

更新可能な用語抽出機能を持つ小規模研究グループ向け論文・図表データベースの構築

Construction of Paper/Figure Database for Small Research Group using PDF papers

北海道大学大学院情報科学研究科¹, 北海道大学国際連携研究推進局², 北海道大学大学院情報科学研究科³, 北海道大学量子集積エレクトロニクス研究センター⁴, 物質・材料研究機構⁵

吉岡 真治^{1,2}, 大久保 好章¹, (M1) 尹 磊³, 原 真二郎⁴, 鈴木 晃⁵, 高山 英紀⁵, 石井 真史⁵

Faculty of IST, Hokkaido Univ.¹, GI-CoRE, Hokkaido Univ.², Graduate School of IST, Hokkaido Univ.³, RCIQE, Hokkaido Univ.⁴, National Institute for Materials Science⁵

E-mail: yoshioka@ist.hokudai.ac.jp

1 はじめに

本研究では、論文からの情報抽出とその活用を目指したデータベースの提案を行う。具体的には、小規模な研究グループを対象として、そのグループに関連する論文のみを収集し、専門用語抽出の技術 [1] を適用することにより、専門用語辞書の構築を支援すると共に、作成した用語辞書を用いた論文・図表の分析を可能とするデータベースを提案する。

2 論文・図表データベースの構築

2.1 関連論文 PDF を用いたデータベースの構築

本研究では、大学の研究室などの小規模な研究グループを対象とし、そのグループが興味を持つ論文の PDF (関連雑誌や国際会議の PDF 集) を基本データとして、そこから、テキスト情報や図表の情報を抜き出してデータベースの構築を行う。

新しい論文が発表される度に、継続的にデータベースを更新することにより、他の論文データベースと異なり、研究グループに関連する論文が網羅的に収録されていることが期待される。また、最新の論文も含めて、論文の集合が研究領域を分析するための基本データとなるため、特徴的な用語の使われ方に基づいた専門用語候補の生成 [1] が可能になる。この候補から、用語辞書を構築することにより、最新の用語の獲得や、用語の使われ方の時間遷移の分析などが可能となる。

このような研究支援の機能の存在を前提とすることで、研究グループのメンバーに、人での介入が不可欠である用語辞書の構築を行う動機を与えることができると考えている。このような形で、最新の研究成果を含む形で研究支援を行うと共に、今後の情報抽出に役立つ資源を作成するという循環型のシステムの構築を目指す。

2.2 プロトタイプシステムの構築

現在、前節で紹介した機能を持つインハウス論文・図表データベースのプロトタイプシステムを構築している。

本システムでは、最新論文の登録を行うために、PDF からのテキスト・図表データの抽出システム PDFFigures2¹ を利用し、論文抽出後の専門用語抽出のシステムとしては、[1] の考え方に基づく用語抽出システムである termextract² を利用する。それ以外にも、パターンベースで作成した略称 (Acronym) の抽出システムなどを組み合わせることで、同義語辞書などの拡張に利用する。

また、PDF 論文からの情報抽出の品質が与える影響などを議論するために、部分的に、XML 論文からの情報抽出を行う方法も並行して行っている。

本システムによる用語抽出の結果を用いるアプリケーションとしては、先に紹介した論文中の図表データを多観点分析するシステム [2] を利用する。また、用語の経時的な出現頻度の変化を分析するようなシステムについても、構築を予定している。

3 おわりに

現在、プロトタイプシステムを作成すると共に、特定分野の研究者で共著者の研究室、北大量子集積エレクトロニクス研究センター先進ナノ電子材料研究室と物質・材料研究機構において、本システムのセットアップと予備的な運用を始める予定である。

参考文献

- [1] Hiroshi Nakagawa and Tatsunori Mori. A simple but powerful automatic term extraction method. In COMPUTERM '02, pp. 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
- [2] 朱濤, Moustafa Dieb Thaeer, 吉岡真治, 原真二郎. ナノ知識探索プロジェクト: 実験記録からの知識発見 (第 4 報). 2016 年度人工知能学会全国大会 (第 30 回) 論文集, 2016. CD-ROM 1J2-4.

¹<http://pdffigures2.allenai.org/>

²<http://gensen.dl.itc.u-tokyo.ac.jp/pytermextract>