Data analysis and strategy evaluation in national defence science and technology tracking

KanLi
Teaching and research support centre, Dalian Naval Academy, Dalian, Liaoning, 116018, China
Corresponding author’s e-mail: likandl@126.com

Abstract. Science and technology tracking can provide powerful decision-making information support for science and technology management, and it is an important reference unit for national defence science and technology development planning and technical route. In the context of confidentiality, patent and industry competition, there are transmission barriers between organizations for systematic scientific and technological information, and a large amount of disorderly and excessive point information will lead to information disaster, lack of effective data and tracking methods for quantitative analysis, and the demand side is at a loss. Based on the combination of literature retrieval and data analysis, this paper selects the association between texts through co citation analysis, explains the relationship between tracking targets through co word analysis, and finally selects groups with similar or common characteristics as tracking targets. This paper explores a new way of science and technology tracking, which can provide reference for national defence science and technology tracking.

1. Introduction
National Defence Science and technology is the important support of national defence industry, the fundamental to research and produce new equipment, improve the industrial structure of national defence economy, and promote the development of national defence economy, and the cornerstone of the construction of the combat effectiveness of the army; national defence science and technology information is an important reference unit of national defence science and technology development planning and technical route, and an early information resource to carry out national defence science and technology research; National defence technology tracking is the process of obtaining and forming technical information. The purpose of tracking is to comprehensively and systematically collect and analyze national defence science and technology information data at home and abroad, grasp the current situation and development trend of frontier technology in related fields, determine new technical ideas, prevent low-level research, and improve the utilization efficiency of science and technology resources. Data mining usually refers to finding potential valuable information and patterns from existing massive data through algorithms and tools. It is a combination of database, machine learning and other technologies. It is the best technology to solve the current situation reporting disaster of "disorder and excess of data information leading to lack of knowledge". From the aspect of adaptability, the application of data mining logic in national defence science and technology information tracking is appropriate and feasible.
2. Formatting the title, authors and affiliations
With the rapid development of science and technology in the whole world, national defence science and technology has got unprecedented development, and the volume of national defence science and technology information has increased exponentially, which is contained in a variety of carriers. All countries in the world first apply the achievements of science and technology to the military field, so national defence science and technology has the characteristics of technology intensive and rapid update. The demand for national defence and the development trend of science and technology in the world determine the development direction of national defence science and technology. For reasons of confidentiality, patent or industry competition, there will be systematic information blockade between corresponding organizations or even between individuals within the same organization, but scattered information is distributed in Journal meetings and scientific and technological reports. The vast majority of national defence science and technology cutting-edge achievements have eliminated the directional information such as model, code, sensitive vocabulary, etc., which can be found in journals or conference papers in a timely manner in the form of papers after being deconstructed, implied, and commercialized. The institutions of higher learning, research institutes and industrial departments that undertake the task of national defence research all train graduate students, and the papers of graduate students are mostly technical difficulties or new ideas extracted from the process of completing the task undertaken by the tutor. The prerequisite for carrying out national defence pre research is to master domestic and international cutting-edge technologies, which is the significance of national defence science and technology tracking.

3. Logical method and application of data mining

3.1. Establish data source and dynamic database
This paper introduces the citation database Science Citation Index (SCI) of American Institute of scientific information, the engineering index (EI) published by American engineering information company, and the index to scientific & Technical P (index to scientific & Technical P) published by American Institute of scientific information With the help of providing periodical subscription service for US military agencies, EBSCO publishing (EP) has been developed. As well as China's CNKI, China info, VIP database, Superstar Digital Library and other databases, special organizations can also introduce internal journals, special groups of erudite papers and scientific reports.

3.2. Screening target group (excluding interference data)
A large number of interference data are eliminated from a large number of data sources.

The first step is to select high impact factor (if) journals. That is, the total number of citations of papers published in the first two years of a journal in the reporting year divided by the total number of papers published in the two years. The calculation formula is:

\[ IF_k = \frac{n_{k-1} + n_{k-2}}{N_{k-1} + N_{k-2}} \]

Where: FI is the influencing factor, k is the year, N is the number of papers published in the journal, and n is the number of cited papers in the journal.

The second step is to select influential journals. Among the above journals with high impact factors, the journals with large impact index and strong impact are selected. The algorithm formula is as follows:

\[ I_c = \sqrt{2 - \sqrt{(1 - I_f)^2 + (1 - T_c)^2}} \]

Formula (1) is substituted into formula (2), where If and TC are the influence factors and total cited frequency of a periodical after standardization by linear normalization method, and the influence factors are positively correlated with the cited frequency.
The third step is to select journals with high annual download rate. That is to say, the ratio of the total number of articles published and downloaded by the source journal in that year to the total number of documents published and online in that year. The calculation formula of annual download rate of Web is as follows:

$$W_{dr} = \frac{n}{N}$$  \hspace{1cm} (3)

Based on the law of people's information use for carriers, the above mining algorithm has selected the core group in the massive data to reduce the scope of tracking.

4. Technology tracking data analysis

On the premise of reducing the tracking range, we further use the co citation model and the co word model to analyze the data of national defence science and technology, which we call co-occurrence model. Through the analysis of co-occurrence phenomenon, we can know the strength and type of connection between things. Co citation means that two documents are cited by other documents at the same time. It is generally believed that the co cited literature has more or less similarity in subject, so the co cited times (i.e. co cited intensity) can measure the relevance of literature in content. Taking the selected core group literature as the analysis object, using cluster analysis and multi-dimensional scaling methods, and with the help of computer, the complex co indexing network relationship among many analysis objects is simplified into the relationship between several relatively small groups and expressed directly.

4.1. Mathematical expression of CO citation analysis

If there are n references and C = cij references, this citation relationship can form a matrix m, which is called citation matrix. Its matrix element is

$$c_{ij} = \begin{cases} 1 & \text{document i references j} \\ 0 & \text{No citation} \end{cases}$$  \hspace{1cm} (4)

Let $U = (1,1,...,1)^t$, then the number of references i has is

$$\sum_{j=1}^{n} c_{ij} = (CU)_i = (CC^t)_{ii}$$  \hspace{1cm} (5)

The number of citations owned by $j$ is

$$\sum_{i=1}^{m} c_{ij} = (UC^t)_j = (C^tC)_jj$$  \hspace{1cm} (6)

The co citation relationship between $i$ and $j$ is

$$\sum_{k=1}^{m} c_{ki}c_{kj} = (C^tC)_{ij}$$  \hspace{1cm} (7)

Among them, $U^t$ and $C^t$ are transposed matrices.

4.2. Mathematical expression of coinage analysis

From the perspective of CO words, words in literature can basically describe the relationship between text knowledge units or text book topics. The co-occurrence of two words in a document shows the relationship between the concepts they represent, and the co-occurrence of the same word pair shows the relationship between the documents of the same research topic. Compared with the bolometric method of subject words statistics, ranking and analysis of research hotspot, coinage analysis not only seeks high-frequency words, but also pays more attention to the relationship between these words, so as to better reflect the relationship between concepts.

Coinage model is one of the most important models in the field of natural language processing based on statistical methods. The common co word model directly uses the mutual information between words (feature items) and words (feature items) $MI(T_i,T_j)$ as a measure of the correlation between words. In general, mutual information is defined as
\[ \text{MI}(T_i, T_j) = \log \left( \frac{P(T_i, T_j)}{P(T_i)P(T_j)} \right) \] (8)

\[ P(T_i, T_j) = \frac{C(T_i, T_j)}{\sum C(T_i, T_j)} \]

Among them, \( C(T_i, T_j) \) is the frequency of the word \( T_i \) and \( T_j \) in the same window unit; \( C(T_i) \) is the frequency of the word \( T_i \).

There are some differences between the analysis of CO words and the analysis of quoted information. Co-citation analysis can only show that there is a certain association between texts, but not the specific content information about the research object; while CO word analysis can show the content information of texts. Co-citation analysis is the explanation of the correlation between concepts, while CO word analysis is mainly used for \( V \). In other words, the results of literature clustering produced by the two methods are different. In the best case, CO citation analysis can cluster literature according to discipline paradigm or according to the same research topic and hypothesis, while CO word literature clustering is more related to research topic, reflecting the relationship between research topics. When the analysis goal is to find the relative position information and the change trend of domain knowledge, we should choose the co-word analysis as the analysis unit; when the research goal is to analyze the content association of texts, understand the composition degree of intimacy between texts, we should choose the quoted literature information as the analysis unit.

### 4.3. Building data analysis model based on co-occurrence relationship

Select the literature group to be analyzed from the literature to be clustered. Suppose there are \( M \) articles and \( N \) subject words in this literature group. Each article is represented by a \( n \) dimension feature vector \( X = \{x_1, x_2, \ldots, x_n\} \), where \( x_i \) is the measurement of \( i \) subject words. In this way, we can get a matrix \( \epsilon \) of \( M \times N \), that is, \( \times N \) thematic words in \( M \).

\[ \pi_{jj'} = \begin{pmatrix} \frac{\epsilon_{i1}}{D_1} & \frac{\epsilon_{i2}}{D_2} & \cdots & \frac{\epsilon_{in}}{D_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\epsilon_{m1}}{D_m} & \frac{\epsilon_{m2}}{D_m} & \cdots & \frac{\epsilon_{mn}}{D_m} \end{pmatrix} \] (9)

\( \omega_{jj'} \) is the co-occurrence times of the subject words \( T_j \) and \( T_j' \), defined as

\[ \omega_{jj'} = \begin{cases} \sum_{i=1}^{n} \epsilon_{ij} \epsilon_{ij'} \quad & \text{if } j \neq j' \quad 1 \leq j \leq n, 1 \leq j' \leq n \\ 0 \quad & \text{if } j = j' \end{cases} \] (10)

Among them, \( j \) is the theme word; \( \epsilon_{ij} \) and \( \epsilon_{ij'} \) are the co-occurrence relationships between the theme words. After the above analysis, the correlation coefficient matrix \( \{T_{jj'}\}_{n \times n} \) can be generated to measure the similarity degree of the subject words. If we use co-occurrence times to measure the similarity analysis of subject words, we may overestimate or underestimate the similarity. In Figure 1, the area \( S_i, S_2, S_3 \) and \( S_4 \) of the figure are respectively the times of occurrence of the subject words \( T_i, T_2, T_3 \) and \( T_4 \) while the elliptical area \( \omega_{il} \) and \( \omega_{4i} \) are the times of occurrence of the subject words \( T_i, T_2, T_3 \) and \( T_4 \). When the area of \( \omega_{12} \) and \( \omega_{34} \) is the same, the similarity of the two groups of subject words is the same. But \( T_1 \) and \( T_2 \) are less than \( T_3 \) and \( T_4 \). Obviously, \( T_1 \) and \( T_2 \) are more similar than \( T_1 \) and \( T_2 \). Therefore, it is more appropriate to measure the similarity of subject word pairs by the proportion of subject word pairs in the total number of occurrences.

![Figure 1. Evaluate the shortcomings of subject word pair similarity by co-occurrence times.](image)
After the above analysis, the correlation coefficient matrix can be generated to measure the similarity degree of the subject words.

\[
\pi_{jj'} = \begin{pmatrix}
T_1 & T_2 & \cdots & T_m \\
\pi_{12} & 1 & \cdots & \pi_{1m} \\
\vdots & \vdots & \cdots & \vdots \\
\pi_{nm} & \pi_{n2} & \cdots & 1
\end{pmatrix}
\]  

The co-occurrence analysis was carried out with Bib excel software.

4.4. cluster analysis

Classify the data collected on the basis of the above co-occurrence analysis. Generally speaking, the correlation coefficient refers to the correlation coefficient between variables. As a characterization of the correlation between data objects, it can also be similarly defined, that is, the correlation coefficient between the \(i\) data object \(x_i\) and the \(j\) data object \(x_j\) is defined as

\[
\rho_{ij} = \frac{\sum_{k=1}^{p} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{p} (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^{p} (x_{jk} - \bar{x}_j)^2}} \quad (-1 \leq \rho_{ij} \leq 1)
\]

Among them, \(\bar{x}_i = \frac{1}{p} \sum_{k=1}^{p} x_{ik}\); \(\bar{x}_j = \frac{1}{p} \sum_{k=1}^{p} x_{jk}\).

The correlation coefficients of all two data objects can be calculated and arranged into a correlation coefficient matrix, i.e.

\[
R = (\rho_{ij}) = \begin{pmatrix}
\rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\
\rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\
\vdots & \vdots & \cdots & \vdots \\
\rho_{n1} & \rho_{n2} & \cdots & \rho_{nn}
\end{pmatrix}
\]

Among them, \(\rho_{11} = \rho_{22} = \cdots = \rho_{nn} = 1\) can cluster \(n\) data objects according to \(R\).

5. Conclusion

The processed technical data is presented to users through the visualization technology platform, so as to extract key information from complex data in a short time and serve for decision support. Obtaining knowledge and important information from a large number of original data of scientific and technological activities is the purpose of scientific and technological tracking. The most important task of national defence science and technology management is to analyze and evaluate the situation and trend of science and technology development, to carry out the selection of priority areas of science and technology, and to reasonably allocate science and technology resources. The data analysis strategy proposed in this paper can provide support for the demonstration of national defence pre research project.

Acknowledgments

Thank you for the support of "research and development fund" of Dalian naval academy! Thanks for the guidance of Professor Wang Yitao and Xu Xiaogang of Dalian naval academy!

References

[1] Zhang,F.F.,Zhou,N.,Yu,X.S.,Chen,H.Q.(2007) On the visualization of knowledge domain. Forum of digital library, 5:15-17.

[2] Song,S.(2006) Research on the development of management science based on scientific measurement. Nanjing University of Technology Press, Nanjing.

[3] Kang,Y.H.(2007) A research on technology tracking method based on co-occurrence analysis. Dalian University of Technology Press, Dalian.
[4] Zhao, F., Leng, F. H. (2007) On the theory of strategic information research in the practice of information service. Journal of information, 5:27-31.