IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 25, 2021, accepted January 18, 2022, date of publication January 25, 2022, date of current version February 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3145954

SEAL: Semantically Enriched Authoring in LaTeX—A Model for Scientific Discourse

IMRAN IHSAN[1], MOHIB ULLAH[2], RAFI ULLAH KHAN[2], M. IRFAN UDDIN[3], ABDULLAH ALHARBI[4], AND WAEL ALOSAIMI[4]

[1]Department of Creative Technologies, FCAI, Air University, Islamabad 44000, Pakistan
[2]Institute Computer Science and Information Technology, The University of Agriculture Peshawar, Peshawar 25120, Pakistan
[3]Institute of Computing, Kohat University of Science and Technology, Kohat 2600, Pakistan
[4]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia

Corresponding author: Imran Ihsan (iimranihsan@gmail.com)

This work was supported by Taif University Researchers Supporting through Taif University, Taif, Saudi Arabia, under Project TURSP-2020/231.

ABSTRACT Semantic tags can enrich citation graphs by inter-connecting papers with citation reasons. One of the best sources of knowledge to tell the reason for citation is the author himself. Integrating these reasons in an authoring system can help authors to choose a reason while citing. We examined various Human and Automatic authoring systems for integration of citation reasons. However, to the best of our knowledge, no such system exists that facilities authors to integrate the reasons while citing. Same is the case with LaTeX Cite Packages. This research proposes integration of *CCRO*:Citations' Context and Reasons Ontology's (Ihsan and Qadir, 2019) taxonomic hierarchy of reasons within LaTeX document. We have developed a *CCRO* Package to semantically tag citations with reasons and to create an intra-discourse relation between research articles. Furthermore, embedding these structures within *RDF* Data Store enables the creation of knowledge graphs that become a foundation artifact for the Semantic & Scientific Discourse.

INDEX TERMS Semantic annotation, semantic authoring, scientific discourse, citation Indicators, citation behaviors, ontology.

I. INTRODUCTION

The reason to write any scientific scholarly document is to advance the accumulated knowledge in a verifiable way. Authors communicate this knowledge through literature review to form and present scientific claims along with their justifications. The most common method adopted by the authors to form the discourse and express in a document is via citation. Many researchers have discussed the rhetorical and argumentative nature of such discourse in the past by providing insights into why authors cite specific research and a need for a semantic-based research paper authoring tool that can help an author to choose semantic and meaningful tags for citation [22].

Semantic-based authoring can create an ecosystem to alleviate the information overload problem. The scientific community either use LaTeX or traditional processing systems for authoring research papers. Traditional processing systems include Microsoft Word, Google Docs, LibreOffice, Apple Pages, etc. The solution relies on enriching scientific publications with explicit rhetorical and argumentation discourse structures, using ontology by identifying and classifying citation texts within LaTeX files. So, the question is, how many authors use LaTeX typesetting language for authoring. To answer this, a study [2] was conducted to investigate the presentation and adoption of LaTeX across various disciplines. The data was extracted with the help of a paper called ''Don't Format Manuscripts'' [5]. Using Scopus/SciMago,[1] the total number of citable documents from 1996 to 2019 in the field of Mathematics, Physics, and Astronomy, and Computer Science, LaTeX articles are calculated. Table 1 shows the tabulated results.

Based on the study, around 27% researchers used LaTeX typesetting for authoring and an astonishing 11,930,976 citable documents are written using LaTeX in hard sciences. Another plus point for LaTeX typesetting is its availability as open-source as compared to the traditional processing system. Therefore, LaTeX typesetting is selected for

The associate editor coordinating the review of this manuscript and approving it for publication was Chang Choi.

[1]SciMago: https://www.scimagojr.com/countryrank.php

TABLE 1. Summary Statistics of LaTeX in Science Disciplines.

Discipline	Cited Documents	Rate	Documents
Mathematics	4,190,427	96.9	4,022,810
Physics & Astronomy	8,262,894	60.0	4,957,736
Computer Science	6,556,510	45.8	2,950,430
Total	19,009,831		11,930,976

the development of Semantic Publishing Ecosystem using explicit rhetorical and argumentation discourse structures. Furthermore, embedding these structures within *RDF* Data Store enables the creation of semantic publications that lay a foundation artifact for the Semantic Publishing Ecosystem (Knowledge Graphs) and linked resources to become part of the current Web of Data.

A. OUR CONTRIBUTIONS

To meet the requirements of Semantic Publishing Ecosystem and overcome the challenges, we develop SEAL, a Semantically Enriched Authoring in LaTeX framework to support citation reason annotation while authoring a research paper. SEAL brings together the following essential components to semantic authoring and publishing:

- Developed CCRO (Citations' Context and Reasons Ontology) LaTeX Package for Semantic Authoring and Publishing (Section IV-A);
- Enabled Semantically Enriched Authoring in LaTeX (Section IV-B);
- Established Semantic Publishing Process for automatic RDF generation. (Section IV-C);
- Established Semantic Citation Knowledge Graph (Section IV-D);
- Developed CCRO (Citations' Context and Reasons Ontology) Application for Semantic Querying (Section V).

Paper Organization: Section 2 outlines the two surveys performed as literature review, and its findings in detail. Section 3 presents the proposed methodology, based on which Section 4 describes four experiments and their outputs. Section 5 provides a discussion on the proposed semantically enriched authoring with a query as a proof of concept. Section 6 concludes the paper.

II. RELATED WORK

Scientometrics provides insight into scholarly documents and patterns within publications [21]. However, it provides a little support for the qualitative nature of scholarly writing. Authors cite other researches to make claims on their findings or to base their models on certain findings or simply contradict or negate results, commonly known as discourse analysis [22]. We have performed a survey towards the semantic authoring of scientific literature and is comprised of two parts. First, we have examined available semantic tools that provide mechanism form modeling and annotation of semantic authoring. Secondly, we have investigated different packages available for LaTeX that provide support for authors to integrate scientific discourse while authoring the research document.

A. AUTHORING TOOLS FOR CITATION REASONS

1) HUMAN AUTHORING AND ANNOTATION

One of the first applications to develop semantic hypertext for scholarly discourse was developed in 2001-04 and is known as "ClaiMaker[2]". "Claimaker" is the part of "ScholOnto" Project [9], [23] that provides a research prototype for usability testing, modeling and system development issues [9]. Based on "Claimaker" model, "ClaiMapper[3]" was developed. It is a visual based hypermedia tool that can store a claim in research paper in the form of semantic triple. These triples can later be interconnected to form a chain of complex nodes and structures. Similarly, in 2008 another tool knows as "Cohere[4]" was released. It is highly interactive and open source web interface using *RESTful APIs*. It provides facilities to tag semantic annotations such as problem, hypothesis, assumptions etc. using *RDF*. In 2015, *Research Articles in Simplified HTML - RASH Framework* [11] integrated *RDF* with *HTML* to provide set of specifications and a tool for academic authoring. *RASH* is a markup language that provides a restricted *HTML* with only *25* elements and a facility for validation, visualization, conversion, and enhancement. The *'conversion'* feature uses *XSLT* to convert a *RASH* document to LaTeX using *ACM ICPS* and *Springer LNCS* styles whereas '*evaluation*' feature uses Document Component Ontology (*DoCO*) for automatic annotation of markup elements to structural semantics. *RASH* provides easy to use mechanism for semantic annotations however it uses only `cito:cites` [20] for citation and does not provide any markup for scientific discourse.

All the above-mentioned tools except *RASH* provide semantic annotation for scientific discourse independent of the authoring environment. However, the semantic authoring process suggests enriching scientific publications with explicit linear, rhetorical, and argumentation structures while authoring a research publication [22]. In this context, one semantic authoring mechanism was proposed by *SALT* [14]. *SALT*, also known as Semantically Annotated LaTeX, provides plugins for both LaTeX and MS Word, where the author can manually annotate semantic tags while authoring the research document.

[2]Claimaker - `http://claimaker.open.ac.uk/`
[3]ClaiMapper - `http://compendium.open.ac.uk/institute`
[4]Cohere - `http://cohere.open.ac.uk`

LaTeX enables the authoring of documents using a high-quality typesetting system. It also provides a series of commands in a programmatic manner to produce the formatting and styling for the text. Due to its familiarity with authors to write content for publication and its capability for semantic authoring, LaTeX will be our focus of research instead of MS Word. Therefore, a deeper insight into *SALT*'s LaTeX plugin reveals that it provides three different types of annotations.

1) **SALT Rhetorical Ontology:** The author can correspond to a chunk of text as a rhetorical block using this ontology. Two tags are used to define the rhetorical block that is "`\begin{motivation}`" and "`\end{motivation}`".
2) **Elementary Discourse:** These items refer to a smaller sized text chunk along with rhetorical relations. It used LaTeX commands such as "`claim[ID]{ \ldots }`" and "`cause{CLAIM_ID:SUPPORT_ID})`'
3) **Argumentation Elements:** These elements also provide elementary discourse using the positioning of claims within a document. It uses "`position[ID][CLAIM_ID]{ \ldots })`" command for the said purpose.

All these discourse commands require the presence of identification elements such as `ID` or `CLAIM_ID` to create the rhetorical relations. The author, while writing the document using LaTeX, must create, manage, and track these IDs. Therefore, it might become difficult for authors because LaTeX doesn't provide such a provision.

Another application that allows semantic annotation of discourse relationships in the form of a hypothesis, claims, and evidence in the biomedical domain is known as the *SWAN Workbench* [8]. The application uses *RDF* Triples to model and store relationships. However, this application was later replaced by *SWAN* Annotation Framework that integrated text mining algorithms to override manual annotation.

2) AUTOMATIC AUTHORING AND ANNOTATION

Automatic annotation schemes normally use argumentative zoning to detect rhetorical blocks based on the author's language. One of the applications based on this principle is Xerox Incremental Parser (*XIP*) [1] to perform rhetorical analysis on scientific papers. *XIP*'s annotation for rhetorical blocks includes "Summarizing", "Background Knowledge", "Contrasting Ideas", "Novelty", "Significance", "Surprise", "Open Question" and "Generalizing". XIP labels the sentences with annotation tags more rigorously as compared to the reader of the document. However, the annotation list requires more rhetorical functions to describe research problems.

Another framework that provides automatic annotation of scientific discourse is the *SWAN Annotation Framework (AF)*. The framework works in conjunction with the *NIH-supported Neuroscience Information Framework (NIF)* [8]. *AF* is a three-tier application with the client-tier provides embedded web interface, the middle-tier provides text mining functionality and the data-tier provides persistence using *Annotation Ontology (AO)* [7]. Kindly note, the *SWAN Annotation Framework* does not use *SWAN* ontology and rather uses *Annotation Ontology* to make it orthogonal to any domain.

The Above survey reveals that only a handful of applications or researches are available that provide annotation for the scientific discourse nature of research articles. Furthermore, whether the application provides human annotation or automatic, the application only deals with the elements by finding claims and ideas within a single document. Scholarly activities evolve with a passage of time and there is a need to embed the inter-connected nature of scientific literature (citation reasons) at the time of authoring a research document. LaTeX provides "`\cite{paperID}`" command to cite another research and is widely used in all types of LaTeX based authoring tools. However, "`\cite{paperID}`" command only creates a hyperlink without any cognitive link between the citing and the cited paper. A semantic annotation integrated within a "`\cite{paperID}`" command can empower the author to integrate the context and reason to cite. There are some variations available for LaTeX "`\cite{paperID}`" command as well. Let's investigate some of its available variations to find if they provide any provision of semantic annotation for citation reasons or not.

B. LaTeX CITE PACKAGES FOR CITATIONS

1) CITE PACKAGE

The Cite Package is the most basic package for citation in LaTeX started in 1998. It is mostly intended for well-formed numeric citations [3]. The package only needs one command "`\cite{paperID}`" and is the natural behavior of LaTeX. However, there is hardly any documentation available for the complete package. Even Sebastian Rahtz, a long-term contributor to LaTeX typesetting when trying to provide support for it in "hyperref", had to give up trying to understand it [13]. But there are several packages developed based on the Cite Package.

2) HARVARD PACKAGE

The Harvard Package [25] qualifies citations by using the grammatical function of the label in the sentence and provides several commands. For example; when a citation is a noun, it uses "`\citenoun{paperID}`", and when something must be affixed, it uses "`\citeaffixed{paperID}`" command. The package also provides "`\citeyear{paperID}`", "`\citename{paperID}`" and "`\possessivecite{paperID}`" commands as well. However, no semantic-based command is available for citation reasons.

3) ACHICAGO PACKAGE

The Achicago Package [24], aimed at the Chicago Manual of Style, provides several bibliographic elements but doesn't use typeset quotations such as "`{}`" or

"\emp{}". It also provides multiple command such as "\citeNP{paperID}", "\citeYear{paperID}", "\citeN{paperID}", "\citeA{paperID}" etc. However, the main emphasis of the package is on what it is that the citation needs and not why the citation is made?

4) NATBIB PACKAGE

The Natbib Package [10] is the definitive word on author-year bibliography styles with LaTeX. Build upon the Harvard Package, it provides a set of customization possibilities. It provides various commands such as "\citep{paperID}", "\citet{paperID}", "\citeyearpar{paperID}" and "\citeauthor{paperID}" but does not provide plain "\cite{paperID}". Most of the LaTeX templates available on the internet, use the Natbib Package as a chosen family of styles for citation. However, the package just provides citation styling in a variety of formats and does not integrate the semantic nature of citation reasons.

5) APACITE PACKAGE

The Apacite Package [17] provides citations and references according to American Psychological Association rules. The package can be customized in several ways. The apacite citation commands include "\cite{paperID}", "\citeA{paperID}", "\citeAuthor{paperID}", "\citeYear{paperID}", "\citeNP{paperID}" and "\nocite{paperID}" etc. Using the Natbib package, it also provides support for Full and short author lists, masked citations, and ad-hoc citations. However, this package also does not provide any support for the semantic nature of citation reasons.

The analysis of the survey reveals that citation packages create multiple forms of citation styles by providing variations in the basic "\cite{paperID}" command. Survey also reveals that the most commonly used package in LaTeX templates is the Natbib Package. However, no package has integrated semantic or meaningful tags to define the context or reason of citation. Authors cite another research based on some context or reason. Therefore, it is evident to develop a citation package that can integrate functionality to empower the author, to add citation's context and reasons while citing other documents to create a semantic or cognitive link between the citing and the cited paper, and thus making it possible to study the evolutionary paths in the scientific literature.

III. PROPOSED METHODOLOGY

The complete process for Semantic Publishing Ecosystem spans over four steps, "Semantic Annotation", "Semantic Authoring", "Semantic Publishing", and "Semantic Graph". Fig 1 describes the methodology adopted.

A. SEMANTIC ANNOTATION

Based on our survey, it is evident to develop a citation package that can provide semantic annotation. Advance feature of LaTeX allows creation of *.ins* and *.dtx* files for creating and distributing classes and style files [19]. Using the *Natbib* Package, the *CCRO* (Citations' Context and Reasons Ontology) Package is designed integrating Citation's Context and Reason's Ontology object properties. "Citation's Context and Reasons Ontology – CCRO" [16] defines a taxonomic hierarchy of eight object properties distributed among three main sentiment-based reasons. The first three are sub-properties of positive, the next three are of negative and the last two are of neutral reasons. These properties are:

1) "*ccro:Incorporate*"
2) "*ccro:Extend*"
3) "*ccro:BasedOn*"
4) "*ccro:Negate*"
5) "*ccro:Crticize*"
6) "*ccro:Contrast*"
7) "*ccro:Compare*"
8) "*ccro:Discuss*"

To incorporate these eight properties, Semantic Citation commands are created that can be used in any available LaTeX Editor.

B. SEMANTIC AUTHORING

Authors typically use LaTeX to write their research articles. However, to the best of our knowledge, no repository publicly exists that houses LaTeX files. In order to simulate the authoring environment, a set of *40* random research papers is downloaded in *PDF* format from various sources. The selected articles come from four domains, with 10 highly representative papers from each domain. These domains are "H-Index", "Scientometrics", "Ontology" and "Sentiment Analysis". The corpus is then manually converted into LaTeX files using standard LaTeX template. The resultant is a collection of 40 LaTeX files that act as our input corpus.

Using the developed *CCRO* Package for semantic citation, all 40 LaTeX files in the selected corpus are manually converted into semantic-based LaTeX files. Kindly note, each LaTeX file is separately stored after the inclusion of semantic citation tags. In principle, the package enables authors to integrate semantic citation while authoring the paper and providing the reason why he/she is citing a paper. However, the basic concept of integrating a bibliographic entry in a LaTeX file adapts the same procedure as described in the world known as the *Natbib* package.

C. SEMANTIC PUBLISHING

After converting the selected LaTeX files into semantic LaTeX files, the collection can be referred to as "Semantic Corpus - SC". For Semantic Publishing, an application is developed that reads each semantic LaTeX file from the Semantic Corpus, automatically extracts semantic citations, and converts them into an *RDF* Triple. As *RDF* Triple is composed of "Subject – Predicate – Object", therefore each triple contains the Citing Paper as the "Subject", the Cite Paper as the "Object" and the selected *CCRO* Property as the "Predicate" as shown

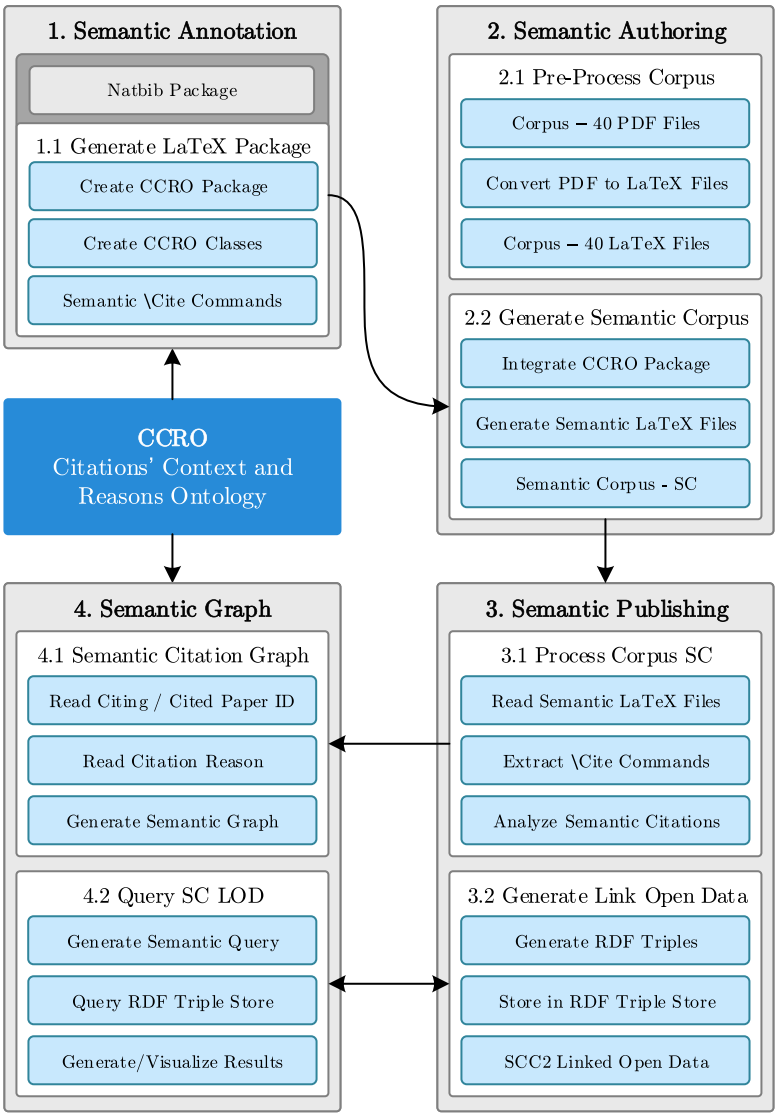


FIGURE 1. Steps towards Semantic Authoring and Publishing.

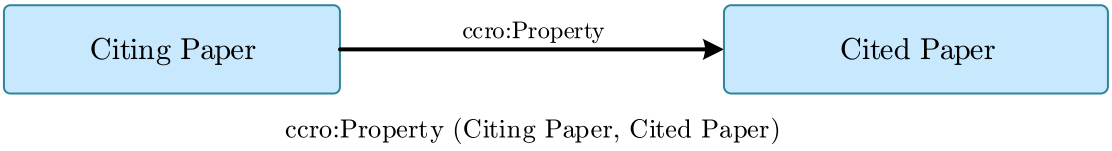


FIGURE 2. RDF Triple in RDF Data Store.

in Figure 2. This collection of all *RDF* Triples is known as *RDF Triple Store* that formulates *Semantic Corpus SC Linked Open Data*.

D. SEMANTIC GRAPH

Semantic Graph can store information in a rich, contextual, and conceptual construct. This construct is commonly called a 'triple'. Using the triples available in *Semantic Corpus SC Linked Open Data* semantic graph is then visualized. Generated *RDF Triple Store* is thus a semantic graph that may contain valuable information regarding how a scholarly activity has evolved during its lifecycle. To find the evolutionary paths between the scholarly activity, *SPARQL* queries are written and executed on *RDF Data Store*. The results are then visualized for discourse analysis.

IV. EXPERIMENTS AND RESULTS

To create an ecosystem of semantic authoring and publishing, four experiments are performed. First experiment is to create semantic annotation by creating the *CCRO* Package for LaTeX.

\citepos{Ihsan2019}{+1}	[Incorporate Ihsan et al., 2019]
\citepos{Ihsan2019}{+2}	[Extend Ihsan et al., 2019]
\citepos{Ihsan2019}{+3}	[Basedon Ihsan et al., 2019]
\citeneg{Ihsan2019}{-1}	[Contrast Ihsan et al., 2019]
\citeneg{Ihsan2019}{-2}	[Criticize Ihsan et al., 2019]
\citeneg{Ihsan2019}{-3}	[Negate Ihsan et al., 2019]
\citeneu{Ihsan2019}{=1}	[Discuss Ihsan et al., 2019]
\citeneu{Ihsan2019}{=2}	[Compare Ihsan et al., 2019]

FIGURE 3. CCRO Package for LaTeX - Syntax and Output.

\citeneg{J93-2003}{-2}	For example, the statistical word alignment in IBM translation models [Criticize Brown et al., 1993] can only handle word to word and multi-word to word alignments.
\citeneu{J93-2003}{=1}	Bilingual word alignment is first introduced as an intermediate result in statistical machine translation (SMT) [Discuss Brown et al., 1993].
\citepos{J97-3002}{+1}	In addition, Wu [Incorporate Wu., 1997] used a stochastic inversion transduction grammar to simultaneously parse the sentence pairs to get the word or phrase alignments.

FIGURE 4. A Semantic LaTeX Sample.

The second is to integrate these semantic tags in LaTeX files. The third is an application to automatically read semantic tags from LaTeX files and create *RDF* Data Store and the Last experiment is to visualize *RDF* Data Store (Semantic Citation Graph) among selected papers.

A. THE CCRO PACKAGE

The "CCRO" Package is an extension to LaTeX "\cite{paperID}" command by integrating semantic-based citations. The package is based on "Natbib" Package and is compatible with the standard bibliographic style files such as "harvard", "apacite" and "chicago" etc.

In contrast to other packages, the "CCRO" Package supports semantic tagging of citations. The Package uses Citation's Context and Reasons Ontology - *CCRO*'s constituent properties to create a meaningful tag between the citing and the cited paper. Like all other packages, it is required to be loaded in the document preamble such as

```
\usepackage{CCRO}
```

The document text itself begins with:

```
\begin{document}
\bibliographystyle{plainnat}
```

"plainnat" specifies the bibliography style used by the "BIBTEX" program to generate the actual bibliography from a database. The style "plainnat" is adapted from "natbib". However, any other bibliographic styles can be used instead of "plainnat"

To make a semantic citation in the text, the following commands are formed

\citepos{paperID}{+1} for "*ccro:Incorporate*".
\citepos{paperID}{+2} for "*ccro:Extend*".
\citepos{paperID}{+3} for "*ccro:BasedOn*".
\citeneg{paperID}{-1} for "*ccro:Contrast*".
\citeneg{paperID}{-2} for "*ccro:Criticize*".
\citeneg{paperID}{-3} for "*ccro:Negate*".
\citeneu{paperID}{=1} for "*ccro:Discuss*".
\citeneu{paperID}{=2} for "*ccro:Compare*".

Where \citepos command defines citations' reason classes in "Positive" context, \citeneg in "Negative" and \citeneu in "Neutral". Fig 3 defines the syntax and its output in detail. Though, using numbers as commands is not user-friendly, including complete names such as \citepos{PaperID}{Incorporate} becomes laborious for authors. For the future version of *CCRO* package, we are working on a smarter way to incorporate citation reasons.

B. SEMANTIC LaTeX

Semantic LaTeX is the extension of LaTeX writing environment that supports the semantic annotation of citations based on citations' context and reasons using the CCRO Package.

Semantic PDF

For example, the statistical word alignment in IBM translation models [Criticize Brown et al., 1993] can only handle word to word and multi-word to word alignments.

Bilingual word alignment is first introduced as an intermediate result in statistical machine translation (SMT) [Discuss Brown et al., 1993].

In addition, Wu [Incorporate Wu., 1997] used a stochastic inversion transduction grammar to simultaneously parse the sentence pairs to get the word or phrase alignments.

Semantic LATEX

```
For example, the statistical word alignment in IBM translation models
\citeneg{J93-2003}{-2} can only handle word to word and multi-word to
word alignments.

Bilingual word alignment is first introduced as an intermediate result in
statistical machine translation (SMT) \citeneu{J93-2003}{=1}.


In addition, \citepos{J97-3002}{+1}. used a stochastic inversion
transduction grammar to simultaneously parse the sentence pairs to get
the word or phrase alignments.
```

Semantic RDF

```
<ccro:C04-1005>       a                       ccro:CitingPaper     .
<ccro:J93-2003>       a                       ccro:CitedPaper      .
<ccro:J97-3002>       a                       ccro:CitedPaper      .

<ccro:citation#1>     a                       nif:string           ;
                      rdf:type                ccro:Citation        ;
                      biro:isReferencedBy     <ccro:C04-1005>      ;
                      biro:References         <ccro:J93-2003>      ;
                      nif:sentimentValue      -1                   ;
                      ccro:MainVerb           "handle"             .

<ccro:citation#2>     a                       nif:string           ;
                      rdf:type                ccro:Citation        ;
                      biro:isReferencedBy     <ccro:C04-1005>      ;
                      biro:References         <ccro:J93-2003>      ;
                      nif:sentimentValue      0                    ;
                      ccro:MainVerb           "introduce"          .

<ccro:citation#3>     a                       nif:string           ;
                      rdf:type                ccro:Citation        ;
                      biro:isReferencedBy     <ccro:C04-1005>      ;
                      biro:References         <ccro:J97-3002>      ;
                      nif:sentimentValue      +1                   ;
                      ccro:MainVerb           "use"                .

<ccro:C04-1005>       ccro:consistsOf         <ccro:citation#1>    ;
                      ccro:Criticize          <ccro:J93-2003>      .

<ccro:C04-1005>       ccro:consistsOf         <ccro:citation#2>    ;
                      ccro:Discuss            <ccro:J93-2003>      .

<ccro:C04-1005>       ccro:consistsOf         <ccro:citation#3>    ;
                      ccro:Incorporate        <ccro:J97-3002>      .
```

FIGURE 5. A Semantic Publishing Process.

Semantic LaTeX lets the author choose a reason from the available list while citing another research in the content of the research paper. The process is more robust than as defined by SALT [14]. In Semantically Annotated LaTeX - SALT, semantic annotation is provided as metadata in the RDF file along with PDF document using *Annotation Ontology*. However, for Semantic LaTeX, the author does not create a separate RDF file for metadata, rather the file can be automatically created along with the process to generate PDF. A sample of Semantic LaTeX using CCRO Package is shown in Fig 4. Using this technique all 40 research papers are initially converted into LaTeX before extending them into Semantic LaTeX.

C. SEMANTIC PUBLISHING PROCESS

The semantic publishing is an application that takes Semantic LaTeX documents as an input and creates a *RDF* file using the guidelines provided in *Citations' Context and Reasons Ontology - CCRO*. The transformation process consists of six steps, that are;

1) Read each Semantic LaTeX document from Semantic Corpus
2) Parse and Extract Semantic Citations
3) Analyze the Semantic Citations \cite commands
4) Fetch IDs for Citing and Cited Paper from Semantic Corpus

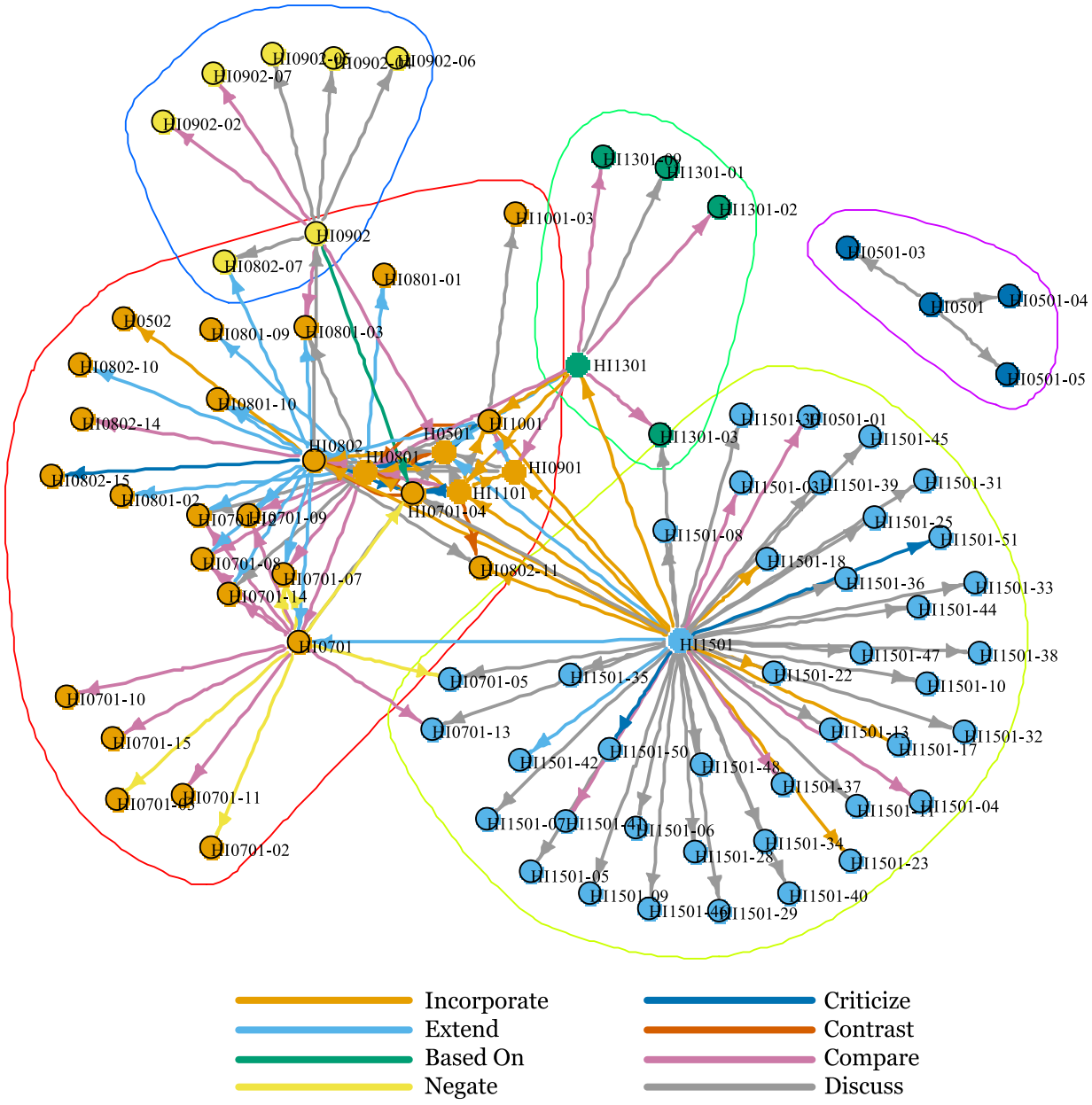


FIGURE 6. CCRO Based Semantic Citation Graph.

5) Generate RDF Files using schema defined in CCRO
6) Generate PDF Files using `CCRO` Package

Figure 5 shows a sample `Semantic` LaTeX document with its *Semantic RDF* and *PDF* counter parts. After the complete process on entire semantic corpus, an *RDF Triple Store* is generated. The application is developed using *Microsoft.NET Framework*. To generate *RDF* Files, an open source *.NET* library for *RDF*, known as "dotNetRDF[5]" is used. The library provides *APIs* for parsing, managing, querying and writing *RDF* and *RDF Triple Stores*.

[5]dotNetRDF - `https://www.dotnetrdf.org/`

D. SEMANTIC CITATION GRAPH

RDF Triple Store for `Semantic` LaTeX documents can be visualized as a semantic citation graph, describing a cognitive link between citing and cited paper. The visualization uses the citing paper and cited paper IDs from semantic corpus as nodes and the assigned CCRO class as the edge between these nodes. These edge classes are assigned different weights and colors. A partial representation of the visualized graph for the semantic corpus is shown in Figure 6.

V. DISCUSSION

With the advent of Knowledge Graphs, the research began for storing scientific data in large scale *RDF* format. One such

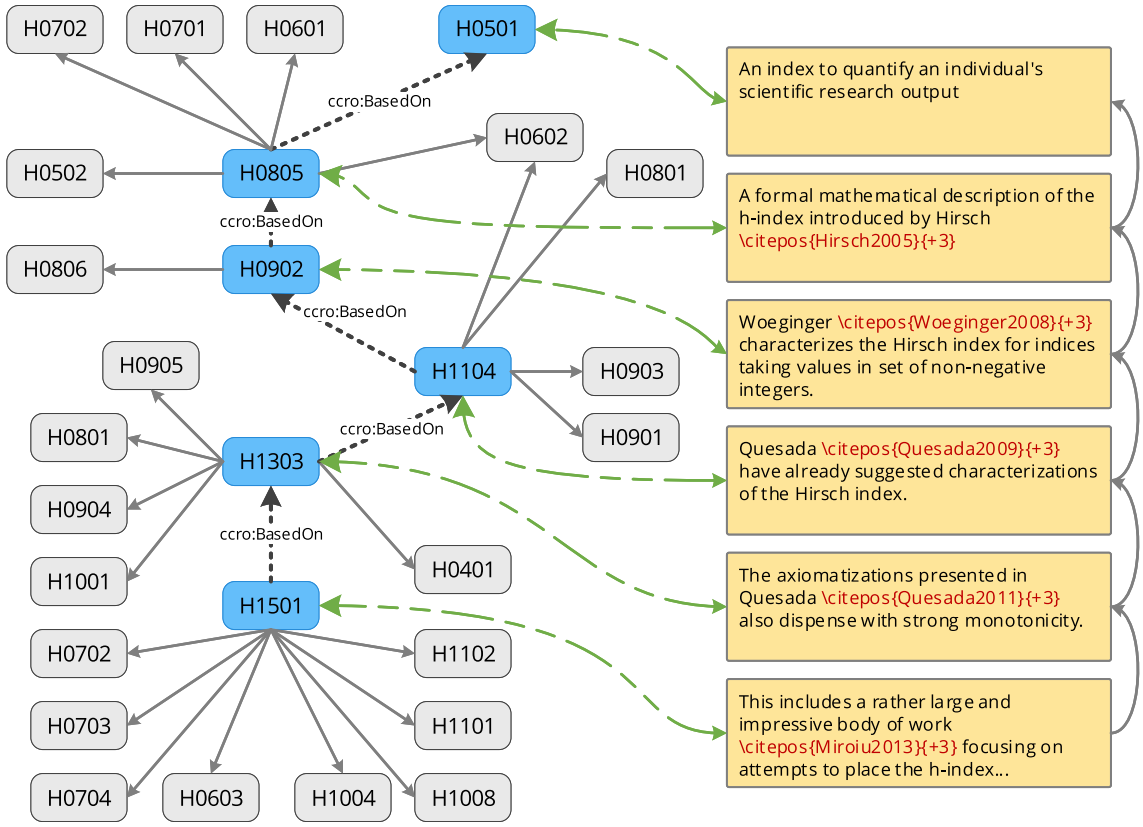


FIGURE 7. SPARQL Query: Results Visualization.

effort is the development of *Microsoft Academic Knowledge Graph (MAKG)* [18] with a huge volume of 8 billion triples and its availability on *Linked Open Data Cloud*. However, it requires the adaptation of various ontologies to encode different parts of a research article. For references, *MAKG* has modeled the citation information using a separate ontology *CiTO* [6]. Due to the coarse-grained 41 properties of *CiTO*, *MAKG* has only used one `cito:citation` as an entity type and leaving the rest. Though *MAKG* believes that citation context for each reference is valuable information for various tasks such as citation recommendation and citation-based paper summarization [18]. Therefore, a minimal set of cognitive-based citations' contexts and reasons in the form of an ontology becomes inevitable and its integration within a futuristic research paper authoring tool can help an author choose a semantic and meaningful tag for citation is an eventual outcome.

A scientific paper contains valuable information about the scholarly activity and its evolution. A citation graph has the potential to reveal important and interesting information about the history of particular scholarly research that has happened during its life-cycle. Using Semantically Enriched Authoring, LaTeX scientific papers become inter-connected with citation reasons. Using such semantic citation graphs, it is possible to infer the evolution of a research area over time, measure relations between research areas, and trace the influence of ideas that appear in the literature. For example, using *CCRO* package command `\citepos{paperID}{+3}` against "ccro:BasedOn", it becomes possible to find how a main algorithm or concept has started or evolved.

After the development of Linked Open Data for Semantic Corpus *RDF Triple Store*, *SPARQL* (A Query language for RDF Triple Stores) queries can be developed to look for the answers. To test the example, a well-known algorithm to measure both the productivity and citation impact of the publications of a scientist or scholar, knows as "h-index" is used. Two distinct papers are taken in the domain of Computer Science that represent the start and the current state of "h-index". These two papers and their assigned IDs are;

1) **H0501 –** "An index to quantify an individual's scientific research output", (2005) by J. E. Hirsch [15]
2) **H0801** – "Completing h", (2015) by Keith R. Dienes [12]

To find the path between these two research papers and the intermediate articles that have cited them, using only the *CCRO Package* command "`\citepos{paperID}{+3}`", the graph can be traversed starting from paper *ID: H1501* as the latest paper and paper *ID: H0501* as the starting paper. To query *RDF Triple Store*, "Virtuoso Universal Server[6]" is

[6]Virtuoso - `https://virtuoso.openlinksw.com/`

TABLE 2. SPARQL Query 1: Results.

?Paper	Paper Title for Information Only
HI0501	//An index to quantify an individual's scientific research output
HI0801	//An axiomatic characterization of the Hirsch-index
HI0901	//Monotonicity and the Hirsch index
HI1101	//Further characterizations of the Hirsch index
HI1301	//Axiomatizing the Hirsch index: Quantity and quality disjoined
HI1501	//Completing h

used. Virtuoso provides a middle-ware and database engine to load and query *RDF* data. Using the guidelines for the digital libraries, the *SPARQL* query is developed. This semantic query is:

```
PREFIX : <http://ccropus/resource/>
PREFIX ccro: <http://ccropus/ontology/>
SELECT ?paper
WHERE {
     ?p ccro:CitedPaper "H0501".
     ?paper ccro:Basedon* ?p.
}
```

Results of *SPARQL* Query are shown in Table 2 with its visualization in Fig 7 along with their citation texts in the selected path, for a deeper understanding and to help the author to see the evolution of an algorithm or research.

Automated research analysis using graph networks is now gaining popularity. Several pieces of research Galke2019 are available that generalize convolution to graphs to conduct experiments on representation in large-scale graphs. However, the purpose of this visualization is to show a proof of concept for the developed CCRO Package. Therefore, a simpler visualization technique is used.

VI. CONCLUSION

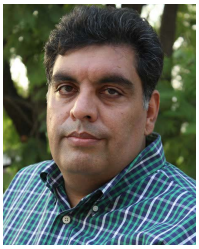
One of the best sources of knowledge to tell the reason for citation is the author of a paper at the time when he/she is writing the paper. Authors of the scholarly articles cite other articles based on certain reasons. We have developed a semantic annotation package for LaTeX to integrate citation reasons at the time of authoring a paper and this package can be integrated into any LaTeX authoring tool. With the help of this package, authors can tag a citation using suitable *CCRO* properties, making a simple LaTeX document as `Semantic` LaTeX document. Afterward, semantic citation tags embedded in a `Semantic` LaTeX document can be stored in a *RDF Triple Store* to formulate a semantically enriched citation graph using citations' context and reasons where Berners-Lee [4] vision of semantic web and giving meanings to hyperlinks can be adapted in its true essence for scholarly publishing.

Development of semantically enriched and machine-understandable citation graphs can become the foundation for many applications, such as the discovery of evolutionary paths in scholarly activity or finding influential papers within a certain domain. In future, we are working on one of the many possible applications of semantic authoring and publishing. This application defines an ontological model for scientific discourse by creating a scholarly knowledge graph enriched with citations' context and reasons. Through this application, it will become possible to extract evolutionary path by understanding the history of a scholar research.

REFERENCES

[1] S. Ait-Mokhtar, J.-P. Chanod, and C. Roux, "Robustness beyond shallowness: Incremental deep parsing," *Natural Lang. Eng.*, vol. 8, nos. 2–3, pp. 121–144, Jun. 2002.
[2] P. Alberto, "How many scholarly articles are written in LaTeX?" Authorea, Brooklyn, NY, USA, Tech. Rep., 2016, doi: 10.22541/au.148771883.35456290.
[3] D. Arseneau, "The cite package: Well formed numeric citation," TeXdoc Online, Tech. Rep., 2015. [Online]. Available: http://ctan.org/pkg/cite
[4] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Sci. Amer.*, vol. 284, no. 5, pp. 34–43, May 2001.
[5] F. Brischoux and P. Legagneux, "Don't format manuscripts," *Scientist*, vol. 23, no. 7, p. 24, 2009.
[6] P. Ciancarini, A. D. Iorio, A. G. Nuzzolese, S. Peroni, and F. Vitali, "Evaluating citation functions in CiTO: Cognitive issues," in *Proc. Eur. Semantic Web Conf.*, in Lecture Notes in Computer Science, vol. 8465, 2014, pp. 580–594.
[7] P. Ciccarese, M. Ocana, L. J. G. Castro, S. Das, and T. Clark, "An open annotation ontology for science on web 3.0," *J. Biomed. Semantics*, vol. 2, no. 2, Dec. 2011, Art. no. S4.
[8] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, and T. Clark, "The SWAN biomedical discourse ontology," *J. Biomed. Informat.*, vol. 41, no. 5, pp. 739–751, Oct. 2008.
[9] C. Mancini and S. J. B. Shum, "Modelling discourse in contested domains: A semiotic and cognitive framework," *Int. J. Hum.-Comput. Stud.*, vol. 64, no. 11, pp. 1154–1171, Nov. 2006.
[10] P. Daly, "Natural sciences citations and references," Texdoc.Net, Tech. Rep. 10.1.1.163.953, 2010.
[11] A. D. Iorio, A. G. Nuzzolese, F. Osborne, S. Peroni, F. Poggi, M. Smith, F. Vitali, and J. Zhao, "The RASH framework: Enabling HTML+RDF submissions in scholarly venues," in *Proc. 14th Int. Semantic Web Conf. (ISWC)*, 2015, p. 4.
[12] K. R. Dienes, "Completing h," *J. Informetrics*, vol. 9, no. 2, pp. 385–397, 2015.
[13] F. Garcia, "LATEX and the different bibliography styles," *PracTEX J.*, vol. 28, no. 2, pp. 1–14, 2007.
[14] T. Groza, S. Handschuh, K. Möller, and S. Decker, "SALT—Semantically annotated LATEX for scientific publications," in *The Semantic Web: Research and Applications* (Lecture Notes in Computer Science), vol. 4519. Berlin, Germany: Springer, 2007, pp. 518–532, doi: 10.1007/978-3-540-72667-8_37.
[15] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 46, pp. 16569–16572, Nov. 2005.
[16] I. Ihsan and M. A. Qadir, "CCRO: Citation's context & reasons ontology," *IEEE Access*, vol. 7, pp. 30423–30436, 2019.
[17] E. Meijer, "The apacite package," English, Tech. Rep. 10.1.1.610.5969, 2009.
[18] F. Michael, "The Microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data," in *The Semantic Web—ISWC 2019* (Lecture Notes in Computer Science), vol. 11779, C. Ghidini *et al.*, Eds. Cham, Switzerland: Springer, 2019, pp. 113–129, doi: 10.1007/978-3-030-30796-7_8.

[19] S. Pakin, "How to package your LaTeX package," Texdoc.Net, Tech. Rep. 10.1.1.169.89, 2015.
[20] S. Peroni, F. Osborne, A. Di Iorio, A. G. Nuzzolese, F. Poggi, F. Vitali, and E. Motta, "Research articles in simplified HTML: A web-first format for HTML-based scholarly articles," *PeerJ Comput. Sci.*, vol. 3, p. e132, Oct. 2017.
[21] R. M. Shiffrin and K. Börner, "Mapping knowledge domains," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5183–5185, 2004.
[22] S. B. Shum, T. Clark, A. D. Waard, T. Groza, S. Handschuh, and A. Sandor, "Scientific discourse on the semantic web: A survey of models and enabling technologies," *Semantic Web J., Interoperability, Usability, Applicability*, pp. 1–34, 2010.
[23] S. Buckingham Shum, E. Motta, and J. Domingue, "ScholOnto: An ontology-based digital library server for research documents and discourse," *Int. J. Digit. Libraries*, vol. 3, no. 3, pp. 237–248, Oct. 2000.
[24] M. Swift, "The achicago LaTeX package—Chicago manual author-date citations," Tech. Rep. 10.1.1.169.9515, 2001.
[25] P. Williams and T. Schnier, "The Harvard family of bibliography styles," Tech. Rep. 10.1.1.478.494, 1994.



IMRAN IHSAN received the Ph.D. degree in knowledge engineering from the Capital University of Science and Technology, Islamabad, in 2020. He has over 20 years of academic (teaching, project supervision, and academic administration) and industry (software, web, graphics, animation, and game development and operational management) experience. Since 2013, he has been working as an Assistant Professor with the Department of Creative Technologies, Faculty of Computing and AI, Air University, Islamabad. He has authored or coauthored research work in various conferences and journals. His research interests include knowledge engineering, semantic computing, natural language processing, and computational linguistics. For more information: log on to (www.imranihsan.com).

MOHIB ULLAH received the Ph.D. degree from the Capital University of Science & Technology Islamabad, Pakistan, and the M.S. degree from Birmingham City University, U.K. He is currently working as a Senior Lecturer at ICS & I.T., The University of Agriculture Peshawar, Pakistan. He has published 17 research articles in well-reputed journals and international conferences. His research interests include the security and privacy issues associated with computer networks, WSN, and the IoT.

RAFI ULLAH KHAN received the B.S. degree in computer science from the Islamia College, University of Peshawar (UoP), Peshawar, Pakistan, in 2007, the M.S. degree in internetworking and digital communication from the Institute of Management Sciences (IMlS), Peshawar, in 2010, and the Ph.D. degree in computer science from the Capital University of Science and Technology, Islamabad, Pakistan, in 2020. He has been working as a Senior Lecturer with the Institute of Computer Science & Information Technology, The University of Agriculture Peshawar, Pakistan, since 2011. His research interests include data mining, machine learning, web user privacy, sentiment analysis, and computer networks.

M. IRFAN UDDIN received the B.Sc. degree in computer science, the M.Sc. degree in computer science, the M.S. degree in grid computing, and the Ph.D. degree in computer science. He worked as a Postdoctoral Research Fellow. He worked as a Faculty Member at different institutes. He has been actively involved with academia and research. He is currently working with the Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan. He has published several articles in reputed journals and conference proceedings. He serves as a reviewer for different journals. His research interests include machine learning, data science, deep learning, convolutional neural networks, reinforcement learning, computer vision, and parallel programming.

ABDULLAH ALHARBI received the Ph.D. degree from the University of Technology Sydney, Australia. He is currently an Assistant Professor with the Information Technology Department, Taif University. His research interests include human–computer interaction, information systems, cybersecurity, and data science.

WAEL ALOSAIMI was born in Saudi Arabia, in 1979. He received the B.Sc. degree in computer engineering from King Abdulaziz University, in 2002, the M.Sc. degree in computer systems security, in 2011, and the Ph.D. degree in cloud security from the University of South Wales, in November 2016. From 2002 to 2004, he worked at Saline Water Conversion Corporation (SWCC) as an Instrument and Control Engineer. Then, he worked as a Trainer at Technical and Vocational Training Corporation (TVTC), until 2008. Next, he joined Taif University as a Teaching Assistant. It provides him with a scholarship to pursue his studies in U.K. Since 2017, he has been an Assistant Professor with the Computer Engineering Department, Taif University. He has many publications in peer-reviewed conferences and journals. His research interests include cloud computing, cloud security, information security, network security, e-health security, Internet of Things security, and data science.