S, Bhatnagar V (2019) sCAKE: Semantic connectivity aware keyword extraction. Inf Sci 477:100–117
Erhard L, Klein A, Rizvi STR, Zander S, Zumstein P (2019) RefDet dataset: additional labeled reference data from the linked open citation database (LOC-DB) project. https://madata.bib.uni-mannheim.de/id/eprint/283. https://doi.org/10.7801/283
Esuli A, Sebastiani F (2006) Determining term subjectivity and term orientation for opinion mining. In: 11th conference of the European chapter of the association for computational linguistics
Florescu C, Caragea C (2017) A position-biased pagerank algorithm for keyphrase extraction. In: Thirty-first AAAI conference on artificial intelligence
Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826. https://doi.org/10.1073/pnas.122653799
Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci 102(46):16569–16572
Hulth A (2003) Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of the 2003 conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 216–223
Kim SN, Medelyan O, Kan M-Y, Baldwin T (2010) Semeval-2010 task 5: automatic keyphrase extraction from scientific articles. In: Proceedings of the 5th international workshop on semantic evaluation, pp 21–26
Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning, ICML '01. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 282–289. http://dl.acm.org/citation.cfm?id=645530.655813
Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: a lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942
Lauscher A, Eckert K, Galke L, Scherp A, Rizvi STR, Ahmed S, Dengel A, Zumstein P, Klein A (2018) Linked open citation database: enabling libraries to contribute to an open and interconnected citation graph. In: Proceedings of the 18th ACM/IEEE on joint conference on digital libraries, JCDL '18. ACM, New York, NY, USA, pp 109–118. https://doi.org/10.1145/3197026.3197050
Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
Litvak M, Last M, Aizenman H, Gobits I, Kandel A (2011) Degext—a language-independent graph-based keyphrase extractor. In: Advances in intelligent web mastering, 3. Springer, pp 121–130
Liu Z, Li P, Zheng Y, Sun M (2009) Clustering to find exemplar terms for keyphrase extraction. In: Proceedings of the 2009 conference on empirical methods in natural language processing, vol 1, pp 257–266
Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
Lopez P, Romary L (2010) HUMB: automatic key term extraction from scientific articles in GROBID. In: Proceedings of the 5th international workshop on semantic evaluation. Association for Computational Linguistics, pp 248–251
Mahata D, Shah RR, Kuriakose J, Zimmermann R, Talburt JR (2018) Theme-weighted ranking of keywords from text documents using phrase embeddings. In: 2018 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, pp 184–189 . https://doi.org/10.31219/osf.io/tkvap
Matsuoka D, Ohta M, Takasu A, Adachi J (2016) Examination of effective features for CRF-based bibliography extraction from reference strings. In: 2016 eleventh international conference on digital information management (ICDIM), pp 243–248. https://doi.org/10.1109/ICDIM.2016.7829774
Mercier D, Rizvi S, Rajashekar V, Dengel A, Ahmed S (2021) ImpactCite: an XLNet-based solution enabling qualitative citation impact analysis utilizing sentiment and intent. In: Proceedings of the 13th international conference on agents and artificial intelligence-volume 2: ICAART. INSTICC, pp 159–168 . https://doi.org/10.5220/0010235201590168
Mercier D, Rizvi STR, Rajashekar V, Ahmed S, Dengel A (2022) Utilizing out-domain datasets to enhance multi-task citation analysis. In: Rocha AP, Steels L, van den Herik J (eds) Agents and artificial intelligence. Springer, Cham, pp 113–134
Mihalcea R, Tarau P (2004) Textrank: bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing
Munikar M, Shakya S, Shrestha A (2019) Fine-grained sentiment classification using BERT. In: 2019 artificial intelligence for transforming business and society (AITB), vol 1, pp 1–5
Nikolentzos G, Meladianos P, Stavrakas Y, Vazirgiannis M (2017) K-clique-graphs for dense subgraph discovery. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 617–633
Ohsawa Y, Benson NE, Yachida M (1998) Keygraph: automatic indexing by co-occurrence graph based on building construction metaphor. In: Proceedings IEEE international forum on research and technology advances in digital libraries-ADL'98. IEEE, pp 12–18 . https://doi.org/10.1109/adl.1998.670375
Pay T, Lucci S (2017) Automatic keyword extraction: an ensemble method. In: Conference: IEEE Big Data 2017, at Boston
Rabby G, Azad S, Mahmud M, Zamli KZ, Rahman MM (2018) A flexible keyphrase extraction technique for academic literature. Procedia Comput Sci 135:553–563
Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems, vol 28. Curran Associates Inc., Red Hook, pp 91–99
Rizvi STR, Lucieri A, Dengel A, Ahmed S (2019) Benchmarking object detection networks for image based reference detection in document images. In: 2019 digital image computing: techniques and applications (DICTA), pp 1–8
Rizvi STR, Dengel A, Ahmed S (2020) A hybrid approach and unified framework for bibliographic reference extraction. IEEE Access 8:217231–217245. https://doi.org/10.1109/ACCESS.2020.3042455
Rousseau F, Vazirgiannis M (2015) Main core retention on graph-of-words for single-document keyword extraction. In: European conference on information retrieval. Springer, pp 382–393
Salatino A, Thanapalasingam T, Mannocci A, Osborne F, Motta E (2018) The computer science ontology: a large-scale taxonomy of research areas. In: 17th international semanticweb conference,

Monterey, CA, USA, October 8–12, 2018, proceedings, Part II, pp 187–205

Sautter G, Böhm K (2012) Improved bibliographic reference parsing based on repeated patterns. In: Zaphiris P, Buchanan G, Rasmussen E, Loizides F (eds) Theory and practice of digital libraries. Springer, Berlin, pp 370–382

Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B (2014) Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers), pp 1555–1565

Thongtan T, Phienthrakul T (2019) Sentiment classification using document embeddings trained with cosine similarity. In: Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop. Association for Computational Linguistics, Florence, Italy, pp 407–414. https://doi.org/10.18653/v1/P19-2057

Tkaczyk D, Szostek P, Fedoryszak M, Dendek PJ, Bolikowski Ł (2015) CERMINE: automatic extraction of structured metadata from scientific literature. Int J Doc Anal Recogn (IJDAR) 18(4):317–335. https://doi.org/10.1007/s10032-015-0249-8

Tkaczyk D, Gupta R, Cinti R, Beel J (2018) Parsrec: a novel meta-learning approach to recommending bibliographic reference parsers. arXiv:1811.10369

Wang R, Liu W, McDonald C (2015) Using word embeddings to enhance keyword identification for scientific publications. In: ADC. Springer, pp 257–268

Ware M, Mabe M (2015) The STM report: an overview of scientific and scholarly journal publishing. Technical report, International Association of Scientific, Technical, and Medical Publishers

Xied Q, Dai Z, Hovy EH, Luong M, Le QV (2019) Unsupervised data augmentation. CoRR arXiv:1904.12848

Xu J, Zhang Y, Wu Y, Wang J, Dong X, Xu H (2015) Citation sentiment analysis in clinical trial papers. In: AMIA annual symposium proceedings, vol 2015. American Medical Informatics Association, p 1334

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems, pp 5754–5764

Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short Papers), pp 207–212