Characterizing Knowledge Entity Extracted from Citation Sentences

Dongin Nam
Library and Information Science
Yonsei University
Seoul, Republic of Korea
dongin.nam@yonsei.ac.kr

Jiwon Kim
Library and Information Science
Yonsei University
Seoul, Republic of Korea
jihas203@yonsei.ac.kr

Jeeyoung Yoon
Library and Information Science
Yonsei University
Seoul, Republic of Korea
jeeyoungyoon9@yonsei.ac.kr

Chaemin Song
Library and Information Science
Yonsei University
Seoul, Republic of Korea
chaemin11@yonsei.ac.kr

Seongdeok Kim
Library and Information Science
Yonsei University
Seoul, Republic of Korea
ystetjdejr@yonsei.ac.kr

Min Song†
Library and Information Science
Yonsei University
Seoul, Republic of Korea
min.song@yonsei.ac.kr

ABSTRACT

This paper proposes a novel entitymetrics analysis by exclusively focusing on citation sentence. Since citation sentence offers both citing and cited author's research interest, knowledge entity that appears in this sentence can be considered as a key entity. To characterize such key entities, we conduct an entitymetrics analysis on citation sentences that are extracted from full-text research articles collected from PMC. We use "opioid" as our search query since it is an actively studied domain, which indicates that rigorous amounts of knowledge entities and entity pairs are available for examination. After which, we construct two novel citation sentence-based networks, namely the direct citation sentence (DCS) network and the indirect citation sentence (ICS) network. The DCS network is built upon direct entity pairs that are captured within citation sentences. The ICS network, on the other hand, utilizes indirect entity cooccurrences based on cited author information that appears inside a citation sentence. To demonstrate the usefulness of the DCS and ICS network, a conventional full-text network is formed for comparison analysis based on network features and opioid-related bio-entity pairs. The results show that DCS and ICS network demonstrate distinct network characteristics and provide unobserved top-ranked bio-entity pairs when compared to traditional method. This indicates that our method can expand the base of entitymetrics and provide new insights for knowledge structure analysis.

CCS CONCEPTS

• **Information systems** → *information extraction;*
• **Information systems applications** → *knowledge entity extraction*

KEYWORDS

Entitymetrics; Citation sentence; Direct Citation Sentence Network; Indirect Citation Sentence Network; Network analysis; Opioid; Knowledge structure

1 Introduction

In accordance with exponential increase of scientific publication, importance of knowledge entities as a means to extract meaningful and structured knowledge from mass literature is growing. Entitymetrics is an approach enabling entity level analysis on scientific literature and related researches are being actively conducted since its proposal in 2013 [1][2][3]. Entitymetrics was initially suggested to utilize article title and abstract for knowledge structure analysis [1]. While a lot of entitymetrics related studies had a tendency of focusing on title and abstract [1][2][3][4][5], several entitymetrics studies attempted to conduct full-text data [6][7][8][9] based research based on development of text-mining techniques. Using full-text data is considered to be significant, since it contains more comprehensive entities compared to title and abstract [9][10]. Thus, the scope of data for entitymetrics analysis leads to disparate research outcomes. Based on this idea, our study suggests a new approach for entitymetrics by focusing on citation contexts in an article. Such method is expected to extend the base of bibliometric knowledge discovery.

Entitymetrics has been actively utilized on biomedical literatures. Zhu et al. [5] proposed a framework to build a paper-entity/entity-entity cooccurrence, and entity-specific network from the title and abstract of the paper for identifying relationships between liver cancer related disease, drugs, and gene entities. In addition, drug repurposing studies have also actively utilized entitymetrics [11][12]. Such previous works define bibliometric indicators upon bio-entities such as drugs, diseases, and symptoms extracted from biomedical literatures. This enables an efficient literature-based knowledge discovery in the field of biomedicine, where publication is conducted actively with large volume. In this context, the current paper also uses biomedical literature to suggest the usefulness of our suggested approach. More specifically, the current study targets the domain of opioid research. The opioid domain is an actively

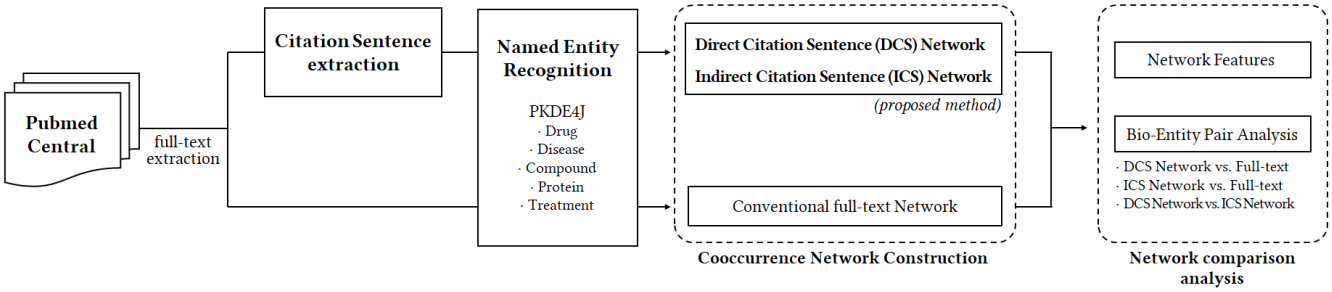

Figure 1: The Overall Schematic Research Workflow for the Proposed Methods

studied research field, especially after the rise of the opioid crisis due to the overprescription of opioid medications from the 1990s until today [13]. According to the study of Sweileh et al [14], research productivity of tramadol, which is a widely used opioid pain medication, has risen significantly since the 1970s and 80s and has sparked in the year 2008. This indicates that there will be abundant rate of knowledge entities and entity pairs to explore in opioid-related publications.

Our study focuses on citation sentence, where reference information is included. Citation sentences include both the citing and cited author's intention for the corresponding contents. In the perspective of the citing author, he or she takes advantage of the citation sentence in order to obtain credibility for their research. For the cited author, citation sentence is a channel of recognition receiving recognition from other researchers for their established research findings. Based on such characteristics, analysis of citation sentences can provide insights, which reflects both citing and cited author's interest.

The current study suggests a novel entitymetrics based approach in three aspects. First, we conduct entitymetrics analysis using citation sentence to expand the base of entitymetrics. Second, based on this citation sentence based entitymetrics, we propose two novel networks, namely the direct citation sentence network (DCS) and the indirect citation sentence (ICS) network, which enables us to analyze the knowledge entities and knowledge structures of a certain research domain in different perspectives. Third, to demonstrate the usefulness of our suggested method, we conduct knowledge structure analysis towards scientific literature in the opioid research domain. During this process, we aim to investigate three specific research questions as follows: 1) How can we construct a citation sentence-based cooccurrence network by using entitymetrics? 2) What is the key difference between our suggested networks and a conventionally built cooccurrence network? 3) What kind of new aspects can we find from the knowledge structure of opioid domain by utilizing citation sentence-based networks?

This paper explores opioid research domain using bio-entities extracted from citation sentences. Also, we construct two novel cooccurrence networks derived from the corresponding entitymetrics method and conduct a bio-entity pair analysis. The results show that suggested method yields the identification of unobserved bio-entity pairs with significant difference. The rest of the paper is organized as follows: Section 2 explains detailed methodology, Section 3 presents research results, and Section 4 concludes the research and suggests future work.

2 Methodology

Figure 1 shows the overall workflow of the current study.

2.1 Data Collection & Parsing

To conduct citation sentence entitymetrics and construct DCS and ICS network, we collected the total full-text research papers that were published until March 2022 by using "opioid" as the search query in PubMed Central (PMC). A total of 118,808 papers were collected in this process. After the paper retrieval stage, we parsed each article's PMID/PMCID and full-text. For each sentence in the collected full-text data, we designated distinct ID. After which, we identified citation sentences that referred to other journal articles or documents. In addition, the referred authors (cited authors) in the citation sentences were collected and matched with the corresponding sentences.

2.2 Bio-Entity Extraction

To extract bio-entities from the collected citation sentences, PKDE4J [15] was employed for named entity recognition (NER). PKDE4J is a framework designed for dictionary-based NER tasks, which consists of two major modules: entity extraction module and relation extraction module. For the current study, we only used the entity extraction module. The extraction module contains another four sub-modules: dictionary loading, pre-processing, entity annotation, and post-processing module. Dictionary-wise, it is possible to add multiple dictionaries for the entity extraction process. To obtain comprehensive findings from the collected dataset, we used a total of five biomedical entity dictionaries based on drug, disease, compound, protein, and treatment. These dictionaries were built from biomedical and clinical databases such as BioGrid, PharmGKB, NCBI taxonomy, PubChem, Drugbank, Medical Subject Headings (MeSH), and ClinicalTrials.gov.

2.3 Network Construction

Based on the idea that entity pair that co-occurs within a unit is considered to have a strong associtation, we computed

cooccurrence of bio-entities and construct an entity-entity network based on their relationships. In contrast with conventional entitymetrics cooccurrence network built upon full-text, this research employed a novel approach by constructing a network only using citation sentence. Such approach is expected to provide novel insights that have not yet been addressed regarding knowledge structure. We formed two networks in accordance with differently set cooccurrence window. To be specific, one is built upon entity cooccurrence within same citation sentences, whereas the other is formed based on entity cooccurrence within author information that is included in the citation sentence. Since the former method considers direct cooccurrence within the citation sentences, it is defined as a direct citation sentence network (DCS network). The latter approach is defined as an indirect citation sentence network (ICS network) because it captures cooccurrence beyond sentence-based occurrence instance by generating indirect pairs employing author information. The construction framework is presented in Figure 2.

2.3.1 Direct Citation Sentence (DCS) Network. For the first approach, we calculated the sentence-level cooccurrences of bio-entities using the citation sentences. For instance, in Figure 2, bio-entities that appeared in the same citation sentence is computed as a cooccurrence pair. For instance, since <Citation Sentence 1> in Figure 2a has four bio-entities (dopamine, neurotransmitter, midbrain, and striatum), a total of six bio-entity pairs are provided (dopamine-neurotransmitter, dopamine-midbrain, dopamine-striatum, neurotransmitter-midbrain, neurotransmitter-striatum, and midbrain-striatum) as denoted in Figure 2b. This approach is adopted under the idea that the entity pairs in the citation sentences contain key information the citing author intended to emphasize, despite its small volume. Citation is an act of authors attempt to obtain credibility for his or her assertion, and at the same time, it is an act of assigning credit to the cited author. This indicates that entities included in citation sentences can be considered as key entities. Thus, utilizing DCS network can identify key bio-entities and entity cooccurrence pairs more concretely, which

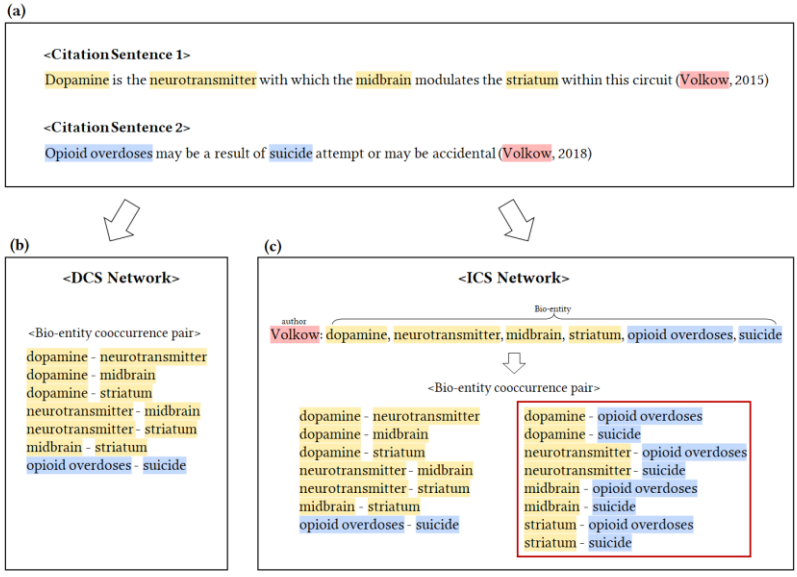

Figure 2: Cooccurrence Network Construction Process for the Direct and Indirect Citation Sentence Network. Newly captured entity pairs are highlighted (red box)

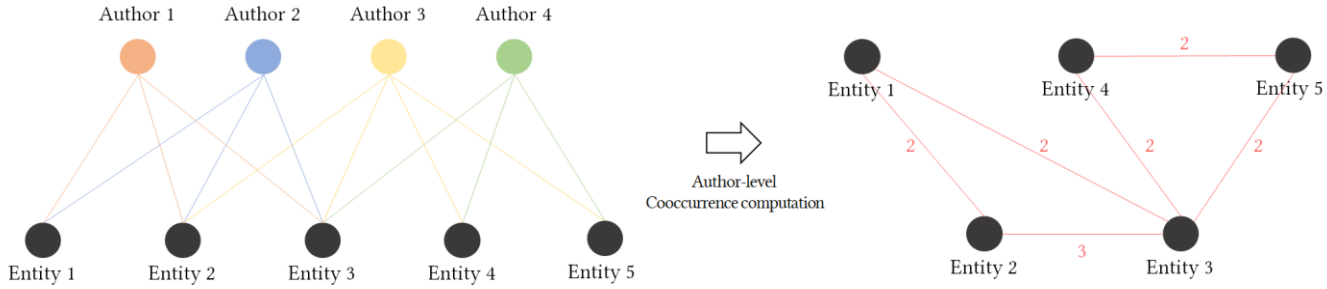

Figure 3: Citation sentence-based author-entity bipartite network converted into an entity-entity network

expands the scope of entitymetrics and knowledge structure analysis.

2.3.2 Indirect Citation Sentence (ICS) Network. Like the previous method, citation sentences are utilized to construct ICS network. However, unlike the DCS network, this network considers the author-level cooccurrence of bio-entities. Doing so enables construction of indirect connection between key bio-entities in a way which conventional methods could not. For this, we linked a connection between the author and bio-entity using citation sentences with the cited author. That is, the bio-entities appearing in citation sentence were regarded to belong to the cited author. After which, we counted the frequency of the bio-entity pairs that belong to each author. For example, <Citation Sentence 1> and <Citation Sentence 2> in Figure 2a are both citing the same author (i.e., Volkow). Since <Citation Sentence 1> includes four bio-entities (dopamine, neurotransmitter, midbrain, and striatum) and <Citation Sentence 2> contains two bio-entities (opioid overdoses and suicide), a total of six bio-entities are belonged to the cited author (Figure 2c). This process represents an author-entity bipartite network (Figure 3). Then, the corresponding bipartite network is converted into an entity-entity network by computing the bio-entity cooccurrence pair within a cited author. This approach has considerable advantages as it extends the window of cooccurrence through author information in specific sentences (in this case, citation sentences) of individual papers. To be more precise, since ICS network considers the cited authors' corresponding works, knowledge entities extracted from this network can be thought as carrying several authors' research key points. Moreover, in the aspect of entity cooccurrence pair, unobserved entity pairs that have not yet been scrutinized can also be detected.

2.4 Network Comparison Analysis

To demonstrate our proposed method's usefulness, we compared DCS and ICS network with a conventionally built full-text cooccurrence network in two different aspects. First, we compared network features of our suggested networks with the traditional full-text network. In this process, we provided network features such as network density, average path length (i.e., geodesic length), average clustering coefficient, and modularity to examine network characteristics. Then, we explored the bio-entity pairs derived from the cooccurrence results from the DCS and the ICS network. Based on this, we conducted a comparison analysis with the conventional full-text network to observe the distinguishing entity pairs, which are only shown in our proposed networks.

2.4.1 Network Features. Density of a network represents the overall degree of connection between nodes in a network. It is measured as the ratio of the number of links that are actually present to the maximum number of possible connections in the network. The calculation for network density is as follows:

$$D = \frac{\sum L_w}{{}_N C_2}$$

where D is the density, N is the number of nodes, and L_w is the weighted link between two distinct nodes.

Average Path Length (APL) is a network feature that is calculated by the average number of steps among the geodesic paths (i.e., shortest paths) for all possible pairs in a network [16]. This measure can be expressed as:

$$APL = \frac{1}{\frac{1}{2}N(N+1)} \sum_{i>j} d_{ij}$$

where N is the total count of nodes and d_{ij} is the shortest path length between node i and node j.

The Average Clustering Coefficient (ACC) indicates the degree of association between local communities that are comprised of nodes and the degree of aggregation of a network [17]. Higher ACC means that there is also a higher tendency of topological clustering in the network. ACC can be described as follows:

$$C = \frac{3 \times number\ of\ \ triangles}{number\ of\ \ all\ triplets}$$

Modularity is another common macro-level network feature that measures the strength of network community characteristic [18]. Similar to ACC, high modularity level indicates better community detection. Modularity algorithm is computed as follows:

$$M = \frac{1}{2m} \sum_{i,j} \left[L_{ij} - \frac{k_i k_j}{2m} \right] \delta\left(c_i, c_j \right)$$

where M represents the modularity, L_{ij} is the weight of the edge between i and j, k_i is the sum of the weights of the edges linked to node i (same goes with node j), c_i is the community node i is assigned to, $\delta(c_i, c_j)$ is 1 when $c_i = c_j$ and is 0 when $c_i \neq c_j$, and $m = \frac{1}{2}\sum_{ij} L_{ij}$ [19].

2.4.2 Bio-Entity Pair Analysis. After examining network characteristics based on different network indicators, we compared the top-20 bio-entity pairs that were observed in the suggested DCS and ICS network with the top-ranked bio-entity pairs in the traditional full-text cooccurrence network. Top-20 bio-entity pairs from DCS network and ICS network were also compared with each other to further distinguish each networks' advantage. Difference in entity cooccurrence pair results from distinctness of analysis scope, and the suggested citation sentence-based network construction yields novel insights regarding knowledge entities and knowledge structures.

The domain of opioid research was explored through this process. Opioid is a heavily studied biomedical concept since it is extremely necessary for surgical contexts [20][21] and highly addictive [22] at the same time. This indicates that opioid is an extremely sensitive topic due to its double-edged sword feature. For this reason, it was worth examining knowledge entities and knowledge structure of the opioid domain to analyze the thoroughly studied research field.

We used conventional full-text cooccurrence pair that was collected using sentence-level cooccurrence extraction. Network comparison result is presented in the next section.

3 Results

We constructed two cooccurrence networks based on opioid-related bio-entities extracted from citation sentences: the direct citation sentence (DCS) network and the indirect citation sentence (ICS) network. Based on different power-law distribution studies [23][24][25], we excluded bio-entities and bio-entity cooccurrence pairs that showed unusually low-frequencies (frequency less than 10) in order to obtain reasonable results by getting rid of non-informative data. Also, a total of 75 bio-entities that were in the top-100 entity frequency list were excluded due to their ambiguous and overly general characteristics. For instance, entities such as "opioids" were excluded since it was obvious for us to observe such words due to the fact that we used "opioid" as our search query. Other examples include inexact terms such as "treatment," "drug," "human," and so on (see Appendix A). DCS network consists of 6,105 bio-entities and 45,087 links, whereas ICS contains 13,525 bio-entities and 1,831,917 links. For comparison, a cooccurrence network was formed based on full-text data, which consists of 13,292 bio-entities and 144,800 links. The fact that ICS network has more identified bio-entities than the full-text network indicates that the author-entity bipartite network indirectly connects significant number of bio-entities throughout the whole dataset.

3.1 Network Features

3.1.1 Conventional Method. As mentioned above, a conventional full-text cooccurrence network was built for comparison. The density of this network is 0.00164, which means that 0.164% of the whole possible links are presented. This particular metrics represents the average strength of the possible connections among the entire network. The average path length is 3.351, which indicates that the shortest path among the entire entity pairs is about 3 step long.

The full-text network's ACC is 0.601 and the modularity is 0.407. According to Newman [26], a network with modularity greater than 0.3 is considered to have significant community structures. For this reason, we examine the major clusters of this network. According to Table 1, there are three major clusters, which are: 1) tumor and disease related, 2) psychological disorder and reward system related, 3) anesthetic and analgesic related.

Table 1: Top-10 bio-entities for each cluster in full-text network (Cluster 1: anesthetic and analgesic related, Cluster 2: tumor and disease related, Cluster 3: psychological disorder and reward system related)

Cluster 1	morphine; anesthesia; fentanyl; infusion; saline; sedation; propofol; ketamine; analgesics; postoperative pain
Cluster 2	tumor; liver; glucose; mrna; dna; hypertension; hcv; obesity; rna; il-6
Cluster 3	substance use disorder; withdrawal; chronic pain; addiction; dopamine; cocaine; amp; neuropathic pain; methadone; mental health

3.1.2 DCS Network Features. The density of the DCS network is 0.00242, which indicates that 0.242% of all possible links are presented in the current network. The DCS network has a larger density than the conventionally built network, which shows that our method suggests a more connective network. The average path length of the DCS network is 3.434, which is similar to the full-text network. This means that both networks' shortest path of all bio-entity pairs is approximately 3 steps. Based on the study of Ding et al. [1], this can be interpreted as both networks having an efficiency regarding knowledge flow.

ACC for the current network is 0.612, which shows that nodes in the DCS network tends to form clusters with each other since it is even higher than the traditional full-text network. Also, the modularity for the DCS network is 0.455, which is higher than the full-text network. These results indicate that the DCS network has a higher clustering tendency than the conventionally formed full-text network. Hence, we investigated the major topological clusters that appeared in this network. In the suggested DCS network, there are four major clusters, which are: 1) pain management related, 2) tumor and disease related, 3) anesthetic and analgesic related, and 4) psychological disorder and reward system related. This result shows that the DCS network provides more specific topological clusters than the conventional full-text network. Table 2 shows DCS network's top-10 bio-entities for each cluster based on the weighted degree.

Table 2: Top-10 bio-entities for each cluster in direct citation sentence network (Cluster 1: pain management related, Cluster 2: tumor and disease related, Cluster 3: anesthetic and analgesic related, Cluster 4: psychological disorder and reward system related)

Cluster 1	chronic pain; neuropathic; pain postoperative pain; analgesics; pain management; pain relief; quality of life; hyperalgesia; paracetamol; painful
Cluster 2	tumor; liver; calcium; mrna; mitochondrial; proliferation; oxidative stress; dna; nmda; il-6
Cluster 3	morphine; anesthesia; sedation; fentanyl; ketamine; infusion; propofol; epidural; dexmedetomidine; adverse effects
Cluster 4	substance use disorder; dopamine; addiction; withdrawal; cocaine; reward; amp; mental health; gaba; methadone

3.1.3 ICS Network Features. Unlike the DCS network, the ICS network represents a much more compact characteristic in the sense of network connectivity. The density of the ICS network is 0.02010, which refers to the fact that 2.01% of all possible linkages are provided in the corresponding network. This

measure is significantly higher than both conventional full-text and DCS network. Also, the average path length of the ICS network is 2.336. In other words, the average of every shortest path of all entity couples is approximately little more than 2 steps, which is even shorter than the conventional full-text and DCS network. This suggests that the ICS network also has a structure for highly efficient knowledge flow since it presents the lowest average path length.

The ACC for the ICS network is 0.918, which indicates that the current network is extremely connective while having a great clustering tendency. However, the ICS network has the lowest modularity (0.148). Despite the low modularity, this network has three major topological clusters in the domain of opioid research, which are: 1) psychological disorder and reward system related, 2) pain disorder related, and 3) tumor and disease related. This seems reasonable since the ICS network suggests a significantly high ACC. Table 3 suggests the top-10 bio-entities for all three clusters ordered by the weighted degree.

Table 3: Top-10 bio-entities for each cluster in indirect citation sentence network (Cluster 1: psychological disorder and reward system related, Cluster 2: pain disorder related, Cluster 3: tumor and disease related)

Cluster 1	SUD; addiction; dopamine; withdrawal; mental health; reward; amp; perception; emotional; psychological
Cluster 2	chronic pain; morphine; neuropathic pain; anesthesia; analgesics; adverse effects; quality of life; pain relief; persistent; postoperative pain
Cluster 3	tumor; calcium; gaba; liver; mrna; obesity; proliferation; progression; toxicity; oxidative stress

These findings suggest that our proposed methods have distinct advantages when compared with the traditional full-text cooccurrence network. All three networks showed decent average path length, which indicates efficient knowledge flow. However, the DCS and ICS network represented higher density and ACC. To be more specific, the ICS network was significantly more connective than the other two networks. This supports the fact that our proposed methods can provide a much more compact network in the perspective of network linkage. Also, focusing on topological cluster, the DCS network provided the most concrete result by presenting four specific topic clusters while the other networks offered three.

3.2 Bio-Entity Pair Analysis

Since our citation sentence-based networks (DCS & ICS network) show certain strengths, it is worth investigating the highly ranked bio-entity pairs extracted from these networks to analyze knowledge entities and knowledge structures. We compare the top-20 opioid-related bio-entity pair rank with another entity pair rank derived from the conventional full-text network (Appendix B).

3.2.1 DCS Network vs. Full-Text Network. In Table 4, there are a total of five bio-entity pairs that only appear in the DCS network when compared with the traditional full-text network. This means that utilizing citation sentence-based entitymetrics enables us to capture unobserved bio-entity cooccurrences. These entity pairs are dopamine-reward, hyperalgesia-allodynia, withdrawal-morphine, gabapentin-pregabalin, amphetamine-cocaine, and SUD-cocaine. Among these top ranked bio-entity pairs, dopamine-reward is the most highly observed cooccurrence. This entity dyad is highly important in the field of opioid (especially opioid addiction) since it represents the concept of "reward system". The reward system (also known as mesolimbic dopamine system) is a brain region that is comprised of several cortical and subcortical brain regions that mediates complex incentive learning and promotes motivation [27]. This system is known to be closely related to the human's endogenous opioid system and addictive disorders [28].

Hyperalgesia-allodynia is also exclusively included in the top-20 opioid-related bio-entity pair from the direct citation sentence network. Hyperalgesia and allodynia are both pain disorders that are associated with severe neuropathic pain. While hyperalgesia is an escalated pain from a stimulus that normally occurs pain, allodynia is linked to pain that usually does not provoke pain [29]. Though the relationship between opioid and these two pain-related disorders are not yet understood in the molecular-level, it is thought to be that high dosage of opioid administration induces such symptoms due to opioid tolerance [30][31]. The fact that an entity pair that was newly introduced in the top cooccurrence list needs further investigation indicates that the corresponding bio-entity pair is being thoroughly studied in the domain of opioid.

Table 4: Top-20 bio-entity pairs in the direct citation sentence network (bolded pairs represent exclusive pairs compared with the full-text network)

Entity 1	Entity 2	Freq
SUD	addiction	1993
mental health	SUD	1992
methadone	buprenorphine	1989
morphine	fentanyl	1942
dopamine	**reward**	**1544**
anesthesia	propofol	1482
cbd	thc	1290
glucose	insulin	1206
heroin	cocaine	1203
hyperalgesia	**allodynia**	**1172**
postoperative pain	pain management	1162
withdrawal	morphine	1146
gabapentin	**pregabalin**	**1132**
propofol	sedation	1126
naloxone	overdose	1119
morphine	oxycodone	1107
anterior	posterior	1092
hip	fracture	1052
amphetamine	**cocaine**	**1050**
SUD	**cocaine**	**1048**

Gabapentin-pregabalin pair is also newly observed in the top-20 list derived from the DCS network. Manufactured by Pfizer, both these anticonvulsants (i.e., drugs meant for epilepsy reduction and prevention) are known to be significantly associated with opioid use disorder patients since it has a likelihood of high co-prescription with opioid medications [32][33]. It is not yet certain whether these drugs are harmful as opioids when they are misused [34][35] or whether they affect opioid receptors in the human body [34]. Considering this situation, the fact that gabapentin-pregabalin pair is being frequently observed highlights its research value in the domain of opioid research.

Another bio-entity pairs that appeared in the top-20 list are amphetamine-cocaine and SUD-cocaine. The appearance of these cooccurrences is explainable since amphetamine and cocaine are psychostimulants that are highly addictive [36].

3.2.2 ICS Network vs. Full-Text Network. Compared with the traditional full-text network, the ICS network has 17 unique bio-entity pairs among the top-20 list (Table 5). This indicates that the ICS network, which utilized author information in the process of the network construction, offers an additional understanding regarding the knowledge structure of the domain of opioid research. These entity dyads fall into the scope of addictive disorder and pain disorder. To be more specific, while chronic pain-neuropathic pain, chronic pain-morphine, analgesics-chronic pain, and chronic pain-pain management are linked to pain disorder, the other bio-entity pairs are connected to addictive disorder. For instance, SUD-heroin pair represents heroin addiction, which is one of the most prevalent opioid drug addictions nowadays [37]. Analgesics-chronic pain, on the other hand, suggests the close connection with opioid administration for pain management. Due to its highly controversial and addictive nature, opioid analgesics for pain treatment is being thoroughly studied [38]. This shows the two-sided characteristic of opioid use since it emphasizes both the positive (treatment for pain disorder) and the negative (opioid addiction) aspect of opioid. In addition, this tendency suggests that the corresponding features of opioid use receive robust attention from various authors in the field of opioid-related studies.

Table 5: Top-20 bio-entity pairs in the indirect citation sentence network (bolded pairs represent exclusive pairs compared with the full-text network)

Entity 1	Entity 2	Freq
SUD	addiction	8684
mental health	SUD	6480
chronic pain	**neuropathic pain**	**6080**
SUD	**cocaine**	**5814**
dopamine	**reward**	**5363**
withdrawal	**SUD**	**4819**
withdrawal	**addiction**	**4507**
addiction	**reward**	**4448**
abuse	**SUD**	**4246**
chronic pain	**morphine**	**4227**
analgesics	**morphine**	**4210**
morphine	fentanyl	4177
analgesics	**chronic pain**	**4159**
dopamine	**addiction**	**4135**
SUD	**heroin**	**4120**
SUD	**reward**	**4110**
addiction	**cocaine**	**4108**
dopamine	**SUD**	**4105**
chronic pain	**pain management**	**4002**
SUD	**relapse**	**3930**

3.2.3 DCS Network vs. ICS Network. It has been suggested in the previous sections that both DCS and ICS network provide us with meaningful insights towards the knowledge structure of opioid-related research. However, this has been done at different levels. Since the ICS network reflects many different authors' topological research key points, the bio-entity pairs observed in this network tend to be much more general and broader than the bio-entity pairs from the DCS network. Generality and broadness can be explained by ICS network's characteristic, which can be related to our findings in section *3.1*, where it was suggested that the ICS network offered a highly connective network structure (high density and ACC) while obtaining knowledge flow efficiency (low average path length) at the same time. On the other hand, the top-ranked bio-entity cooccurrences that appeared in the DCS network deals with more narrowed-down and detailed opioid-related concept pairs. This can also relate with our results in the previous subsection, where it was highlighted that the DCS network provides a much more specific topological clusters based on modularity measure (Table 2). This can be explained by the fact that the bio-entity pairs from the DCS network are under the realm of those derived from the indirect citation sentence network. For example, hyperalgesia-allodynia pair is included in the domain of pain disorder, while pain disorder is heavily covered by the bio-entity pairs observed in the ICS network. At the same time, amphetamine-cocaine pair provides specific examples of opioid based psychostimulant, which is immensely associated with addictive disorders. Also, CBD (cannabidiol) and THC (tetrahydrocannabinol) are both the main psychoactive components of marijuana, which is also an actively studied addictive substance that appeared in the DCS network. Likewise, addictive disorder is greatly dealt by the top-ranked bio-entity pairs from the ICS network.

Table 6: Top-20 bio-entity pair comparison between DCS and ICS network (bolded pairs represent exclusive pairs compared with each other)

DCS Network		ICS Network	
Entity 1	Entity 2	Entity 1	Entity 2
SUD	addiction	SUD	addiction
mental health	SUD	mental health	SUD
methadone	**buprenorphine**	**chronic pain**	**neuropathic pain**
morphine	fentanyl	SUD	cocaine
dopamine	reward	dopamine	reward

anesthesia	**propofol**	**withdrawal**	**SUD**
cbd	**thc**	**withdrawal**	**addiction**
glucose	**insulin**	**addiction**	**reward**
heroin	**cocaine**	**abuse**	**SUD**
hyperalgesia	**allodynia**	**chronic pain**	**morphine**
postoperative pain	**pain management**	**analgesics**	**morphine**
withdrawal	**morphine**	morphine	fentanyl
gabapentin	**pregabalin**	**analgesics**	**chronic pain**
propofol	**sedation**	**dopamine**	**addiction**
naloxone	**overdose**	**SUD**	**heroin**
morphine	**oxycodone**	**SUD**	**reward**
anterior	**posterior**	**addiction**	**cocaine**
hip	**fracture**	**dopamine**	**SUD**
amphetamine	**cocaine**	**chronic pain**	**pain management**
SUD	cocaine	**SUD**	**relapse**

4 Conclusion

This paper proposes a novel approach to entitymetrics which utilizes citation sentences so that we can measure the impact of knowledge entities and analyze the knowledge structure with a different perspective, compared to conventional approaches where abstract or full-text is utilized. To be more specific, we suggest two citation sentence-based networks, namely the direct citation sentence (DCS) network and the indirect citation sentence (ICS) network. Both networks are cooccurrence networks that are constructed based on citation sentences that were extracted from full-text data collected from PMC. The DCS network is built by calculating cooccurrence pair within a citation sentence. The ICS network, on the other hand, is formed in a way where indirect connection between entities are being captured based on cited author information. That is, we first compute an author-entity bipartite network, then we convert this network into an entity-entity network (i.e., ICS network).

When compared with a conventionally built full-text network, it was clear that DCS and ICS network respectively hold different advantages regarding network features. Both the DCS and ICS network show denser network connectivity than the traditional full-text network. This is especially prominent in the ICS network since it has the highest density and ACC measures. DCS network provides the most detailed topic cluster compared to ICS and conventional full-text network based on the highest modularity.

Furthermore, to examine whether our proposed methods can provide us with novel knowledge entity/structure analysis results, we explore the domain of opioid research. This study compares each network's top-20 bio-entity pairs with the conventional full-text cooccurrence network. The comparison results show that our suggested methods successfully capture novel entity pairs in the rank, which are not presented in the top list of the network constructed with conventional approach. Even though our suggested networks provided unobserved bio-entity pairs in the top list compared with the traditional method, significant differences were also captured between DCS and ICS network. While the ICS network tends to provide much more general and broader bio-entity pairs, the DCS network offers much more specific and specialized bio-entity pairs. This can be linked to our previous findings regarding network feature-based characteristics. The generality and broadness of bio-entity pairs in the ICS network can be supported by ICS network's high connectiveness and efficient knowledge flow. The detailed feature of the bio-entity pairs extracted from the DCS network can be explained by DCS network's topological specificity. The novel approach of citation sentence-based entitymetrics thus provide insights which cannot be captured via conventional method. These methods can support the need for the use of citation sentences in future entitymetrics studies when a more in-depth knowledge structure analysis is needed.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A2B5B02002359).

REFERENCES

[1] Ying Ding, Min Song, Jia Han, Qi Yu, Erjia Yan, Lili Lin, and Tamy Chambers, 2013. Entitymetrics: Measuring the impact of entities. *PloS one* 8, 8 (Aug, 2013), e71416. DOI: https://doi.org/10.1371/journal.pone.0071416

[2] Allan Peter Davis, Thomas C. Wiegers, Phoebe M. Roberts, Benjamin L. King, Jean M. Lay, Kelley Lennon-Hopkins, Daniela Sciaky, Robin Johnson, Heather Keating, Nigel Greene, Robert Hernandez, Kevin J. McConnell, Ahmed E. Enayetallah, and Carolyn J. Mattingly, 2013. A CTD–Pfizer collaboration: manual curation of 88

[3] Xuelian Pan, Erjia Yan, Ming Cui, and Weina Hua, 2018. Examining the usage, citation, and diffusion patterns of bibliometric mapping software: A comparative study of three tools. *Journal of informetrics* 12, 2 (May, 2018), 481-493. DOI: https://doi.org/10.1016/j.joi.2018.03.005

[4] Yoo Kyung Jeong, Qing Xie, Erjia Yan, and Min Song, 2020. Examining drug and side effect relation using author–entity pair bipartite networks. *Journal of informetrics* 14, 1 (Feb, 2020), 100999. DOI: https://doi.org/10.1016/j.joi.2019.100999

[5] Yongjun Zhu, Min Song, and Erjia Yan, 2016. Identifying liver cancer and its relations with diseases, drugs, and genes: A literature-based approach. *PLoS One* 11, 5 (May, 2016), e0156091. DOI: https://doi.org/10.1371/journal.pone.0156091

[6] Bahaa Ibrahim, 2021. Statistical methods used in Arabic journals of library and information science. *Scientometrics* 126, 5 (Mar, 2021), 4383-4416. DOI: https://doi.org/10.1007/s11192-021-03913-2

[7] Yuzhuo Wang and Chengzhi Zhang. 2018. *Using Full-Text of Research Articles to Analyze Academic Impact of Algorithms*. In: Chowdhury, G., McLeod, J., Gillet, V., Willett, P. (eds) Transforming Digital Worlds. iConference 2018. Lecture Notes in Computer Science, vol 10766. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-78105-1_43

[8] Yuzhuo, Wang and Chengzhi Zhang, 2020. Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of Informetrics* 14, 4 (Nov, 2020), 101091. DOI: https://doi.org/10.1016/j.joi.2020.101091

[9] Mengnan Zhao, Erjia Yan, and Kai Li, 2017. Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology* 69, 1 (Sep, 2018), 32-46. DOI: https://doi.org/10.1002/asi.23919

[10] Yuzhuo Wang, Chengzhi Zhang, and Kai Li, 2022. A review on method entities in the academic literature: extraction, evaluation, and application. *Scientometrics*, (Mar, 2022), 1-42. DOI: https://doi.org/10.1007/s11192-022-04332-7

[11] Yanhua Lv, Ying Ding, Min Song, and Zhiguang Duan, 2018. Topology-driven trend analysis for drug discovery. *Journal of Informetrics* 12, 3 (Aug, 2018), 893-905. DOI: https://doi.org/10.1016/j.joi.2018.07.007

[12] Xin Li, Justin F. Rousseau, Ying Ding, Min Song, and Wei Lu, 2020. Understanding drug repurposing from the perspective of biomedical entities

and their evolution: Bibliographic research using aspirin. *MIR medical informatics* 8, 6 (Jun, 2020), e16739. DOI: https://doi.org/10.2196/16739

[13] Nora D. Volkow, Emily B. Jones, Emily B. Einstein, and Eric M. Wargo, 2019, Prevention and treatment of opioid misuse and addiction: A review. JAMA Psychiatry, 76, 2 (Feb, 2019) 208–216. DOI: https://doi.org/10.1001/jamapsychiatry.2018.3126

[14] Waleed M. Sweileh, Naser Y. Shraim, Sa'ed H. Zyoud and Samah W. Al-Jabi, 2016, Worldwide research productivity on tramadol: A bibliometric analysis. SpringerPlus, 5, 1108 (Jul, 2016) DOI: https://doi.org/10.1186/s40064-016-2801-5

[15] Min Song, Won Chul Kim, Dahee Lee, Go Eun Heo, and Keun Young Kang, 2015. PKDE4J: Entity and relation extraction for public knowledge discovery. *Journal of biomedical informatics* 57, (Oct, 2015), 320-332. DOI: https://doi.org/10.1016/j.jbi.2015.08.008

[16] Guoyong Mao, and Ning Zhang, 2013. Analysis of Average Shortest-Path Length of Scale-Free Network. *Journal of Applied Mathematics* 2013, (Jul, 2013). DOI: https://doi.org/10.1155/2013/865643

[17] Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications* (1st. ed.). Cambridge University Press, Cambridge, England.

[18] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (Oct, 2008), P10008. DOI: https://doi.org/10.1088/1742-5468/2008/10/P10008

[19] M. E. J. Newman, 2004. Analysis of weighted networks. *Physical review E* 70, 5 (Nov, 2004), 056131. DOI: https://doi.org/10.1103/PhysRevE.70.056131

[20] Kate Flemming MSc, RN, 2010. The Use of Morphine to Treat Cancer-Related Pain: A Synthesis of Quantitative and Qualitative Research. *Journal of Pain and Symptom Management* 39, 1 (Jan, 2010), 139-154. DOI: https://doi.org/10.1016/j.jpainsymman.2009.05.014

[21] Garrett Enten, Mina A. Shenouda, David Samuels, Naomi Fowler, Maha Balouch, and Enrico Camporesi, 2019. A Retrospective Analysis of the Safety and Efficacy of Opioid-free Anesthesia versus Opioid Anesthesia for General Cesarean Section. *Cureus* 11, 9 (Sep, 2019), e5725. DOI: https://doi.org/10.7759/cureus.5725

[22] Andrew Kolodny, David T. Courtwright, Catherine S. Hwang, Peter Kreiner, John L. Eadie, Thomas W. Clark, and G. Caleb Alexander, 2015. The Prescription Opioid and Heroin Crisis: A Public Health Approach to an Epidemic of Addiction. *Annual review of public health* 36, (Mar, 2015), 559-574. DOI: https://doi.org/10.1146/annurev-publhealth-031914-122957

[23] Steven T. Piantadosi, 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review* 21, 5 (Mar, 2014), 1112-1130. DOI: https://doi.org/10.3758/s13423-014-0585-6

[24] Álvaro Corral, Gemma Boleda, and Ramon Ferrer-i-Cancho, 2015. Zipf's law for word frequencies: word forms versus lemmas in long texts. *PloS one* 10, 7 (Jul, 2015), e0129031. DOI: https://doi.org/10.1371/journal.pone.0129031

[25] Staša Milojević, 2010. Power law distributions in information science: Making the case for logarithmic binning. *Journal of the American Society for Information Science and Technology* 61, 12 (Nov, 2010), 2417-2425. DOI: https://doi.org/10.1371/journal.pone.0129031

[26] M. E. J. Newman, 2004. Fast algorithm for detecting community structure in networks. *Physical review E* 69, 6 (Jun, 2004), 066133. DOI: https://doi.org/10.1103/PhysRevE.69.066133

[27] Lauri Nummenmaa, Tiina Saanijoki, Lauri Tuominen, Jussi Hirvonen, Jetro J. Tuulari, Pirjo Nuutila, and Kari Kalliokoski, 2018. μ-opioid receptor system mediates reward processing in humans. *Nature communications* 9, 1 (Apr, 2018), 1-7. DOI: https://doi.org/10.1038/s41467-018-03848-y

[28] Julie Le Merrer, Jérôme A. J. Becker, Katia Befort, and Brigitte L. Kieffer, 2009. Reward processing by the opioid system in the brain. *Physiological reviews*, (Oct, 2009). DOI: https://doi.org/10.1152/physrev.00005.2009

[29] Troels S Jensen and Nanna B Finnerup, 2014. Allodynia and hyperalgesia in neuropathic pain: clinical manifestations and mechanisms. *The Lancet Neurology* 13, 9 (Sep, 2014), 924-935. DOI: https://doi.org/10.1016/S1474-4422(14)70102-4

[30] Marion Lee, Sanford Silverman, Hans Hansen, Vikram Patel, and Laxmaiah Manchikanti, 2011. A comprehensive review of opioid-induced hyperalgesia. *Pain Physician* 14, 2 (Jan, 2011), 145-161.

[31] Maurice H. Zissen, Guohua Zhang, Alvin McKelvy, John T. Propst, Joan J. Kendig, and Sarah M. Sweitzer, 2007. Tolerance, opioid-induced allodynia and withdrawal associated allodynia in infant and young rats. *Neuroscience* 144, 1 (Jan, 2007), 247-262. DOI: https://doi.org/10.1016/j.neuroscience.2006.08.078

[32] Tara Gomes, David N. Juurlink, Tony Antoniou, Muhammad M. Mamdani, J. Michael Paterson, and Wim van den Brink, 2017. Tolerance, opioid-induced allodynia and withdrawal associated allodynia in infant and young rats. *PLoS Med* 14, 10 (Oct, 2017), 247-262. e1002396. DOI: https://doi.org/10.1371/journal.pmed.1002396

[33] Emma Morrison, Euan A. Sandilands, and David J. Webb, 2017. Gabapentin and pregabalin: Do the benefits outweigh the harms? *The Journal of the Royal College of Physicians of Edinburgh* 47, 4 (Dec, 2017), 310-313. DOI: https://doi.org/10.4997/JRCPE.2017.402

[34] Preeti Manandhar, Bridin Patricia Murnion, Natasha L. Grimsey, Mark Connor, and Marina Santiago, 2021. Do gabapentin or pregabalin directly modulate the μ receptor? *PeerJ* 9, (Apr, 2021), e11175. DOI: https://doi.org/10.7717/peerj.11175

[35] Sinead McNamara, Siobhan Stokes, R. Kilduff, and Aine Shine, 2015. Pregabalin Abuse amongst Opioid Substitution Treatment Patients. *Ir Med J* 108, 10 (Nov, 2015), 309-310. DOI: https://doi.org/10.7717/peerj.11175

[36] European Monitoring Centre for Drugs and Drug Addiction and Kateřina Škařupová, 2014 *The levels of use of opioids, amphetamines and cocaine and associated levels of harm : summary of scientific evidence*. Publications Office. DOI: https://data.europa.eu/doi/10.2810/49447

[37] Wilson M. Compton, Christopher M. Jones, and Grant T. Baldwin, 2016. Relationship between nonmedical prescription-opioid use and heroin use. *New England Journal of Medicine* 374, 2 (Jan, 2016), 154-163. DOI: https://doi.org/10.1056/NEJMra1508490

[38] Howard L. Fields, 2011. The Doctor's Dilemma: Opiate Analgesics and Chronic Pain. *Neuron* 69, 4 (Feb, 2011), 591-594. DOI: https://doi.org/10.1016/j.neuron.2011.02.001

Appendix

A Excluded Entity List

Table 7: Excluded 75 bio-entities

Entity	Frequency	Entity	Frequency
treatment	526686	lead	57961
pain	456305	oral	57447
drug	211165	procedure	57276
Fig	206186	procedures	56578
care	201449	affect	56073
response	174147	neuronal	54537
drugs	168170	severity	54409
surgery	161686	protocol	54118
brain	157926	condition	54087
dose	137096	like	53988
reduced	127899	key	53656
function	127052	medications	52639
therapy	117437	end	51958
human	117366	sensitivity	51654
blood	106818	interest	50794
disease	106162	secondary	49189
opioids	101146	rat	48710
symptoms	96081	distribution	48064
support	89745	strategies	46310
chronic	82604	adult	44728
stimulation	82076	disorders	43925
severe	80434	delivery	43609
exposure	80109	line	42817
impact	77239	side effects	42242
general	76767	right	41062
expressed	72793	injury	41590
acute	71777	understanding	40342
normal	71392	moderate	39323
management	68786	focus	37051
medication	67976	diseases	36224
measures	67668	light	35968
association	67400	onset	35871
concentrations	64023	finding	35413
central	63070	strategy	35201
food	62410	nervous	34967
block	62157	activated	34227
set	61710	content	33724
intensity	59397		

B Conventional Full-Text Network Cooccurrence Information

Table 8: Top-20 bio-entity pair in the full-text network

Entity 1	Entity 2	Freq
anesthesia	propofol	5902
methadone	buprenorphine	5708
morphine	fentanyl	5326
mental health	SUD	5240
SUD	addiction	4292
anterior	posterior	3697
morphine	oxycodone	3463
naloxone	overdose	3447
heroin	cocaine	3410
kit	rna	3329
glucose	insulin	3225
hip	fracture	3199
propofol	remifentanil	3182
anesthesia	isoflurane	3152
postoperative pain	pain management	3116
anesthesia	sevoflurane	3093
propofol	sedation	3081
hypotension	bradycardia	3057
cbd	thc	3040
withdrawal	morphine	3035