Topic evolution analysis of radar research using a dynamic topic model based on latent Dirichlet allocation

Xiaoguang Huang¹, Hui Fang¹ *

¹School of Electronic Science and Engineering, Nanjing University, Nanjing, Jiangsu, 210023, China

*Corresponding author’s e-mail: fanghui@nju.edu.cn

Abstract. This work applies a dynamic topic model based on latent Dirichlet allocation to investigate the evolution of radar research and applications. To address the problem that processing a large number of papers exceeds the capacity of computers, we propose a method to sample documents according to their citation count. Of 104,428 sampled articles, 108 research topics are extracted. The evolution of topics is analysed from two dimensions: content and strength. The results show that radar technology, which arose mainly from physics and engineering science, has been widely applied in studies in the civil engineering, geographical, environmental, meteorological, geological, agricultural, ecological, among others. In the long-term development of radar, new technologies have continuously been produced. At the individual topic level, the research content has changed over time. The development objectives of radar systems are to enhance functionality, extract more information and improve clarity.

1. Introduction

The number of academic publications has continued to increase, and interdisciplinary research has led to an increasing diversity of topics in the literature. With a vast number of scientific publications, determining how to comprehensively and quickly assess development in various disciplines has become increasingly important. Researchers have applied topic evolution analysis[1] to the scientific literature to grasp trends and various hotspots over time. Methods, such as co-citation analysis, co-word analysis, and word frequency analysis, have been applied in topic evolution analysis.

With the development of data mining technology, topic modelling methods have been gradually applied in topic evolution analysis. Topic modelling methods have advantages over traditional methods. The semantic features and distribution characteristics of topic words in the text are considered to mine semantic associations and identify the semantic topics among words. The latent Dirichlet allocation (LDA) model[2] is an effective topic modelling method for statistically analysing document topics[3] and overcomes the drawbacks of previous models. The LDA model has a clear hierarchical structure and is a complete generative model that can effectively mine the implicit internal relations associated with semantic information. This approach has been widely used. For example, LDA was used to assess traffic topics from Twitter data and obtain valuable information about traffic[4], and to analyse people’s opinions and emotions towards entities[5]. LDA analysis has also been used in research topic evolution analyses, such as analysing the evolution of research in long-term evolution (LTE) technology[6], and biochemistry[7].

In this paper, the LDA model is applied to radar research, which has a long history. Besides military usage, radar technology has been widely used in civil activities, with considerable benefits and convenience. Many technologies and application scenarios in the field of radar technology have changed
dramatically. Due to the continuous improvement and interdisciplinary development of radar technology, the number of related articles in this field has continued to increase. In the context of this large body of research, it is important to understand the development trend of radar technology.

We used a dynamic topic model (DTM) based on LDA to conduct topic modelling for publications in the field of radar technology and then perform topic evolution analysis and popular topic mining to provide an objective reference for research in this field. The analysis results reveal the development status of the radar field.

2. Methods

2.1. LDA model
LDA is an unsupervised machine learning technique that is now extensively used in identifying latent topic information in collections of documents. LDA is a Bayesian model that consists of three layers: topic, document, and word layers. Each document is represented by the probability distribution of the topics, and each topic is represented by the probability distribution of the words. The generation process of each document in a corpus is: (1) For each document, select a topic from the distribution of topics; (2) Sample a word from the distribution of the words corresponding to the chosen topic; (3) Repeat (1) and (2) for all words in the document.

2.2. DTM based on the LDA model
Topic evolution refers to the process of literature topics changing over time. Suppose that the time series of a corpus is \([C_1, C_2, ..., C_T]\), where \(C_t\) is the corpus in time window \(t (t = 1, 2, ..., T)\) and consists of a series of documents. In the LDA model, each topic has a corresponding topic-word distribution. However, in the process of topic evolution, the topic-word distribution changes over time. This work adopted discrete time evolution DTM because it can identify the changes in topic content in an academic research corpus over time[8]. Therefore, the literature corpus is divided according to the publication year to reveal the topic evolution.

The DTM was implemented using Python’s GenSim library, and the parameters were set according to the GenSim standard.

2.3. Data set
In this investigation, the titles and abstracts of English publications in the radar field were collected as the research data set. The data were collected from Scopus with the Scopus API. Data collection was completed prior to March 31, 2020. The searched keyword was “radar”, the publication year was before 2019, and the document type was “article”, “conference paper” or “letter”. Some papers retrieved using the keyword “radar” were not related to radar. These papers were removed with a customized Python program to exclude any paper if it contained any of certain phrases (listed in the Supplementary Materials) in the title or abstract. The total number of papers used here was 216,802.

An analysis of all these documents would exceed our computational capacity. To address this type of problem, Shao et al. randomly selected documents for analysis[9]. This work selected papers according to their citations, because citations can reflect the popularity of the corresponding research to a certain extent. The papers from each year were sorted in descending order of number of citations. If there were fewer than 1000 papers in a certain year, all of them were included in the analysis. If there were more than 1000 papers but fewer than 2500 papers in a certain year, we selected papers with at least as many citations as the 1000th paper. If there were more than 2500 papers in a certain year, we selected papers with at least as many citations as the paper ranked in the 40th percentile based on citations. After sampling, a total of 104,428 papers were obtained. The distribution of the number of papers per year is shown in figure 1. In the corpus, there was only one publication in 1945, the year for which the first document on radar with an abstract was available in Scopus when we downloaded the data. Therefore, this publication was combined with 16 publications in 1946.
2.4. Determining the number of topics for DTM analysis

In this work, the number of topics in the DTM is determined according to perplexity, which is calculated for different numbers of topics[2]. Perplexity is the integral degree of uncertainty regarding the topic of a document. The lower the perplexity value is, the better the clustering effect.

For the dataset used in this work, the perplexity decreases with increasing number of topics before a minimum point and increases slowly after that minimum point. The number of topics is set to 108, corresponding to the minimum perplexity.

DTM modelling was then performed for the sampled corpus to analyse radar topics. We obtained the document-topic distribution and topic-word distribution for the 108 topics in each time slice.

3. Results and discussion

3.1. Topic content analysis

Based on the trained model, the topic-word distribution of the 108 topics corresponding to each time slice can be obtained. For example, Table 1 lists the word distribution of some topics in 2018. The words are ranked in descending order of word probability, and the top ten words are displayed.

| Topic 1          | Topic 2 | Topic 3          | ... | Topic 107 | Topic 108 |
|------------------|---------|------------------|-----|-----------|-----------|
| microwave        | image   | forecast         | ... | fusion    | wind      |
| composite        | noise   | observation      | ... | data      | speed     |
| frequency        | method  | assimilation     | ... | spectral  | direction |
| Ghz              | propose | data             | ... | detection | measurement |
| absorption       | segmentation | model       | ... | hyperspectral | data   |
| Loss             | speckle | scheme           | ... | image     | field     |
| absorber         | filter  | system           | ... | sensor    | retrieval |
| structure        | pixel   | radar            | ... | fuse      | ocean     |
| resonator        | sar     | ensemble         | ... | multisensor | farm     |
| application      | edge    | simulation       | ... | color     | sea       |

Word clouds are used to intuitively show the main content associated with a topic. In a word cloud, font size reflects the probability of a word in a given topic. The greater the word probability is, the larger the font size. Each picture shows the 20 words that contribute most to the topic. The word clouds of all topics and time slices are provided in the Supplementary Materials.

The topic can be inferred from the word distribution. For example, figure 2 shows the word distribution for topic 7 in 1998. According to the figure and the word clouds for other years, this topic is inferred as “automatic recognition of radar targets”.

Figure 1. Distribution of the literature used in the analysis by publication year.
Table S1 (provided in the Supplementary Materials) lists the inferred results for the 108 topics. The authors inferred these topics alone. When the title of a certain topic could not be clearly determined, the authors read the papers containing the words in the word clouds to understand the content of the topic and then inferred the topic. This took substantial time. The authors eliminated differences in topic inference through discussion.

![Figure 2. Word distribution for topic 7 in 1998.](image)

3.2. Distribution of radar topics across subject areas
The radar topics mainly involve the following subject areas (classified by Scopus): engineering, computer science, Earth and planetary sciences, physics and astronomy, mathematics, environmental science, social sciences, agricultural and biological sciences, energy, and medicine. Studies in engineering, computer science, mathematics, and physics have contributed to the development of radar technology. Radar technologies are mainly applied in research involving Earth and planetary sciences, environmental science, agricultural and biological sciences, physics and astronomy, and medicine. For example, radar is based on physical principles, such as the transmission and reflection of electromagnetic waves. In turn, radar can be used in physics and astronomy, such as in investigations of gravity waves. In Scopus, the main subject area in the social sciences includes “geography, planning and development”, which covers most papers that apply radar in the social sciences. In addition, ground-penetrating radar can be applied in studies of “archeology”, which is generally classified as a social science.

3.3. Categories of radar topics
According to the results of subject inference, the topics can be divided into two categories, namely, radar development and radar application.

Radar development can be classified from four perspectives; the first is basic research on radar, which mainly involves solving problems in radar development, such as theoretical analysis and calculation, simulation, signal processing, and interference reduction. The second perspective involves the radar technology developed for general purposes, such as synthetic aperture radar, Doppler radar, phased-array radar, and lidar. The third perspective involves specialized radars; for example, precipitation radar and weather radar were specifically developed to detect precipitation and monitor weather. The final perspective involves combining radar with other technologies. For example, a radar satellite is a remote sensing satellite that carries a radar system for Earth observation.

Nearly half of the 108 topics involve radar applications. Each application topic includes one or more radar technologies; some use general-purpose radars. For example, “geological applications of radar” and “application of radar in underground exploration” use “ground-penetrating radar”. Some radar application topics use specialized radars. For example, “radar observation of mesoscale convective systems” and “radar monitoring of meteorological disasters” both utilize “weather radar”. Some radar application topics use radar combined with other technology. For example, “radar detection of floods”, “Arctic sea ice monitoring” and “radar remote sensing” are applications of “radar satellite”.

3.4. Topic content evolution analysis
By analysing the changes in the word distribution of a topic in all time slices, we can identify the changes in the research emphasis and technology for relevant research directions. For example, in the word clouds of topic 45, the probability of “radar” was high before 1980; “Synthetic” and “aperture” began
to appear in 1963; and after 1980, the probability of “SAR” was high. This finding reflects the transition from ordinary methods to synthetic aperture radar (SAR) imaging in radar imaging.

Consider the word clouds of topic 100. Before 1980, “lunar” had a high probability. From 1990 to 2000, “Venus” had a high probability. However, after 2005, “lunar” again had a higher probability than “Venus”. This result reflects the change in the focus of celestial bodies to be investigated using radar technology.

The change in research content for a topic can be reflected by a word evolution curve. For example, figure 3 shows the evolution of radar technologies in the application in topic 101. In 1990, the probability of “SAR” exceeded that of “radar”. In 2007, “PalSAR” (“phased-array-type L-band SAR”) began to have the highest probability. After 2013, “PolSAR” (“polarized SAR”) was the most important technology. This finding indicates that applications of radar in the estimation of vegetation coverage have gradually shifted from radar to SAR systems and then to PolSAR. In this figure, “SAR” has a probability higher than 0 in the years before it was developed due to LDA error.

![Figure 3. Evolution curves of radar technologies applied in vegetation coverage type estimation.](image)

3.5. Topic strength evolution analysis

Topic strength[10] describes the degree to which a topic is researched in a certain time window; that is, the greater the proportion of a topic in the document set of a certain time window, the greater is the strength of the topic. According to the strength of each topic in different time slices (provided in the Supplementary Materials), we find different types of evolution trends for radar topics.

An obvious upward strength trend reflects an increase in research interest. For example, InSAR has the advantages of strong continuous observation ability, a high imaging resolution and accuracy, and wide coverage; therefore, it is widely used in topographic map imaging and crustal deformation research[11]. Correspondingly, the strength trend of topic 22 has continuously increased. The topic evolution curve of topic 77 suggests that the public has paid increasing attention to problems related to Earth climate change.

The strength of topic 45 increased sharply in the 1970s and has remained at a high level. This trend indicates that radar imaging is still a popular research topic. Topic 101 displays a similar strength curve, suggesting that people attach importance to environmental protection.

An obvious downward strength trend reflects a relative decrease in research interest compared to that of all radar studies. For example, ranging, angle measurement and radar backscattering were studied early in the radar field and have relatively matured. Correspondingly, the strength trends of topics 43, 72 and 94 are decreasing.

The strength trend in some topics increased at one time but then decreased. These topics, such as topics 39, 48, 89, and 93, are mainly related to specialized radars. The strength trend in topic 26 “ground-penetrating radar” and topic 97 “Doppler radar” also increased and then decreased.

A relatively overall stable strength trend indicates that developments in these research areas have been relatively stable. For example, radar system design is critical for various radar applications. Thus, the strength evolution of topic 73 has been stable. Radar space-time adaptive processing technology can
effectively suppress clutter and improve the detection performance of radar. Correspondingly, the strength of topic 78 has remained steady.

3.6. The changes in popular topics over time
The strength distribution for all the years is provided in the Supplementary Materials. In early years, radar research concentrated on a few topics. With the emergence of many topics related to radar, strength has been distributed more evenly among radar topics in recent years. We determine a topic to be a “hot topic” in a certain year if it ranks in the top five topics in that year.

In radar basic research, “radar system design” is a hot topic. “Circuits in radar” is a hot topic sometimes. “Radar cross-section analysis” has long been a hot topic. “Radar ranging” and “radar data processing” were hot topics in some years in the 1960s, 1970s and 1980s. The “design of ultrawideband antennas in radar” has become a hot topic since 2009.

“Gaussian models in radar systems” was a hot topic until 2012. “Solving mathematical problems for radar systems” and “radar ambiguity function” were hot topics in some early years. “Simulation verification in radar research” has become a hot topic since 2005. These results suggest that improved computational and simulation methods have replaced traditional mathematical analyses in solving radar problems.

In radar technology, “monopulse radar” ranks high until 1971. “VHF radar” was a hot topic before 1960. “Ground-penetrating radar” was a hot topic between 2000 and 2004. “Incoherent radar” was a hot topic in some years between 1958 and 2004. “MIMO radar” has been a hot topic in recent years. “Radar imaging” has been a hot topic since 1988. Topics 2 and 9, which involve solving problems related to SAR, have emerged as hot topics in recent years.

Among the specialized radars, “weather radar” was a hot topic between 1978 and 1984, “aurora radar” was a hot topic between 1998 and 2002, and “meteor radar” was a hot topic in some years between 1997 and 2009.

In radar applications, “application of radar in air traffic control” was a hot topic until 1987. “Radar monitoring of meteorological disasters” was a hot topic between 1981 and 2006.

4. Conclusion
In this paper, we use a DTM based on LDA to conduct a topic content evolution analysis of radar research. To address the problem that processing a large number of papers exceeds the capacity of computers, we propose a method that samples scientific publications based on the citation count. Overall, 108 research topics were extracted from the documents.

We discussed the evolution of topics from the dimensions of content and strength. By studying the changes in the word distribution of the topics in different time slices, we identified the changes in the research focus in related research directions. According to topic strength, the evolution of radar topics over time has displayed rising, falling and stable trends. Hot topics in each period were explored according to the strength value. By setting the time slice length as a year, the times of growth for some topics with rapid development periods were identified according to topic strength curves.

Analysis using the DTM based on LDA provided a panoramic perspective of radar research, including the development and application of radar. The development of radar includes basic research on radar and radar technology. The latter includes general-purpose radar technology, specialized radar technology, and radar technology combined with other technology. Nearly half of the radar topics involved applying these radar technologies in many fields.

In the long term, the function of radar systems has been greatly enhanced. Notably, this function has evolved from detecting objects and determining a target location to depicting two-dimensional and three-dimensional scenes, as well as from obtaining surface object information to extracting information from inside objects. The development trends of radar involve improved clarity and the extraction of more information than ever before.
Supplementary Materials
The Supplementary Materials can be found in https://data.mendeley.com/datasets/m8wf9p7rc9/2.

References
[1] He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., Giles C.L. (2009) Detecting topic evolution in scientific literature: how can citations help? In Proceedings of the 18th ACM conference on Information and knowledge management. Hong Kong. pp. 957–966.
[2] Blei, D.M., Ng, A., Jordan, M.I. (2003) Latent Dirichlet allocation. J. Mach. Learn. Res., 3: 993–1022.
[3] Bian, W., Tao, D. (2009) Dirichlet mixture allocation for multiclass document collections modeling. In 2009 Ninth IEEE International Conference on Data Mining. Miami. pp. 711–715.
[4] Hidayatullah, A.F., Ma’Arif, M.R. (2017) Road traffic topic modeling on Twitter using latent Dirichlet allocation. In 2017 International Conference on Sustainable Information Engineering and Technology. Malang. pp. 47–52.
[5] Bashri, M.F.A., Kusumaningrum, R. (2017) Sentiment analysis using latent Dirichlet allocation and topic polarity wordcloud visualization. In 2017 5th International Conference on Information and Communication Technology. Melaka. pp. 1–5.
[6] Wang, B., Liu, S., Ding, K., Liu, Z., Xu, J. (2014) Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology. Scientometrics, 101: 685–704.
[7] Kang, H.J., Kim, C., Kang, K. (2019) Analysis of the trends in biochemical research using latent Dirichlet allocation (LDA). Processes, 7: 379.
[8] Blei, D.M., Laferty, J.D. (2006) Dynamic topic models. In Proceedings of the 23rd international conference on machine learning. Pittsburgh. pp. 113–120.
[9] Shao, Y., Zeng, Q.T., Chen, K.K., Shutes-David, A., Thielke, S.M., Tsuang, D.W. (2019) Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. BMC Med. Inform. Decis. Mak., 19: 128.
[10] Wang, J., Fan, Y., Zhang, H., Feng, L. (2021) Technology Hotspot Tracking: Topic Discovery and Evolution of China’s Blockchain Patents Based on a Dynamic LDA Model. Symmetry, 13: 415.
[11] Ren, X., Tang, J., Li, Z., Tong, G., Hu, D. (2008) Measuring Three-dimensional Deformation of Earth Surface Using InSAR. Hydrographic Surveying and Charting, 28: 45–48.