Topic Evolution and Emerging Topic Analysis Based on Open Source Software

Xiang Shen¹†, Li Wang²

¹National Science Library, Chinese Academy of Sciences, Beijing 100190, China
²Department of Library, Information and Archives Management of University of Chinese Academy of Sciences, Beijing 100190, China

Abstract

Purpose: We present an analytical, open source and flexible natural language processing and text mining method for topic evolution, emerging topic detection and research trend forecasting for all kinds of data-tagged text.

Design/methodology/approach: We make full use of the functions provided by the open source VOSviewer and Microsoft Office, including a thesaurus for data clean-up and a LOOKUP function for comparative analysis.

Findings: Through application and verification in the domain of perovskite solar cells research, this method proves to be effective.

Research limitations: A certain amount of manual data processing and a specific research domain background are required for better, more illustrative analysis results. Adequate time for analysis is also necessary.

Practical implications: We try to set up an easy, useful, and flexible interdisciplinary text analyzing procedure for researchers, especially those without solid computer programming skills or who cannot easily access complex software. This procedure can also serve as a wonderful example for teaching information literacy.

Originality/value: This text analysis approach has not been reported before.

Keywords  Topic evolution; Emerging topics; Text mining; Thesaurus; VOSviewer

1  Introduction

For scientific discovery and technological foresight, mapping the evolution of topics and detecting emerging topics in science and technology has been of interest to governments, companies, and individual scientists for many years.

With the development of natural language processing and information extraction techniques, text mining research is attracting more and more attention. Swan and
Jensen (2000) presented the TimeMines system, which can detect, rank, and group terms or keywords from data-tagged free text corpora based on their statistical properties. Havre et al. (2002) developed the ThemeRiver visualization system to depict thematic variations over time within a large collection of documents. These systems, however, are neither open source nor freely available to other researchers. In recent years, many methods for topic evolution research based on probabilistic topic models have been published, and such models include, in particular, the seminal topic models of PLSA (probabilistic latent semantic analysis) and LDA (latent dirichlet allocation). Zhou, Yu, and Hu (2017) reviewed notable research on topic evolution based on probabilistic topic models from multiple aspects over the past decade. A certain level of computer programming skills is required, however, to establish and improve probabilistic topic models. It is thus necessary to standardize the synonyms and abbreviations for identical terms or to stop using certain terms to improve analysis results and to make them more accurate, but this is difficult to automate.

VOSviewer is an open source computer program for creating, visualizing, and exploring bibliometric maps (Van Eck & Waltman, 2010). It has been widely used by researchers from different research domains, such as entrepreneurship (Sassetti et al., 2018; Santos, Marques, & Ferreira, 2018), road safety (Zou, Yue, & Vu, 2018), biochemical engineering (Khudzari et al., 2018), psychology (Zyoud et al., 2018), knowledge management (Gaviria-Marin, Merigo, & Popa, 2018) and medical informatics (Chen et al., 2019). Since the release of version 1.4.0, based on the Apache OpenNLP toolkit (http://incubator.apache.org/opennlp/) and linguistic filter, extensive text mining functionality has been included in VOSviewer (Van Eck & Waltman, 2011). Little research, however, has mentioned both the text mining and thesaurus in VOSviewer’s data-cleaning functionality. Meanwhile, VOSviewer offers a direct visualization of a keywords map, giving a vague impression about temporal evolution and assigning colors to terms according to the average year of the publications that use them. The temporal evolution map for “perovskite solar cells research” generated directly by VOSviewer is shown in Figure 1; it only includes the dominant keywords and can neither detect nor reveal emerging topics in great detail.

In this paper, through natural language processing techniques, information extraction techniques and thesaurus functionality provided by the open source bibliometric software VOSviewer, as well as some general functions provided by the easily accessible software Microsoft Office, we develop a flexible approach for topic evolution and emerging topic analysis on all kinds of data-tagged text. Our approach can be applied by researchers in both the bibliometric and scientometric communities, as well as in specific science and technology research domains; in addition, computer programming skills are not necessarily required.
2 Method

The mapping and clustering results of VOSviewer can be saved as map and network files. These file types can be viewed and edited using a text editor or a spreadsheet program (e.g. Excel) (Van Eck & Waltman, 2018). We make the most of the map file in our method, since the map file contains information about the items in a map, and items are characterized by a number of attributes. Each column in a map file corresponds to an attribute, as Figure 2 shows. The “label,” “cluster,” “weight <Total link strength>,” and “weight <Occurrences>” data in the map files are applied during topic evolution and emerging topic analysis in this paper, where “label” represents the terms of the research domain. For a given item, the “weight <Links>” indicates the number of links of an item with other items, whereas the “weight <Total link strength>” indicates the total strength of the links of an item with other items.

The general approach includes:

- Data collection: Text data can be collected from commercial scientific literature databases like Web of Science and Scopus, and they can also be collected from
project and newspaper websites. Through text pre-processing, both scientific literature and non-scientific literature can be analyzed. When putting the full text context in the “title” or “abstract” attributes, text mining analysis can be theoretically done in full text.

- Data slicing: To construct time series text data for trend analysis, text data should be sliced into several subperiods with an appropriate time span. This time span can be selected flexibly according to the amount of literature and the purpose of the research.

- Thesaurus construction: For a better analysis performance, a thesaurus file should be established during the pre-text mining procedure to merge terms and clean data. A thesaurus file has a “label” column and a “replace by” column, and the content in the “label” column will be replaced by the context in the “replace by” column. When the “replace by” column is empty, the content in the corresponding “label” column will be ignored (Van Eck & Waltman, 2018). This function can be utilized to stop the use of terms. In addition, some repeated construction processes are needed for an accurate thesaurus.

- Text mining and map files generation: Based on the thesaurus file, every meaningful text mining result by VOSviewer is saved as a map file for further analysis.

- Map files treatment and analysis: To identify the emerging terms of each subperiod, the “LOOKUP” function of Excel (Microsoft Office, 2019) is applied. By searching the “label” column in the map files, comparing the “label” column to previous subperiods and returning the information in the “weight <Occurrences>” column, emerging terms are detected. In addition, the content “#N/A” means that the relevant label or term does not exist in the compared subperiod’s text data.
• Visualization: This topic evolution and emerging topic analysis can be further visualized using other pieces of open source and easily accessible software such as Microsoft Office PowerPoint and Excel. Variations of term occurrences can also be accounted for and visualized, and future trends of terms can be forecasted by analyzing the increase or decrease of term occurrences.

3 Application

To show our analytical topic evolution and emerging topic method in a detailed and intuitive way, the “perovskite solar cells” (PSCs) research domain is used as an example.

3.1 Data collection

Data collection was conducted on December 23, 2018, using the Scopus database. The central theme in this study consisted of research articles containing PSCs in their titles, abstracts, or keywords. The query string used for this search was: TITLE-ABS-KEY (“perovskite solar cells”) AND DOCTYPE (ar OR re) AND PUBYEAR < 2019. This query string resulted in 7,125 documents, and the oldest publication was from 1999, as Figure 4 shows.

![Graph of annual number of research articles on PSCs indexed in Scopus from 1999 to 2018.](image)

Figure 4. The annual number of research articles on PSCs indexed in Scopus from 1999 to 2018.
3.2 Data slicing

Regarding PSCs, the total amount of literature in the subperiod and the need to highlight temporal evolution as much as possible were taken into consideration during slicing. The total number of articles published before 2013 were small and sliced as one subperiod, and the articles published between 2014 and 2018 were sliced as a separate subperiod, since the amount of this literature was appropriate for a yearly analysis.

3.3 Thesaurus construction

An example of a thesaurus file for VOSviewer is shown in Figure 5. These “label” terms like “cm2,” “c cm2,” and “future” are stop-words. We figure there are two different ways to conduct pre-text mining and to generate map files. One is to pre-text mine the entire set of text data at once, and the other is to pre-text mine each subperiod’s text data. In our opinion, the second method is more reasonable, as the minimum occurrences of keywords to be analyzed can be set flexibly and the distribution of keywords in the entire set of text data is quite different from that of each subperiod’s text data. In the PSCs example, pre-text mining is performed for each subperiod’s text data, and the resulting thesaurus file is used and updated each time until a final version is achieved.

| label                          | replace by                                      |
|--------------------------------|------------------------------------------------|
| x ray diffraction              | x ray diffraction analysis                      |
| perovskite solar cell          | perovskite solar cells                         |
| mhb                            | morphotropie phase boundary                     |
| htm                            | hole transport material                         |
| cm2                            |                                               |
| compact ti02                   | compact ti02 layer                             |
| c cm2                          |                                               |
| dft calculation                | density functional theory                      |
| dssc                           | dye-sensitized solar cells                      |
| dsscs                          | dye-sensitized solar cells                      |
| efficient planar heterojunction solar cell | efficient planar heterojunction perovskite solar cell |
| electricity conversion efficiency | electric conversion efficiency                  |
| electronic property            | electrical property                             |
| eis                            | electrochemical impedance spectroscopy          |
| future                         |                                               |
| low temperature solution process | low temperature solution processing             |
| mw cm                          | mw cm2                                         |
| mm dimethylformamide           | N,N-dimethylformamide                           |
| dmf                            | N,N-dimethylformamide                           |

Figure 5. A part of the thesaurus constructed for PSCs in VOSviewer.
3.4 Text mining and map files generation

With the final version of the thesaurus file, a series of subperiod text data are processed using the natural language processing, information extraction, co-occurrence term analysis, clustering, and visualization techniques offered by VOSviewer, and terms are extracted from titles and abstracts. The corresponding map files are generated and saved for further analysis.

3.5 Map files treatment and analysis

Each cluster in the map file of every subperiod’s text data can be seen as a topic. By analyzing the total link strength of terms in each cluster, we label the cluster with the most dominant and meaningful terms as the topic, and these topics are shown in Figure 6. As the most dominant terms are not always the ones that best describe a cluster’s general nature, a careful manual review by a researcher on the scientific domain could be useful. The number of topics in each subperiod’s text data is dependent on the minimum occurrences of a term to be analyzed, which we set.

![Figure 6. Topic evolution for PSCs.](image)

The emerging terms in PSCs were detected by the “LOOKUP” function of Excel, where “#N/A” means that the label or the term does not exist in the compared
subperiod’s text data. As shown in Figure 7, “inorganic perovskite solar cell” was one of the emerging terms in 2017. In addition, “halide perovskite” and “gap” were two of the emerging terms in 2016, and “tin oxide” “module” “tandem solar cell” and “photodetector” were four of the emerging terms in 2015.

| label                                      | weight<Occurrences> |
|---------------------------------------------|---------------------|
| halide perovskite                           | 173 135 107 N/A N/A N/A |
| gap                                         | 148 112 16 N/A N/A N/A |
| spiro ometad                                | 102 107 79 N/A N/A 2  |
| tin oxide                                   | 78 26 49 8 N/A N/A  |
| inorganic perovskite solar cell             | 71 40 N/A N/A N/A N/A |
| pcblm                                       | 70 57 46 N/A 5 N/A |
| simulation                                  | 69 N/A N/A N/A N/A 2 |
| inverted perovskite solar cells             | 68 41 18 7 N/A N/A |
| density functional theory                   | 67 46 43 12 N/A N/A |
| calculation                                 | 64 45 46 13 6 N/A |
| diode                                       | 61 69 23 N/A N/A N/A |
| perovskite structure                        | 60 53 36 N/A 12 N/A |
| lead halide perovskite                      | 59 N/A 39 N/A N/A N/A |
| pedotpp                                     | 55 52 45 N/A 2 N/A |
| module                                      | 54 33 21 11 N/A N/A |
| tandem solar cell                           | 53 43 35 8 N/A N/A |
| photodetector                               | 52 39 21 3 N/A N/A |

Figure 7. Detecting emerging terms using Excel’s “LOOKUP” function.

### 3.6 Visualization

We used Microsoft Office PowerPoint and Excel to visualize the analysis results. The Top 10 emerging terms for PSCs were picked by researchers in descending order of occurrence, and they were visualized by PowerPoint, as shown in Figure 8.

In addition, we also used Excel to visualize variations in term occurrence and to forecast the future trends of terms. As Figure 9 shows, by analyzing the increase or decrease of term occurrence in PSCs research, the “module” “photodetector” and “flexible perovskite solar cells” topics may receive more attention from researchers in the near future, while the “carbon nanotube” topic may receive less attention.
4 Verification and discussion

The emerging terms detected by our method were verified by comparing them to the search results in the Scopus database and by combining “perovskite solar cells” with emerging terms. Generally, few or only a small percentage of articles were published before the year for which our method detected the term in use. Taking the emerging term “flexible perovskite solar cells” in 2015 as an example, there was only one paper published in 2014 (Jung, Williams, & Jen, 2014), but the number of papers published containing “flexible perovskite solar cells” rose to 21 in 2015, indicating that this method is quite effective.

For better performance analyzing emerging terms and topic evolution by this approach, the following ideas should be considered: (1) an appropriate threshold of minimum occurrences of a term to be analyzed should be set, and it should be as small as possible; (2) a detailed thesaurus file should be constructed, and it should be as detailed as possible. It takes time to merge the synonyms and abbreviations of identical terms, and it is quite difficult to clarify the relationships between the upper and lower levels of a concept, even for specific scientists in a particular research domain. Certainly, with the development of natural language processing, semantic processing and text mining techniques, the performance of topic evolution and emerging term analysis will improve.

5 Conclusion

In this paper, through natural language processing, information extraction, thesaurus utilization, co-occurrence term analysis, clustering techniques, and the use of the LOOKUP function, a flexible approach for topic evolution, emerging topic...
analysis and visualization based on VOSviewer and Microsoft Office has been presented. As an example, the topic evolution and emerging topics of perovskite solar cells research was analyzed. The new attempts and applications of open source bibliometric software in this paper could be a valuable supplement to text mining analysis. Also, they could be quite effective for interdisciplinary researchers, especially those with low or no skills in computer programming. In addition, the proposed approach can be used to analyze large amounts of text data, including both scientific text and non-scientific text (e.g. project files, policy files, and newspaper articles). Finally, this process of analysis can be optionally adjusted by researchers according to their specific needs.

Acknowledgments

We would like to thank Professor Van Eck and Professor Waltman for developing the useful, open source software VOSviewer, and we would also like to thank the anonymous reviewers of this paper for their valuable comments.

Author contributions

Proposing the research problems: Xiang Shen (shenx@mail.las.ac.cn); performing the research: XS; designing the research framework: XS, Li Wang (wangli@mail.las.ac.cn); collecting and analyzing the data: XS; writing and revising the manuscript: XS, LW.

References

Chen, D.H., Zhang, R.T., Zhao, H.M., & Feng, J.Y. (2019). A Bibliometric Analysis of the Development of ICD-11 in Medical Informatics. Journal of Healthcare Engineering, 2019(10059), 1649363.
Gaviria-Marin, M., Merigo, J.M., & Popa, S. (2018). Twenty years of the Journal of Knowledge Management: A bibliometric analysis. Journal of Knowledge Management, 22(8), 1655–1687.
Havre, S., Hetzler, E., Whitney, P., & Nowell, L. (2002). Theme River: Visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics, 8(1), 9–20.
Jung, J.W., Williams, S.T., & Jen, A.K.Y. (2014). Low-temperature processed high-performance flexible perovskite solar cells via rationally optimized solvent washing treatments. RSC Advances, 4(108), 62971–62977.
Khudzari, J.M., Kurian, J., Tartakovsky, B., & Raghavan, G.S.V. (2018). Bibliometric analysis of global research trends on microbial fuel cells using Scopus database. Biochemical Engineering Journal, 136, 51–60.
Microsoft Office. LOOKUP function. Retrieved from https://support.office.com/en-us/article/LOOKUP-function-446D94AF-663B-451D-8251-369D5E3864CB
Santos, G., Marques, C.S., & Ferreira, J.J. (2018). A look back over the past 40 years of female entrepreneurship: Mapping knowledge networks. Scientometrics, 115, 953–987.
Sassetti, S., Marzi, G., Vincenzo, C., & Cavaliere, V. (2018). Entrepreneurial cognition and socially situated approach: A systematic and bibliometric analysis. Scientometrics, 116, 1675–1718.

Swan, R., & Jensen, D. (2000). TimeMines: Constructing timelines with statistical models of word usage. In the Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 73–80). New York, USA: ACM Press.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84, 523–538.

Van Eck, N.J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. ISSI Newsletter, 7, 50–54.

Van Eck, N.J., & Waltman, L. (2018). VOSviewer Manual. Retrieved from https://www.vosviewer.com/download

Zhou, H.K., Yu, H.M., & Hu, R. (2017). Topic evolution based on the probabilistic topic model: A review. Frontiers of Computer Science, 11, 786–802.

Zou, X., Yue, W.L., & Vu, H.L. (2018). Visualization and analysis of mapping knowledge domain of road safety studies. Accident Analysis and Prevention, 118, 131–145.

Zyoud, S.H., Sweileh, W.M., Awang, R., & Al-Jabi, S.W. (2018). Global trends in research related to social media in psychology: Mapping and bibliometric analysis. International Journal of Mental Health Systems, 12(1), 4.

This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (http://creativecommons.org/licenses/by-nc-nd/4.0/).