Scientific Topic Evolution Pattern Recognition and Analysis
—— A Case Study of Medical Topics

Keywords: Medical Topics；Clustering Analysis；Topic Similarity；Topic Evolution Pattern

Nowadays, researchers can easily obtain literatures in a certain professional field. Under such a background of massive data, the topic evolution model can help researchers quickly sort out the development of technical topics in this field. In previous studies, scholars have divided the thematic relations in the adjacent sub-periods into five types , including emergence、division、integration、inheritance and extinction(Zhang et al., 2017). On this basis, the analysis of the relationship between the maturity or attention of the topic and the evolution patterns can help researchers discover the potential law of technology development and make more accurate predictions. Consider possible relationships between topics at the semantic level, this paper proposes a topic evolution pattern analysis method based on the improved similarity calculation model. This paper is organized as follows:

The first step, we extract the MeSH terms of the literatures in the field from Pubmed database and obtain the document-topic frequency matrix（Danielle et al., 2013）.

The second step, we divide the time period according to the amount of literature published. And multidimensional evaluation is carried out for each time stage to make the analysis of evolution more accurate and specific. The k-means algorithm is used to cluster the documents of each time period, so as to obtain a set of documents of a group of terms, namely the topics. Then evolution ability of each topic was evaluated in each time period. Density and heat index are used and mapped to two-dimensional coordinates, and the index of density represents the degree of relevance of the topic words within the topic set, while the index of heat represents the number of document within the topic(Qi et al., 2016). In this way, the maturity and attention of the topic in each period can be identified to facilitate the subsequent evolution analysis.

The third step is to identify the topic association relation of adjacent time periods, so as to construct the topic evolution diagram. Because the traditional topic similarity calculation does not take into account the semantic similarity between terms within the topic, the similarity between topics in different time periods is very small. In this paper, the cosine similarity formula is improved by using the semantic similarity value between terms, so as to obtain more accurate topic correlation between adjacent sub-periods. The improved formula is shown in (1).

$$\mathrm{Similarity}(x,y)=\frac{\sum_{i=1}^{n}\frac{x_i y_i}{\delta_i}}{\sqrt{\sum_{i=1}^{n}{x_i}^2}\sqrt{\sum_{i=1}^{n}{y_i}^2}} \qquad (1)$$

Where x、y indicates two topics, x_i、$y_i$respectively represent the occurrence frequency of the

$\mathrm{i^{th}}$ term of the two topics, δ_i indicates the average similarity value between the $\mathrm{i^{th}}$ term and other terms. The bigger the value of δ_i, the closer the relationship between the term and other terms, the less important the difference of x_i and y_i values is. In this paper, the improved Rodriguez semantic similarity model(Rodriguez et al., 2003) is used to calculate the similarity value between MeSH terms (formula (2)), so the calculation formula of δ_i is shown in formula (3).

$$S_{semantic}(i^p, j^q) = w_w S_w(i^p, j^q) + w_u S_u(i^p, j^q) + w_n S_n(i^p, j^q) \tag{2}$$

$$\delta_i = \frac{\sum_{j=1}^{i-1} S_{semantic}(i^p, j^q) + \sum_{j=i+1}^{n} S_{semantic}(i^p, j^q)}{n-1} \tag{3}$$

Where i、j represent two terms, S_w、S_u、S_n respectively represent the similarity of set of entry words, annotation similarity and semantic distance of i and j.

According to the similarity value calculated by formula (1), the threshold value is set reasonably to filter the correlation relationship, and the topics of the adjacent sub-periods are correlated to get the topic evolution map. On this basis, we combine the results of the topic evolutionary capability evaluation to identify and analyze the evolution regular pattern of the four topic evolution modes: emergence, division, integration and inheritance, so as to discover the possible relationship between topic maturity or attention and evolution patterns. The research results can provide a reference for grasping the regular pattern of technology development and forecasting. This paper takes Alzheimer disease as the object of empirical study to verify the effectiveness of the evolutionary model.

The technological framework is shown in Figure 1.

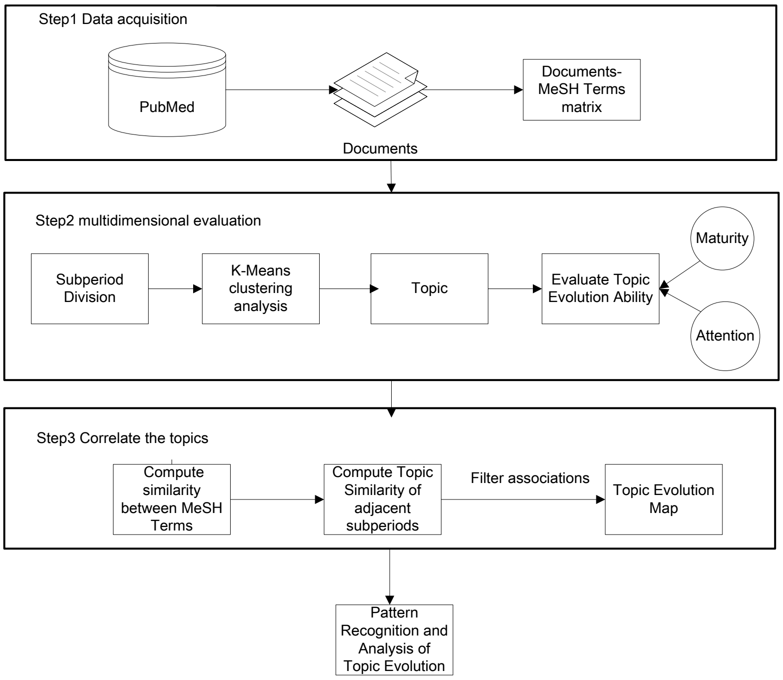
Step1 Data acquisition
PubMed
Documents
Documents-MeSH Terms matrix
Step2 multidimensional evaluation
Subperiod Division
K-Means clustering analysis
Topic
Evaluate Topic Evolution Ability
Maturity
Attention
Step3 Correlate the topics
Compute similarity between MeSH Terms
Compute Topic Similarity of adjacent subperiods
Filter associations
Topic Evolution Map
Pattern Recognition and Analysis of Topic Evolution

Figure 1 Framework for identifying and analysing topic evolution pattern