DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

Corpus construction and mining for Citation Context Analysis

Danqun Zhao, Qianying Guo*, Hongpu Chen, Zhujuan Cai and Xiangyu Wang

Department of Information Management, Peking University, Beijing, China

ABSTRACT

Citation Context Analysis (CCA) is a typical data-driven research field based on full-text information, which breaks the limitations of traditional citation analysis using only bibliographic data, and benefits further studies on various citation behaviors and other core issues behind them, such as citation motivation, citation function and citation sentiment. Corpus for CCA is the most important guarantee and support for these issues. This paper attempts to discuss the corpus construction and mining for CCA in order to comprehensively review the research significance, research status and existing deficiencies in this area. Two main sections in our paper are:1) corpus construction for CCA, its three building tasks, such as citation sentence extraction, citation-reference mapping and citation context extraction, are discussed; 2)corpus mining and utilization for CCA, following related topics or situations are explored, including classification of citation motivation (or behavior) and citation sentiment, indexing and retrieval based on citation, citation recommendation and evaluation, citation-based abstracting and review generation automatically, and domains knowledge metrics. Finally, some suggestions and future research directions are briefly listed.

KEYWORDS

Citation Context Analysis; Citation Content Analysis; Citation Corpus; Citation Analysis

1 Introduction

As one of the core research fields of Bibliometrics, Scientometrics and Informetrics (call these different research areas "information metrics" and abbreviate it as "iMetrics" (Milojecvic & Leydesdorff, 2013)), citation analysis has been studied for more than half a century since it was founded by E. Garfield in 1960s. Looking back, the research development of citation analysis has gone through the following four stages: ①Citation Analysis 1.0, the stage of citation counting by using papers or their bibliographic elements as the quantitative analysis units or objects (before 1970s); ②Citation Analysis 2.0, the stage of clustering analysis of bibliographical relationships, such as bibliographical coupling and co-citation analysis (from 1970s to 1990s); ③Citation Analysis 3.0, the stage of citation network analysis by using complex network theory and SNA tools(since 2000s); ④Citation Analysis 4.0, the stage of citation context (or content) analysis based on full-text information (since 2010s).

Citation Context Analysis (CCA) is one of new research frontiers of citation analysis. By using the full text of a citing paper, CCA tries to obtain and use all citation information about

*Corresponding author: guoqianying@126.com

every reference listed in the end of the citing paper, such as citation position or section, citation frequency (or strength) and citation context, and make the quantitative analysis of citation content on a finer granularity. The synonymous terms of CCA include Citation Content Analysis (Zhang et al., 2013), Full-text Citation Analysis (Liu et al., 2013), etc. Although these terms are different in expression, there is no obvious difference in what they are referring to.

Booming and fast-growing of CCA are mainly due to the joint influence and promotion by the following factors: ①the popularity of full-text database makes it no longer difficult to obtain full-text corpus, which lays the data foundation for CCA; ②the progress of NLP technology, such as text mining, sentiment analysis, Named Entity Recognition (NER), Knowledge Graph(KG) and various advanced machine learning algorithms, has provided strong technical support for CCA; ③the promotion of Open Citation and Initiative for Open Citation ("I4OC"). "I4OC" advocates semantic publishing and citation opening, and is committed to using semantic Web technology to publish and open citation information in RDF format, so that it can be easily tracked and accessed like Web links information and can be understood and used by machines."I4OC" ensures that citation data is exposed and accessed unrestricted in more disciplines (or fields), and the OpenCitations Corpus (OCC) created based on SPAR Ontologies (OpenCitations, 2020) can gradually alleviate the long-standing problems of citation data, such as being difficult to parse, inconvenient to track continuously and unable to be understood by machines, until it is completely solved.

Due to the limitation of bibliographic data, traditional citation analysis is always difficult to accurately identify various citation behaviors and their hidden (implied) motivations, purposes and sentiments behind them, and also difficult to effectively judge ecological environment and quality of citations in academic literature collection. The advent of CCA has greatly broken the constraint of traditional research and become a leading direction in the field of NLP-based Bibliometrics (Amjad et al., 2013).

As a kind of typical data-driven research, CCA's corpus construction is the most important guarantee and support for its research. Through the in-depth cognition and complete mining of the implied value of CCA corpus, it is not only to maximize the value of the corpus, but also to lead the innovation and development of citation analysis both in theory and methodology. This paper attempts to comprehensively discuss the construction, mining and utilization of CCA corpus, which is mainly divided into two sections: one is the construction or building of CCA corpus, and its three building tasks, such as citation sentence extraction, citation-reference mapping and citation context extraction, are analyzed respectively; the other is the mining and utilization of CCA corpus, five related research topics or situations are discussed, including classification of citation motivation (or behavior) and citation sentiment, indexing and retrieval based on citation, citation recommendation and evaluation, citation-based abstracting and review generation automatically, and domains knowledge metrics.

2 Corpus Construction for CCA

Before discussing the corpus construction for CCA, let us clarify the basic concepts and terminology related to it.

①Reference & Citation. These are two closely related concepts based on the citing and/or cited relationships among academic papers. They are also a pair of basic concepts in the entire field of citation analysis and many other concepts are defined or derived from them. Generally, "reference" refers to a cited paper, which usually appears in the reference list (at

the end or after the body of a paper) and is described or annotated in the standard format; "citation" refers to a phrase, clause or sentence where the citation marker (or reference anchor) is located when a reference is cited or mentioned in the body of a citing paper. There is a many-to-many relationship between them, that is, a reference has at least one or more citations in the body of a citing paper, and a citation will also correspond to or be associated with one or more references in its reference list. Due to this, for "reference", there are Unitary Mentioned Reference (UMR, a reference that be mentioned only once in a citing paper) and Multiple Mentioned Reference (MMR, a reference that be mentioned more than once in a citing paper); for "citation", there are Unitary Reference Citation (URC, a citation only includes one reference) and Multiple References Citation (MRC, a citation includes more than one reference), as shown in Figure 1 (Lin et al., 2019).

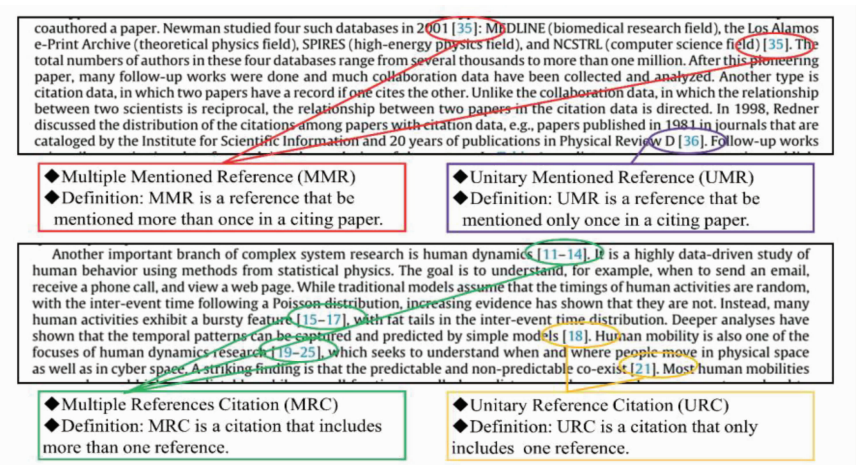

Figure 1 The Definition and example of MMR, UMR, MRC, URC

It should be noted that only bibliographic data can be used in early citation analysis while "citation" is often regarded as a synonym of "reference" (or cited paper). After entering the stage of full-text citation analysis (4.0), the differences (or semantic differences) between them gradually become clear. Many scholars have had in-depth discussions on "reference", "citation" and their semantic differences (Wouters, 1999), and the formation of the reference tradition and the establishment of the citation mechanism based on them have become the institutional guarantee for citation analysis. Robert K. Merton (1988), the founder of American sociology of science, emphasized that-- "We thus begin to see that the institutionalized practice of citations and references in the sphere of learning is not a trivial matter. While many a general reader--that is, the lay reader located outside the domain of science and scholarship--may regard the lowly footnote or the remote endnote or the bibliographic parenthesis as a dispensable nuisance, it can be argued that these are in truth central to the incentive system and an underlying sense of distributive justice that do much to energize the advancement of knowledge." Merton's view laid an important foundation for establishment of the Normative Theory of citation.

②Citation sentence. Nakov et al. (2004) first used a new term "citance" (abbreviation of "citation sentence") in 2004 to refer to the sentence surrounding the citation marker in a citing paper. "Citation sentence" can be understood in both narrow and broad sense, in which the narrow one is the sentence itself where the citation marker is located, and the broad can

be extended to the sentence where the citation marker is located and its surrounding text. In this paper, a complete sentence with a citation marker is taken as a citation sentence (or "explicit citation sentence"). Obviously, a citation sentence contains one or more citations, and each citation is associated with one or more references.

③Citation context. Small (2011) defined "citation context" as the text surrounding the references, which means the text content around the citation marker when a reference is cited in a paper (synonymous with the aforementioned generalized "citation sentence"). For the convenience of research, it is usually necessary to set a citation window to identify or extract citation context. Terms with the same meaning or similar to the "citation context" include "citation area", "citation site", "scope of influence of citation", and "citation statement", etc. In our paper, we use the term "citation context" uniformly and take the following understanding: the remainder of a specific citation window which removed the (explicit) "citation sentence" and some other sentences unrelated to this citation sentence. Ideally, the remainder has no sentences (or text) unrelated to this citation sentence, that is, all remaining sentences (or text) in the given window have a high semantic similarity with the specific citation sentence.

Normative Theory of citation believes that citation system is one of academic norms and basic professional standards that need to be consciously abided by scholarly community. Citation behavior can be regarded as an active way of communication (also known as "formal communication") in academic activities, which is important to maintain knowledge accumulation and discipline development from inheritance and transcendence diachronically to supplement and enrichment synchronically. From this theoretical perspective, citation, reference, citation sentence and citation context are all indispensable source of CCA corpus. Therefore, a fully functional CCA corpus should focus on the following three tasks for its construction: citation sentence extraction, citation-reference mapping and citation context extraction.

2.1 Citation Sentence Extraction

Citation sentence extraction is the primary task of CCA's corpus construction, and it is also the basis of the latter two. Generally speaking, the writing of peer-reviewed papers has strict requirements or description standards for the annotation of references (in or after the body of a paper). Therefore, it is not too difficult to identify and extract citation sentences from the full-text of citing papers in most cases, especially from those with structured full-text (such as XML format). Of course, if a paper with structured full-text cannot be obtained directly, extracting its citation sentences will be relatively difficult because such full-text needs to be parsed and preprocessed.

In fact, due to the different writing habits of authors, the diversity of description standards and differences in citation styles, etc. there are still many detailed problems to be solved in accurately identifying and extracting citation sentences from the full text, and the complexity of these details cannot be ignored. One of the most important problems is the identification of citation markers (or reference anchors), so we need to investigate citation styles first. The widely used citation styles are as follows: Numbered (which use numbers or other abbreviations to refer to an entry in the reference list) and Author-Date (which use an "author-year" pair to uniquely identify an entry in the reference list, also known as "Harvard Style"). Different journals or publishers make some adjustments for themselves based on these citation styles. For example, different conventions are involved in the Numbered style, such as whether numbers are superscripted? How to represent consecutive numbers? And how to

select the separators between discontinuous numbers? etc. Powley & Dale (2007) have made a more comprehensive investigation on the citation styles used in journal papers and found that there are five kinds of citation styles more representative (see Table1).

Table1 Some examples of citation styles

Citation styles	Examples
Textual Syntactic	**Levin (1993)** provides a classification of over 3000 verbs according to…
Textual Parenthetical	…are WordNet **(Miller et al., 1990)** and Levin classes **(Levin, 1993)**
Prosaic	**Levin** groups verbs based on…
Pronominal	**Her** approach reflects the assumption…
Numbered	…of behavior among verb groups [1]

Citation sentence extraction also frequently meets the problems of informal citations or implicit (non-explicit) citations (Powley & Dale, 2007; Qazvinian & Radev, 2010), which are more common especially when the Harvard style is used for marking (or annotating) citations. For example, two cases in Table 1 (the third and fourth), there are no obvious citation markers, instead of a person's name or personal pronoun used to describe or cite a specific reference. Many studies confirmed that such implicit citations often contain richer semantic information and have higher value in citation analysis (Athar & Teufel, 2012). Our paper adopts a narrow understanding of citation sentence, so here only focuses on extraction of explicit citation sentences, while the identification and extraction of implicit citation sentences will be discussed in Section 2.3.

After a citation sentence extracted, it can be stored in a database table and the fields of the table are as follows: citation sentence number (unique), citation sentence type, citation sentence content (text) and citation sentence source, etc. The "source" field can be further subdivided into several subfields to comprehensively record its number of citing paper, numbers of chapter (or section) position, paragraph and sentence where the citation sentence is located in the given citing paper.

2.2 Citation–Reference Mapping

The second construction task of CCA corpus is to scan each citation sentence stored in the database table and make mapping between each citation in the sentence and its corresponding reference (s) to form a citation-reference mapping record until all citation sentences are processed. Here the key problem is still the parsing of citation markers. Obviously, if a citation sentence contains only one URC citation, one citation-reference mapping record will be created; if a citation sentence contains more than one URC or MRC citations, multiple citation-reference records will be created. When all citation-reference mapping records of all citation sentences in a citing paper are stored in the database table, their metadata of citations for a citing paper can be aggregated. The main fields in citation-reference mapping table include: citation sentence number (unique), reference number (unique), reference author, reference title, reference publication year, reference source, etc. Furthermore, if reference (s) and citing paper are from the same literature database, more fields of reference(s), such as abstract, keywords, and author institutes, etc., can be considered to write into citation-reference mapping records.

Citation-reference mapping table is one of the important parts of CCA corpus, which can

provide effective supplementary or supporting information for accurately understanding their content and semantics of citation sentences, and evaluating rationality or interdisciplinary of citation behaviors, etc.

Normally, a citation appears in the body of a citing paper with reference anchor(s), and the corresponding reference (s) appears at the end of the citing paper (corresponding to the reference list). The common way to establish a citation-reference mapping record is to use regular expression to identify the various styles of citation markers located in the citation sentences. Figure 2 gives a simple example for the "Author-Date" citation style (Harvard style). It can be seen that how to accurately extract fields of each reference corresponding to citation (such as author, title, publishing year, journal, conference, etc.) from the reference list is quite complicated. When the Harvard style is used for citation marking, many NER problems will be involved due to the different abbreviations, translations and personal pronoun anaphora of author names, all which are fairly difficult to identify and parse accurately.

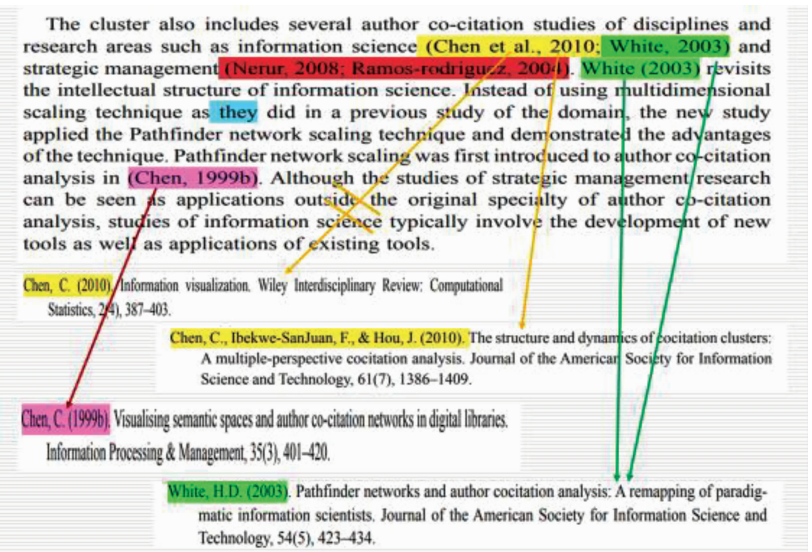

Figure 2 Examples of "Citation-Reference" Mapping

2.3 Citation Context Extraction

The third construction task of CCA corpus is to identify and extract its citation context around a citation sentence, or it can be understood as extraction of implicit citations. Current extraction strategies of citation context can be divided into the following three types: ①taking the whole paragraph where the citation sentence is located as its context; ②selecting a fixed number of sentences before and after the citation sentence as its context, or calculating the physical distances between them and the citation sentence, adding them different weights, and then considering whether to use them as context; ③calculating the semantic similarity between the citation sentence and its surrounding sentences (in given citation window), and selecting the sentences with high similarity as its context.

Among these strategies, the first two are simple and easy, while the third is the most ideal that requires a large computational cost and time consumption (deeply rely on NLP algorithms). Literature investigation found that there is no consensus on how to extract citation

context: the first two strategies are extensively used to resolve this problem; some directly extract citation sentences as an alternative; and a few researches also began to pay attention to the third strategy which tried to make elastic adjustments to the number of extracted sentences to improve extraction quality of citation context. For example, through calculating semantic similarity between each sentence (in the given citation window) and the citation sentence, the sentence(s) with high similarity (more than a certain threshold) but not necessarily adjacent to the citation sentence can be identified. Finally, the actual extraction result of context can contain different number of sentences by removing irrelevant text to the citation sentence and be divided into different types accordingly, such as No Context, Only Before (the citation sentence), Only After(the citation sentence), Both Before and After (the citation sentence).

Obviously, the challenge and complexity of accurately and completely extracting citation context far exceed the first two construction tasks of CCA corpus. Some researches specialized in the challenge tasks are in progress (Lei et al., 2016), and many other studies, such as citation function classification (Teufel et al., 2006a; Teufel et al., 2006b), citation summarization (Qazvinian & Radev, 2008; Qazvinian & Radev, 2010) and automatically generating review (Nanba et al., 1999; Nanba et al., 2011), also involved massively in it. Results of all these studies have further confirmed that CCA corpus can not only using (explicit) citation sentences. A lot of important information about author's attitude and comments on references (or cited papers) often appear somewhere around the citation sentences and their context information is very important for many CCA research topics.

2.4 Some Related Discussion

Though CCA has become active recently, related discussion focusing on its corpus construction is still rare. Three construction tasks discussed above are all important parts of a complete CCA corpus, among them extracting citation sentence is most critical, and the difficulty of citation-reference mapping and citation context extraction remains incrementally. They are all related to a large number of NLP technical issues, such as Named Entity Recognition (NER), anaphora resolution, definition and calculation of sentence similarity, dictionary building for clue words, and knowledge representment and extraction (for example semantic triple of SPO), etc., which require help of machine learning algorithms. Due to limited space, these issues can only be discussed in another paper.

We believe that the first choice for massive construction of CCA corpus is biomedical field in the current situation, some reasons list as follows: ①The full-text of academic papers in this field has a high degree of open access, especially the structured full-text. For example, only PubMed Central (PMC) has collected more than 3 million articles (in XML format), which is not only free and open, but also easy to analyze and use, which undoubtedly provides sufficient and high-quality sources for corpus construction. ②Research tools are abundant and relatively mature. There are many thesauri and their supporting tools, such as MeSH, UMLS and Medical Text Indexer (to extract medical subject terms), MetaMap (to extract UMLS concepts) (NLM, n.d.); Second, there are Semantic MEDLINE Database (SemMedDB) (Kilicoglu et al., 2012) which stored and represented in SPO triples and its supporting SPO extraction tool Batch SemRep (Kilicoglu et al., 2012); Third, there are tools of extracting citation sentences, such as Colil for extracting PMC citation sentences only and free AI tool Semantic Scholar for extracting beyond PMC. In addition, medical document feature modeling BioBERT (Lee et al., 2020) based on deep learning algorithm and the knowledge graph based on BioBERT (Xu et

al., 2005) are also in open and available. ③A large number of concepts, entities or knowledge objects (such as diseases, drugs, tissues/organs, genes, medical instruments, etc.) existing in this field can be used for bibliometric analysis, and semantic relations and their meanings among these concepts, entities or objects are also very rich and clear, which can provide enough research topics and application scenarios for CCA. Obviously, building its CCA corpus in biomedical field has significant benefits.

3 Corpus mining and utilization for CCA

Citation sentences and their context related to a specific research paper are very valuable because they are peers-reviewed text and have rich evaluation information among them. Especially for highly cited papers, such corpus of text continuously accumulated after a period of publication time has a huge amount of information and utilization value, which is worthy of in-depth mining and utilization. The corpus construction for CCA, on the one hand, can ensure the comprehensive extraction, analysis and storage of this kind of valuable text; on the other hand, it can effectively remove a large number of redundant fragments from full text and facilitate mining and efficient utilization.

After carefully considering and evaluating value of such citation corpus, five important research topics focusing on CCA corpus mining and utilization are listed and discussed as follows.

3.1 Classification of Citation Motivation (or Behavior) and Citation Sentiment

Accurate classification of citation motivation (or behavior) and citation sentiment is the primary research task of CCA, and it is also a difficult problem staying in the field of citation analysis for a long time. CCA corpus derived from full-text of papers makes it possible to solve above problems or achieve breakthroughs and plays an important role in reasonably differentiating weight or impact of every citation, modifying and improving research hypothesis and theory of citation analysis. At the same time, CCA corpus has also laid a solid foundation for a series of further researches, such as evaluation of quality, rationality and ecological environment of citations, as well as citation content-based academic evaluation, etc.

3.1.1 Classification of Citation Motivation (or Behavior)

Psychologists believe that motivation is an internal psychological process or impetus that leads to, inspires and maintains individual activities by goals or objects. The generation of motivation is mainly based on needs at various levels. Citation motivation is a kind of social motivation, mainly due to the needs of academic research. Furthermore, various motivational theories in psychology hold that motivations are the basis of most human behaviors, and there is a close relationship between motivations and behaviors. Therefore, citation behaviors can be regarded as an externalized expression of citation motivations. For the sake of convenience of discussion, the following does not make a strict distinction between them.

Citation motivations (or behaviors) is key to whether the research hypothesis of citation analysis is tenable and whether its theoretical basis is complete, so it has attracted great attention of scholars as early as 1960s. Through the investigation and observation, Garfield (1964), founder of citation analysis, first summarized citation motivations appearing in the writing of scientists' academic papers into the following 15 types: 1) paying homage to pioneers; 2) giving credit for related work (homage to peers); 3) identifying methodology, equipment, etc.; 4) providing background reading; 5) correcting one's own work; 6) correcting the work of others; 7) criticizing previous work; 8) substantiating claims; 9) alerting to forth-

coming work; 10) providing leads to poorly disseminated, poorly indexed, or uncited work; 11) authenticating data and classes of fact-physical constants, etc.; 12) identifying original publications in which an idea or concept was discussed; 13) identifying original publication or other work describing an eponymic concept or term as, e.g. Hodgkin's Disease, Pareto's Law, Friedel-Crafts Reaction, etc.; 14) disclaiming work or ideas of others (negative claims); 15) disputing priority claims of others (negative homage). Since then, many scholars have had more discussions on citation motivations (or behaviors) from different dimensions and different classification frameworks (Bornmann & Daniel, 2008; Brooks, 1985; Thorne, 1977; Weinstock, 1971). In recent years, domestic scholars have also introduced ecological perspective for classifying motivations of citations into different types, such as parasitism, mutualism, competitive symbiosis, commensalism, amensalism and irrelevant symbiosis (Li & Liang, 2012).

According to different research methods of citation motivations(or behaviors), existing related works or research outputs can be summarized into the following four categories: ①experience judgment of fields experts for simple classification of citation functions and motivations, most of early studies falls in this category; ②using questionnaire survey and interview to understand users' real citation motivations or behaviors; ③empirical research based on small-scale scientific literature data sets to verify or improve existing citation classification models; ④automatically identifying citation motivations (or behaviors). Among them, the fourth is one of the frontier topics in the CCA research which especially relies on the construction of large-scale citation corpus, and can effectively make up for the shortness and defects of the first three types of research.

Further literature investigation shows that the fourth type of research can also be roughly divided into rule-based, statistical-based and ontology-based methods. Among them, rule-based method is simple and effective, but it is time-consuming and laborious to build a base of rules manually by experts in advance, not easily shifting across different fields also leads to its poor flexibility; statistical-based method mainly uses various machine learning algorithms (such as Naïve Bayes, N-gram, Support Vector Machine, etc.) to train classifiers, which needs to build a manually annotated citation corpus in advance; ontology-based method needs to build an ontology using for citation classification description, there is now only Semantic Publishing and Ontologies (SPAR) can be compatible with ontology frameworks such as Citation Typing Ontology (CiTO) (Iorio, 2013; Shotton, 2010) and can be used for reference in practice.

In short, classification of citation motivations (or behaviors) is confronted with the difficulty of how to "enter the author's head", which a lot of problems of distinguishing psychological cognition and emotional attitude involves in. Now, various automatic recognition or classification methods are commonly encountered such corners as follows: inconsistent classification frameworks (no consensus or little agreement has not reached about it), poor corpus quality (small corpus, inaccurate and incomplete extraction of citation sentences and context, etc.), and algorithm limitations or heavy burden of manual annotation. For the future, it is necessary to strengthen the integrated utilization of different research methods, and how to construct a more comprehensive citation classification framework (or model) through extensive investigation or by reference to ontology tools such as CiTO is also key path in order to complete automatic classification of citation motivations (or behaviors) with higher accuracy according to some valuable cue words extracted from the citation sentence and its context.

Finally, it should be emphasized that explorations on citation motivations (or behaviors)

have always been as a central topic in field of citation analysis throughout all stages from 1.0 to 4.0. If summarizing all these studies theoretically, two schools or cliques for cognition about citation motivation gradually formed as below: one is the Normative Theory of citation, emphasizing that citation system is the academic norm abided by all scientists in practice, thus citations from peers represent obtaining recognition, and more citations represent more recognition. So, citation analysis as a method can be used to evaluate achievements and impacts of scientists and their works. The second is Social Construction of citation, which only regards citation as a rhetorical device (it has nothing to do with Merton's social norms theory). It holds that citations among papers are a kind of information utilization behaviors taken by individuals because of their perceived needs, and citation motivation is a complex, uncertain and private operation with certain propensity, so the usefulness of citation analysis is questionable. Many scholars have made a lot of theoretical discussions and experimental analyses around these topics (Baldi, 1998; Collins, 1999; Nicolaisen, 2003; Nicolaisen, 2007; Small, 1978; Small, 1998), some of them have tried to put forward new citation theories (Small, 2004; Nicolaisen & Frandsen, 2007). Up to now, although the two theories have their own achievements, their views or opinions about citation are in sharp opposition. How to eliminate or balance their disputes and seek their combination in the future has become an important research mission of CCA.

3.1.2 Classification of Citation Sentiment

Research for citation motivation affected by many subjective and objective factors is extremely complex, which has a natural difficulty in its accurate recognition. In contrast, the distinction of citation sentiment or emotions not only depends on the recognition of citation motivation, but also on the accurate extraction and correct understanding of the cue words representing various emotional attitudes in the citation corpus, which is also very challenging and difficult same as study for citation motivation.

Sentiment analysis is the task of identifying positive and negative opinions, sentiments, emotions and attitudes expressed in text. Although there has been a growing interest in this field in the past few years for different text genres such as newspaper text, reviews and narrative text, relatively less emphasis has been placed on extraction of opinions from scientific literature, more specifically, citations (Athar, 2011). Different from the highly personal emotional commentary texts on hot topics (or events) published by Web users, emotional expression is mostly considered to be relatively neutral when it comes to literature citation in academic papers, such as citing some facts or data, or objectively introducing the design idea and working principle of an algorithm. Only when it comes to the subjective evaluation of previous research, the more implicit and euphemistic emotional expression (positive or negative) will appear in the writing. CCA corpus (mainly involving citation sentences and their context) has rich evaluation information related to author's emotional expression for the cited paper. Therefore, the analysis or classification of citation sentiment can mainly use this part of the CCA corpus, but the difficulty and complexity of recognition of citation sentiment has been greatly increased undoubtedly due to the author's relatively cautious and careful wording.

The classification of citation sentiment has been involved in the research of citation context extraction and citation motivations (or behaviors) classification, etc. (Teufel et al., 2006a; Teufel et al., 2006b), and some studies adopted artificial methods directly (Yu, 2014).With rapid development of NLP technology and related machine learning algorithms, the research on automatic classification of citation sentiment is gradually developed. Its basic procedure

or main steps for experimental study can be described as follows:

①Construct or establish classification model (framework) of citation sentiment. One of the primary important challenges is how to define the implied sentiment in the citation corpus, including two dimensions of sentiment polarity and sentiment strength. Generally, the determinants of citation sentiment polarity are mostly related to citation motivation, and can be simply divided into three categories: positive, negative, and neutral; while citation sentiment strength can be divided into strong and weak. The elements combination of the two dimensions can form a preliminary citation sentiment classification model.

Similar to the classification of citation motivation belonging to the same category of psychological activities, there is no consensus classification framework for citation sentiment, especially the fine-grained framework. Therefore, a classification model that meets the requirements of citation sentiment recognition task can be constructed by referring to the relevant research results of citation motivation recognition.

②Based on the sentiment classification model and citation corpus, manually annotating citation sentiment and creating a training set by using annotation results. In fact, because most of the emotional expressions in citation corpus are implicit and euphemistic, the subjectivity of manual annotation is inevitable. How to ensure the quality and consistency of results of emotional annotation needs to be paid more attention to.

③Compile a dictionary of clue words for citation sentiment recognition. The compiling of the dictionary of clue words is a complicated NLP task, and there is no available one for citation sentiment recognition yet at present. HowNet (CNKI) is a universal emotional dictionary which is difficult to use directly because its coverage is not enough for fully covering specific fields. Therefore, it is necessary to start with the selection of seed words (mostly adjectives) and complete the compilation through continuous expansion of the set of seed words. It is worth noting that emotional words are not only adjectives, some nouns, adverbs, negative words and even transition conjunctions are all useful and important ones. For example, adverbs may be an important basis for judging sentiment strength, while transition conjunctions may directly determine or change the sentiment polarity of a sentence. So, it is very important for recognition task of citation sentiment in given field to building a fitting dictionary of clue words with high coverage (or wide range).

④Design or apply classification algorithm to complete the task of citation sentiment classification. The common machine learning algorithms which can be selected to use mainly include Naïve Bayes, Support Vector Machine (SVM), etc. According to the emotional score of each word in the emotional Dictionary (to be preset after compiling), emotional scores of all clue words existing in given citation sentence or its context can be used for calculating emotional score of the whole citation sentence or its context. Finally, all citation sentences and their context can be classified into different emotional categories by using their emotional scores respectively.

Deep learning algorithms have been developed and applied rapidly on various NLP tasks during recent years, but related research on citation sentiment recognition by these algorithms has not been founded easily. Relative lag of corpus construction for CCA is one of the major constraints because it is very difficult to estimate a large number of parameters that these deep learning algorithms demanded on existing small-scale citation corpus.

3.2 Indexing and Retrieval Based on Citation

Automatic indexing by using CCA corpus can optimize traditional indexing methods and

its results, and lay a foundation for citation retrieval, especially for citation context retrieval. Furthermore, the implementation of indexing and retrieval based on citation will accelerate the research of citation analysis.

3.2.1 Citation Indexing

The indexing value of citation corpus is mainly reflected in whether new keywords or subject words reflecting paper's content, theme and academic contribution can be extracted from such text or corpus, so as to add and enrich indexing terms obtained from its title, abstract and keywords in a given paper to a certain extent.

The research steps of citation indexing can be described as follows: ①selecting target papers (usually some highly cited papers) as a sample set of papers(D); ②for each paper dj in D, gathering all citation sentences and their contexts from each main body of its citing papers (set), and further extracting all keywords (set CKj) from these citation sentences and their contexts; ③comparing keywords in CKj with the original keywords of paper dj (set Kj, to be usually composed of keywords from title, abstract and some descriptors written by authors of dj), to find out whether or not new keywords are extracted in CKj? How many and their quality of these new keywords? Can they reveal the content or theme of the target paper as enriched indexing terms? ④Repeating steps② and③until all sample papers in D are finished.

It has been found that some new keywords with good indexing value can usually be extracted from citation corpus, and they are very useful for optimizing the characterization and disclosure of important contents of the target literature (Zhang et al., 2017). Generally speaking, the higher the cited times of a target paper, the more abundant citation sentences and contexts can be obtained, and correspondingly, the greater the possibility of finding additional index terms.

3.2.2 Citation Context Retrieval

The problem about citation context retrieval (or retrieval based on citation context)is described briefly below: according to the user's query or his/her search input of words, returning immediately online some citation sentences matched with this query or searching words in citing papers, and detailed information for each hit citation sentence in the answer set also includes its context (before or/and after the hit), position in the paper (for example IMR&D), all references occurred in the hit (with co-citation relationship), etc.

Objectively speaking, citation context retrieval does not require too many technical breakthroughs. It only needs to extract feature terms with index value (or significance) from citation sentences in CCA corpus and organize them into inverted files in order to matching user's queries. Hu (2016) has designed and implemented the SOS (search of sentence) system in his doctoral dissertation which search results can provide each hit citation and its following citation information: position (of chapter or section), context and all co-cited references. It is not only convenient for users to understand more citation details and realize the progress from "what to cite" to "how to cite", but also lay a research foundation for more fine granular co-citation analysis (for example, co-citation analysis at the level of single citation sentence, or single paragraph or section of papers, etc.).

Semantic Scholar, a free search tool launched by Allen Institute for AI (AI2) in 2016, also provides some citation retrieval services, such as References (referenced by this paper), Citations (citing this paper), and Figures, Tables and Topics (search for charts and their titles in given paper). According to recent visit and investigation, Semantic Scholar has collected nearly 200 million papers in its database, which can be used as a citation extraction tool in

addition to search service. Compared with Colil (DBCLS, n.d.), an OSS (open source software) that only aims at extracting citation sentences from abstract of articles in PMC, Semantic Scholar can be applied to more literature databases except for PMC, such as Microsoft Academic, Springer, Nature and ArXiv, etc.

3.3 Citation Recommendation and Evaluation

Citation recommendation is mainly devoted to providing timely and effective help for the author's academic paper writing and successful publication, including immediate recommendation and evaluation recommendation. The former serves for the author's academic paper writing, and provides a list of suitable citing reference(s) related to the current content written online by author, while the latter serves to comprehensively evaluate the quality of references cited in the paper after finishing it, including relevance, comprehensiveness, academic quality of these references; citation ecological environment of the written paper, and gives some comments (such as reservation/ deletion, ecological health/sub-health/pollution, etc.) or provides important references that failed to cite or omitted by author(s). Both the former and the latter need to be based on an accurate understanding of the content or semantic relationship between the cited papers (references) and citing papers, especially the former, which is more dependent on high-quality citation corpus, and is also an important source of high-quality citation corpus. At present, some service systems with citation recommendation function have appeared. For example, after uploading a paper in PDF format or its URL in Website, such system like Citeomatic, can return a list of references already cited by author (s) and a list of papers which are deserved to be cited by Citeomatic but not yet cited by author(s).

Citeomatic is a typical citation evaluation and recommendation tool. Although its function is limited (no immediate recommendation now), there is a large room for service enrichment or function expansion for expected citation evaluation in future, especially for academic ecological assessment based on CCA corpus. For example, given citing paper(s)(single or collective) and then obtaining all citation sentences and their positions occurring in the citing paper(s), the overall judgment about the single paper or macro monitoring of ecological environment about the collection of papers can be considered comprehensively by analysis of multi-dimensional influencing factors, including citation motivations (or behaviors), citation relevance(combined with using citation-reference mapping records in CCA corpus), quantity and quality of their citations, etc. Finally, some further comments on the citing paper(s) can be concluded, or evaluation results of classification can be given, like "Health", "Sub-Health", "Pollution", etc. This undoubtedly has a great influence and guidance on creating healthy academic environment (or atmosphere), developing academic research activities orderly and improving current mechanism of academic evaluation drastically.

3.4 Citation–Based Abstracting and Review Generation Automatically

An abstract refers to a brief and accurate description of the content of a document (or a document unit), and usually does not include the supplement, explanation or comment to the original. There are various types of abstracts, two kinds of manually writing ones are informative abstract and indicative abstract, while automatically generating ones can be divided into more types according to different standards. For example, according to whether original text is used, it can be divided into Full text-Based Abstract (FBA) and Citation-Based Abstract (CBA); according to whether it is oriented to specific users, it can be divided into

Generic Abstract and Biased Abstract, and the latter can be further subdivided into different subtypes, such as biased document themes, biased user interests and biased user's queries; according to the number of documents processed, there are Single Document Abstract (SDA) and Multiple Document Abstract (MDA). In particular, when the number of documents to be automatically summarized reaches or exceeds a certain threshold level, MDA can be regarded as automatically generating survey or review. Finally, the following four kinds of abstracts based on different research strategies are often referred to, such as statistical-based abstract (or excerpt), NLP-based abstract, information extraction-based abstract and text structure-based abstract.

As one of the important research topics emerged in the field of NLP since 1950s, automatically abstracting has always been focused on the generation technology and method of FBAs for a long time. Although these FBAs (including the author's abstracts) can better reflect the content of the original text, their ability to summarize the influence of the original text is relatively limited, and they can't reflect the diachronic changes of the literature influence after publishing (Mei & Zhai, 2008). So, CBA emerged gradually as a new research and exploration direction since 2008 (Qazvinian & Radev, 2008). It mainly learns from theory of citation analysis in Bibliometrics and uses the citation corpus (especially citation sentences and their contexts) to form a generalization or understanding about main content of a paper to be abstracted which can reflect its academic influence or value from the views of other peer's researches.

All existing theoretical and empirical analyses show that CBA has many advantages compared with its FBA (Bradshaw, 2003; Elkiss et al., 2008; Kan et al., 2002; Mohammad et al., 2009). Especially for the highly cited papers, their citation abstracts are more obvious in objectivity and diversity. In fact, their citation corpus is more general, extended and critical after ever-increasing accumulation and processing by more and more academic peers which can better reflect the significant parts of the original text.

A typical research task for CBA in early time can be described as below: given a paper to be abstracted, gathering its citing papers completely and extracting all citation sentences and their contexts (and regard as a set of citations); then selecting some citation sentences from the set to generate its citation abstract, and ensure that these selected sentences as a subset has sufficient compression rate and good generalization ability. Its main research steps (or key issues involved in)are listed as follows: ① selecting an appropriate full-text database (for its coverage and availability); ②identifying and extracting citation sentences (in broad or narrow sense); ③ classifying and screening these citation sentences by identifying their types and purposes; ④organizing or ranking citation sentences to form an abstract (as a draft); ⑤ post-processing of abstract draft, such as de-duplication, coherence processing for sentences, etc.; ⑥evaluating abstract.

Obviously, CBA's study is closely related to the task of construction of CCA corpus. The key points and difficulties are mainly derived from the second, third and fourth steps. The existing problems are the lack of structured full-text corpus (except PMC), accurate identification or extraction of citations and contexts in non-fixed citation window, semantic de-duplication and reasonable ranking of citation sentences, and new design for suitable evaluation indicators and schemes (such as for evaluation of sentence coherence). Up to now, many kinds of CBAs still focus on the generation of single document abstract (SDA). It should be considered how to extend from SDA to multiple document abstract (MDA) and to generate literature review automatically, which will lead to all old problems, such as classification/cluster-

ing, de-duplication and ranking of citation sentence, becoming more difficult and challenging. Furthermore, the combination of CBA and FBA is also an important research topic because CBA will be greatly restricted or impracticable for mostly low cited papers in academic collections due to the lack of their citation corpus.

3.5 Domains Knowledge Metrics

Metrics and analysis of scientific knowledge has attracted wide attention in recent year, and CCA corpus has become an important guarantee for such researches which are promoting development of iMetrics to Knowledge metrics gradually. Two basic problems need to be solved in knowledge metrics: one is knowledge representation, which defines the knowledge entity (or object) and its representation method for quantitative analysis, that is, determining the basic knowledge unit; the other is meta-knowledge representation, which is used to represent its sources and cognitive states for a given knowledge entity (object) or basic knowledge unit, like known, unknown, etc.

Take an example of the field of medicine. To solve the problem of knowledge representation, there are now three representative research achievements: ①SemMedDB of National Library of Medicine (USA) (Kilicoglu et al., 2012), which transforms the free text describing medical knowledge that can be understood by human into "Subject-Predication-Object" (SPO) triples that can be understood easily by machine, and maps all concepts and semantic relations involved in SPOs to the UMLS (Universe Medical Language System). ②Nano publication model, proposed by The Netherlands Bioinformatics Centre (NBIC), and here "nano publication" refers to the smallest and machine-readable publication unit with scientific significance (Groth et al., 2010). ③ Micro publication model proposed by Harvard Medical School (Clark et al., 2014). Among the three representations (or models) of medical knowledge described above, SPO Triples of SemMedDB is the most influential and most concerned. SemMedDB implements extraction and storage of knowledge units (SPO triples) in very large scale (the latest version contains nearly 100 million SPO triples extracted from abstract of papers in database of PubMed), but has the disadvantage of ignoring or missing most of the meta-knowledge of each SPO triple. This involves the second basic problem, the representation of meta-knowledge. Here, meta-knowledge mainly refers to some scientific judgments related to the cognitive state for a given SPO triple, such as whether the knowledge represented by the SPO triple is pure research hypothesis, or the conclusion of new experiments, or even only objectively citing previous scientific assertions (or opinions), etc. Scientists usually express their academic opinions through carefully selected words in the text, or give a more rigorous explanation and conclusion for it. This type of information or clues expressed in the text, which can provide the certainty (or uncertainty) degree of the knowledge represented in a SPO triple, is abundant in citation sentences or context, especially in medical field (Chen et al., 2018; Murray et al., 2019). Therefore, the representation of meta-knowledge can be considered starting from some valuable clue words in citation sentences or context. By accurately identifying and extracting these clue words and adding them to their corresponding SPO triples, a complete representation of scientific knowledge can be formed and then used for quantitative analysis and evaluation of different knowledge units.

Most researches on the topic of knowledge metrics heavily focuses on the field of biomedicine, and uncertainty measurement of medical knowledge is particularly active among them. For example, a series of studies completed by H. Small et al. (Small & Klavans, 2011; Small, 2018; Small et al., 2019; Small, 2019), proposed a new method to identify scien-

tific breakthroughs in scientific literature by combining co-citation analysis and citation context, and made empirical analyses from various aspects by the indicator of hedging rate. Here, "hedging rate" is defined as the proportion of citation sentences containing words "may", "could" or "might" in all citation sentences in a paper. It is mainly used to quantify the uncertainty of scientific knowledge contained in different types of papers and their citation sentences.

Besides these, the team of Chen Chaomei has also made fruitful findings on the measurement of knowledge uncertainty in the medical field. At the end of 2017, they published their work——"Representing Scientific Knowledge: The Role of Uncertainty" (Chen & Song, 2017), proposed that uncertain information should be regarded as the meta-knowledge of scientific proposition (SPO triples), and that hedging words can't cover all aspects of knowledge uncertainty, emphasized that scientific knowledge in the state of inconsistency, contradiction or controversy is an important driving force for the emergence of new paradigm or scientific reform. H. Small commented that "uncertainty is key to understanding the development of scientific knowledge", "opens up a new area in the study in Scientometrics and Informetrics as well as information visualization, namely the study and measurement of uncertainty of scientific knowledge and how uncertainty is expressed in scientific texts". Domestic scholars such as Du Jian also regard knowledge uncertainty as an important research front (Du, 2019; Du, 2020), and are committed to promoting the development of medical knowledge metrics.

Taking "uncertainty" as the core of topic, our paper believes that CCA and its corpus can also make more researches on domains knowledge measurement or metrics, here are some examples:①analysis of spatial-temporal evolution of knowledge (or visualization), combined with the publication time and authors' affiliations (institutions or countries/regions belong to, etc.) of papers associated with SPO triples, we can observe the change (curve) of knowledge uncertainty from dimensions of time and/or space, and then make judgment on the maturity, development and evolution, and key turning points of scientific knowledge in given triples;②knowledge interaction /cross- analysis, exploring the transformation relationship between scientific knowledge (in papers) and technology (in patents), or between scientific knowledge (in papers) and clinical treatment (in patients' cases). ③construction and analysis of meta-knowledge graph. Here, the meta-knowledge graph (MKG) uses each SPO triple as its node and the relationship between triples of SPOi and SPOj as its edge, it can be understood as a new high-level knowledge graph established on the basis of underlying knowledge graph, in which each "S" or "O" represents different node and each "P" as an edge of nodes. In a specific MKG, the meaning of edge of (SPOi, SPOj) can be defined by using their co-occurrence or co-citation relationship while taking the frequency of co-occurrence or strength of co-citation as its weight of the edge. Furthermore, the underlying knowledge graph (composed by knowledge entities and their semantic relations) and its meta-knowledge graph (MKG) can be integrated to build a two-layer (or even multi-layer) heterogeneous network. It is very noteworthy for the feasibility and significance of information science for studying such two-layer (or multi-layer) heterogeneous network in the future.

4 Conclusion

CCA is booming and developing rapidly at home and abroad during recent years. On the one hand, the large-scale and high-quality corpus of citations is far from being built and completed now, and the existing tools of extracting citation sentences can only provide very

limited data support for CCA; on the other hand, some basic issues, such as the classification of citation motivation and sentiment, etc., have not yet reached a research consensus, more research topics of CCA are waiting to be explored and expanded, which obviously fails to promote and drive corpus construction at once. In fact, the corpus construction and mining for CCA are two intertwined and closely related issues. This paper only discussed them from a macro and holistic perspective. In the future, on the basis of clarifying basic concepts of CCA, it is necessary to make more pragmatic and technical discussion on more detailed levels in close combination with our research jobs supported by fund of NSFC in order to promote the continuous and in-depth development of citation analysis in open full-text era.

References

Abu–Jbara, A., Ezra, J., & Radev, D.(2013). Purpose and Polarity of Citation: Towards NLP–based Bibliometrics. In: *Proceedings of the Main Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 596–606.

Athar, A.(2011). Sentiment Analysis of Citations using Sentence Structure–Based Features. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 81–87.

Athar, A., & Teufel, S.(2012). Detection of implicit citations for sentiment detection. In: *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, 18–26.

Baldi, S.(1998). Normative versus social constructivist processes in the allocation of citations: A network–analytic model. *American Sociological Review*, 829–846.

Bradshaw, S.(2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. *Research and Advanced Technology for Digital Libraries*, 499–510.

Brooks, T. A.(1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science, 36* (4), 223–229.

Bornmann, L., & Daniel, H.(2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64* (1), 45–80.

Chen, C., Song, M., & Heo, GE.(2018). A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. *Journal of Informetrics, 12* (1), 158–80.

Chen, C., & Song, M.(2017). *Representing Scientific Knowledge: The Role of Uncertainty*. Springer.

Clark, T., Ciccarese, P. N., & Goble, C. A.(2013). Micropublications: A semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics, 5* (1), 28.

Collins, H. M.(1999). Tantalus and the aliens: Publications, audiences and the search for gravitational waves. *Social Studies of Science, 29* (2), 163–197.

Du, J.(2019). An Automated Approach for Extracting Uncertain Clinical Knowledge from Published Medical Documents. In: *Proceedings of the 2019 Tianfu International Forum on Scientometrics and Research Evaluation*, Chengdu, China.

Du, J.(2020). Measuring Uncertainty of Medical Knowledge: A Literature Review. *Data Analysis and Knowledge Discovery, 46*, 14–27.

Elkiss, A., Shen S., & Fader, A.(2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology, 59* (1), 51–62.

Garfield, E.(1964). Can citation indexing be automated? *Statistical association methods for mechanized documentation, symposium proceedings*. National Bureau of Standards, Miscellaneous Publication 269, Washington DC, 189–192.

Groth, P., Gibson, A., & Velterop, J.(2010). The anatomy of a nanopublication. *Information Services & Use, 30*, 51–56.

Hu, Z.(2016). F*ull–text Citation Analysis: Theory, Method and Application*. China Science Publishing & Media Ltd.

Iorio, A. D., Nuzzolese, A. G., & Peroni, S.(2013). Towards the Automatic Identification of the Nature of Citations. In: *Proceedings of 3rd Workshop on Semantic Publishing*, 63–74.

Kan, M.Y., Klavans, J. L., & Mckeown, K. R.(2002). Using the annotated bibliography as a resource for indicative summarization. In: *Proceedings of LREC*, 1746–1752.

Kilicoglu, H., Rosemblat, G., Fiszman, M., & Shin, D.(2020). Broad–Coverage Biomedical Relation Extraction with SemRep. *BMC Bioinformatics, 21* (1), Article No.188.

Kilicoglu H, Shin D, Fiszman M., & Shin, D.(2012). SemMedDB: a PubMed–scale repository of biomedical semantic predications. *Bioinformatics, 28* (23), 3158–3160.

Lee, J., Yoon, W., & Kim, S.(2020). BioBERT: a pre–trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36* (4), 1234–1240.

Lei, S., Chen H., Huang, Y., & Lu, W.(2016). Research on Automatic Recognition of Academic Citation Context. *Library and Information Service, 60* (17), 78–87.

Li, Z., & Liang, Y.(2012). The ecology explanation on citation motivation. *Studies in Science of Science, 30* (4), 487–494.

Lin, G., Hou, H., & Hu, Z.(2019). Understanding Multiple References Citation. In: *Proceedings of 17th International Conference on Scientometrics & Informetrics*, 2347–2357.

Liu, X., Zhang, J., & Guo, C.(2013). Full–text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology, 64* (9), 1852–1863.

Mei, Q., & Zhai, C.(2008). Generating Impact–Based Summaries for Scientific Literature. *Association for Computational Linguistics*, 816–824.

Merton, R. K.(1988). The Matthew Effect in Science II. Cumulative Advantage and the Symbolism of Intellectual Property. *ISIS, 79* (299), 606–623.

Milojecvic, S., & Leydesdorff, L.(2013). Information metrics(iMetrics): A research specialty with a socio–cognitive identify? *Scientometrics, 95* (1), 141–157.

Mohammad, S., Dorr, B., & Egan, M.(2009). Using citations to generate surveys of scientific paradigms. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* 584–592.

Murray, D., Lamers, W., Boyack, K., Larivière, V., Sugimoto, C. R., & Van Eck N. J. (2019). Measuring disagreement in science. *17th International Conference on Scientometrics and Informetrics*, 2370–2375.

Nakov, P. I., Schwartz, A. S., & Hearst, M.(2004). Citances: Citation sentences for semantic analysis of bioscience text. In: *Proceedings of the SIGIR´04 workshop on Search and Discovery in Bioinformatics*, 81–88.

Nanba, H., & Okumura, M.(1999). Towards multi–paper summarization using reference information. *International Joint Conference on Artificial Intelligence*, 926–931.

Nanba, H., Kando, N., & Okumura, M.(2011). Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online, 11* (1), 117–134.

Nicolaisen, J.(2003). The social act of citing: Towards new horizons in citation theory. In: *Proceedings of the American Society for Information Science and Technology, 40* (1), 12–20.

Nicolaisen, J.(2007). Citation analysis. *Annual review of information science and technology, 41* (1), 609–641.

Nicolaisen, J., & Frandsen, T. F.(2007). The handicap principle: a new perspective for library and information science research. *Information Research, 12* (4), 12–14.

Powley, B., & Dale, R.(2007). Evidence–based information extraction for high accuracy citation and author name identification. In: *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 618–632.

Qazvinian, V., & Radev, D. R.(2008). Scientific paper summarization using citation summary networks. In: *Proceedings of the 22nd International Conference on Computational Linguistics*, 689–696.

Qazvinian, V., & Radev, D. R.(2010). Identifying non–explicit citing sentences for citation–based summarization. In: *Proceedings of the 48th annual meeting of the association for computational linguistics,* 555–564.

Shotton, D.(2010). CiTO, the citation typing ontology. *Journal of Biomedical Semantics, 1* (Suppl 1), S6.

Small, H.(1978). Cited documents as concept symbols. *Social Studies of Science, 8* (3), 327–340.

Small, H.(1998). Citations and consilience in science–Comments on theories of citation?. *Scientometrics, 43* (1), 143–148.

Small, H.(2004). On the shoulders of Robert Merton: towards a normative theory of citation. *Scientometrics, 60* (1), 71–79.

Small, H.(2010). Referencing through history: how the analysis of landmark scholarly texts can inform citation theory. *Research Evaluation, 19* (3), 185–193.

Small, H.(2018). Characterizing highly cited method and non–method papers using citation contexts: The role of uncertainty. *Journal of Informetrics, 12* (2), 461–480.

Small, H.(2019). What makes some scientific findings more certain than others? A study of citing sentences for low–hedged papers. In: *Proceedings of the 17th International Conference of the International Society for Scientometrics and Informetrics*, 554–560.

Small, H., Boyack, K. W., & Klavans, R.(2019). Citations and certainty: a new interpretation of citation counts. *Scientometrics, 118* (3), 1079–1092.

Small, H., & Klavans, R.(2011). Identifying scientific breakthroughs by combining co–citation analysis and citation context. In: *Proceedings of 13th International Conference of the International Society for Scientometrics and Informetrics*, 783–793.

Teufel, S., Siddharthan, A., & Tidhar, D.(2006a). An annotation scheme for citation function. In: *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, 80–87.

Teufel, S., Siddharthan, A., & Tidhar, D.(2006b). Automatic Classification of Citation Function. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 103–110.

Thorne, F. C.(1977). Citation index–another case of spurious validity. *Journal of Clinical Psychology, 33* (4), 1157–1161.

Weinstock, M.(1971). *Citation indexes, Encyclopedia of Library and Information Science*. New York: Marcel Dekker.

Wouters, P. F.(1999). The citation culture. [Doctoral dissertation, University of Amsterdam]. UvA–DARE(Digital Academic Repository). https://garfield.library.upenn.edu/wouters/wouters.pdf

Xu, J., Kim S., Song M., & Jeong, M.(2020). *Building a PubMed knowledge graph*. Retrieved from https://arxiv.org/abs/2005.04308

Yu, B.(2014). Automated Citation Sentiment Analysis: What Can We Learn From Biomedical Researchers. In: *Proceedings of the American Society for Information Science and Technology*. https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/meet.14505001084

Zhang, G., Ding, Y., & Milojevic, S.(2013). Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content. *Journal of the American Society for Information Science and Technology, 64* (7), 1490–1503.

Zhang, S., Liang, M., & Cao, G.(2017). Research on Subject Extraction of Scientific and Technical Documents Based on Citation. *Information studies: Theory & Application, 40* (6), 122–127.