Construction of In-house Paper/Figure Database System Supporting Research Trend Analysis

°(M)Lei Yin[1], Masaharu Yoshioka[1], Shinjiro Hara[1], Hokkaido University[1],
E-mail: yinlai2612@yahoo.co.jp

Abstract

Automatic construction of the paper/figure databases can support the development of data-driven science. An in-house database system aimed for a specific research group has been proposed in [1]. In this report, we show an application to the nanocrystal device development domain of this database system, and then develop a framework of research trend analysis to help users analyzing research papers by conducting multifaceted analysis using domain terms.

Application of the Database System and Research Trend Analysis

The paper data (PDF files) we used for the proposed system is from a nanocrystal device development group RCIQE[1]. We use pdffigures2[2] and GROBID[3] to extract textual contents and figures from PDF files. Since the difficulty of detecting the location of images in PDFs, some figures were missing. Then we choose a dictionary-based approach for extracting terms from the textual contents. Tables 1 and 2 show the extraction results.

Table 1. Number of PDF files and extracted contents.

PDF files	Fulltext	Caption	Figure
9,980	9,852	37,826	28,441

Table 2. Number of terms extracted from the fulltexts.

Material	Parameter	Method	Characteristic	Artifact
345,921	590,958	101,921	44,187	63,597

Then, we conduct the framework of research trend analysis based on Burst Detection [2]. Figures 1 and 2 show the trends of “Al2O3” based on different sources, i.e., major international conferences of “MNC” and “SSDM”. The trends based on two sources are highly similar, and going upward after 2012/2013. GaN growth was first popular on “Al2O3”, sapphire, substrates before around 2000. Therefore, the bursty trend after 2012 may be inferred that there’s a new technology about Al2O3 attracted some attention.

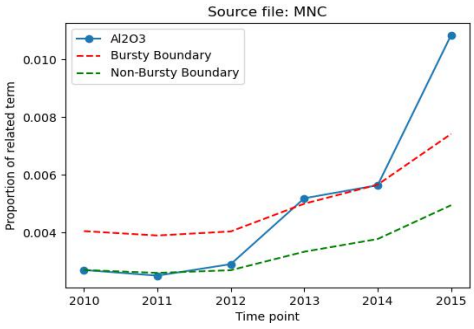

Fig. 1 Trend of “Al2O3” based on MNC.

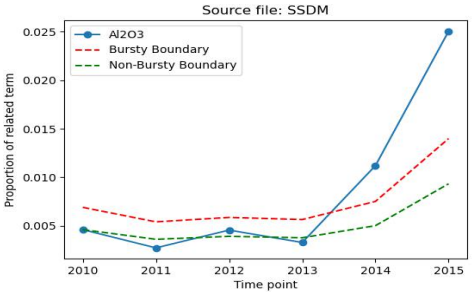

Fig. 2 Trend of “Al2O3” based on SSDM.

Conclusion

After communicating with the domain expert, we believe the proposed system will be useful for the researchers, but it’s still necessary to improve the performance of term extraction and research trend analysis. The dictionary need to be improved continuously, and it can be realized by interaction with users. Another problem is the terms which show a “bursty level” in research trend analysis. Since a high frequency term may coming from papers published by just one single research group, we need to confirm the source of the term to ensure that the term is really popular in a specific year.

References

[1] 吉岡真治, 大久保好章, 尹磊, 原真二郎, 鈴木晃, 高山英紀, 石井真史: 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. 第 80 回応用物理学会秋期学術講演会, 19a-B01-9, 2019.

[2] J. Kleinberg. Bursty and hierarchical structure in streams. In Proc. 8th SIGKDD, pp. 91–101, 2002.

[1] https://www.rciqe.hokudai.ac.jp/
[2] http://pdffigures2.allenai.org/
[3] https://grobid.readthedocs.io/en/latest/