（ 続紙 1 ）

京都大学	博士（情報学）	氏名	Yang ZHANG
論文題目	Citation Knowledge Mining for On-the-fly Recommendations （その場での推薦のための引用知識マイニング）		

（論文内容の要旨）

One of the most frequent questions considered when writing academic papers is: “Which paper should I cite at this place?” However, searching and finding comprehensive, relevant articles for citing from many publications is challenging, even for expert researchers. According to the STM scholarly publishing report (https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf), up to 2018, there were 150 million articles in total published in the Web of Science databases. In addition, the number of newly published papers is also growing at 5-6% per year in recent years. It could imagine that researchers would not find and read all the potential articles relevant to their studies. Exploring the mountains of publications and efficiently recommending them has become a non-trivial problem.

To the end of assisting writing and reviewing academic papers, this work proposes a novel concept of “on-the-fly” citation recommendations to recommend practical candidate papers efficiently and accurately for citing. In addition to providing comprehensive citing candidates, there are two requirements of the “on-the-fly” recommendations: 1) the system could recommend citing candidates to a working manuscript, even it is still underwriting, and 2) the system should be able to recommend not only the papers from the database but also the out-of-dataset papers, i.e., the newly published papers.

Technically, the proposed on-the-fly citation recommendation approach leverages the advantages of the embedding techniques to mine citation knowledge and adapt them for the recommendation tasks. Citation knowledge mining involves three tasks: source representation learning for extracting the citing intents of users, target representation learning for inferring the content semantics of citing candidates from the databases, and citation relationship mining to enhance the recommendation by learning the relation knowledge from the citation network.

(1) Source Representation Learning focuses on representing the source manuscript, including users’ citing intents into semantic space. The

proposed algorithm detects the citing intents from the query contexts and adaptively detects the topic semantics from the continuous updates of the manuscript.

(2) Target Representation Learning is designed to learn the representation of the content semantics of the candidate papers, which the source manuscript may cite. The work proposes a method to construct a “universal content modeling” by adapting content embedding techniques, complied with dynamic content sampling strategies. The constructed content modeling can be adapted for representing and recommending citing candidates from both in-dataset and out-of-dataset (newly published papers).

(3) Citation Relationship Mining leverages the structural information of citations, such as co-citation relationships, co-citation frequencies, and frequencies of their historical appearances, to improve the accuracy and efficiency of citation recommendations. The proposed approach also enables the requests of multiple targets for citing, which is helpful to find comprehensive citing candidates.

Experiments simulating real-world scenarios with two open datasets have verified the proposed methods. Three completing stages with different amounts of finished contents for input manuscripts are adapted to validate the on-the-fly recommendations. In addition, extensive user tests and explainability studies are conducted to demonstrate the usability and rationality of the approaches. The proposed models are additionally analyzed in the ablation study to verify each designed component. Overall, the experiments could demonstrate the framework's effectiveness and rationality from the perspectives of accuracy, rationality, and usability.

（続紙 ２ ）

（論文審査の結果の要旨）

研究と論文作成には、既存研究をサーベイして必要十分な参考文献を適切に引用することが必要不可欠である。急激に増加し続ける膨大な数の論文から如何に効率よく関連文献を網羅的にサーベイするかは極めて重要な課題である。本論文では、既存の文献データベースから引用に関する知識を発見する技術とそれを用いた参考文献や引用を推薦する仕組みに関する以下の研究成果をまとめている。

- 新しい論文も対象としたOn-the-fly引用推薦を提案している。その場の推薦であるOn-the-fly引用推薦は、作成中または完成した直後の原稿を引用元とし、適切な引用先を推薦できる仕組みである。また、従来手法では困難とされているモデルの学習データに含まれていない新しい論文も引用先として推薦できるとしている。

- On-the-fly引用推薦を実現するため、引用元の表現学習、引用先の表現学習と引用関係分析に関する引用知識のマイニング手法を提案している。引用元の表現を効率よく学習するデュアルアテンションネットワーク手法、引用先の表現の学習には事前学習を活かしたコンテンツベースBERT引用推薦モデル（CBERT4REC）、そして、引用関係分析にはコンテキスト情報を活かした分散表現モデルDocCit2Vecや共引用検索手法をそれぞれ開発している。また、学習の効率化のため、動的サンプリング手法やマルチポジティブ目的関数を提案している。

- データ規模が異なる二つの公開データセットを用いて、複数の利用シナリオを想定して実験を行い、提案手法の有効性と有用性について検証を行っている。従来手法と比べて推薦性能の向上を確認できたほか、ユーザ実験を行ってユーザビリティを確認している。また、Ablation Studyを行い、提案手法の主要技術ポイントの有効性を確認している。さらに、モデルの中間出力について分析し、モデルとその推薦結果の解釈可能性を向上させている。

本論文で提案しているOn-the-fly引用推薦は、初心者の論文作成支援やAIによる論文自動生成のほか、審査委員による論文審査やレビューへの支援など幅広く応用でき、学術上、実際上寄与するところが多い。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和４年２月１６日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。なお、令和4年4月1日以降の本論文のインターネットでの全文公開についても支障がないことを確認した。