Classifying Biomedical Text for Mining Keyword Correlations and Technology Opportunities Analysis

Jing Ma[*1], Donghua Zhu[1], Alan L. Porter[2, 3] and Ying Huang[1]

[*]*majing881003@163.com*
[1.] School of Management and Economics, Beijing Institute of Technology, Beijing, China
[2.] School of Public Policy, Georgia Institute of Technology, Atlanta, GA, USA
[3.] Search Technology, Inc., Atlanta, GA, USA

Seeking opportunities for future research and innovation makes great sense for emerging technologies. Technology Opportunities Analysis (TOA) attempts to capture technological dynamics using tech mining to develop Competitive Technical Intelligence (CTI) and grasp rising hotspots in a specific domain. It facilitates researchers with supportive intelligence, not only to keep up with technology developmental trends, but to get inspired from adjacent fields for further R&D.

Biomedical research requires strict procedures from formulation development to clinical trial and only few studies end up with market products. For different formulas and biomedical materials, they may undergo different development phases and prospects. It has drawn much attention from scientists and technology administrators to reveal and study this translational process to reduce R&D cost and to improve research efficiency. Considering the abundant literature resources on biomedical research, it is necessary and beneficial to trace such translational progress by carrying out TOA regarding different research stages in biomedical field so as to grasp more detailed insights.

The ideal profile of TOA requires the ability of providing technology insights with less reliance on experts' contributions, but more efficiency. In this study, we start from biomedical literature. First, we demonstrate how to classify biomedical text based on characteristic descriptions regarding research stages of reported publications using multi algorithms. On one side, we apply natural language process (NLP) and essential cleaning steps to generate keywords. Based on these keywords we are able to introduce topic model and to get a distribution on how these articles are related to different topics. At the same time, a keyword list from domain experts are also introduced as properties to cluster these articles. In fact, it is impossible to classify each article to a specific research stage, since many studies are not always forward, and they may have feedback. And then based on different subsets of classified data and clustered topics, we apply multiple keyword correlations and tech mining approaches to 1) generate technology framework and sub-sectors and 2) identify potential opportunities and directions by analyzing the networking of different research stages.

We illustrate this research framework in how gold nanoparticles (GNPs) have been applied to biomedical domain in different aspects. By comparing developmental pathways of various elements, disease types, and relative nano-materials, similarities and differences are identified. We further discuss which topics are further developed or more close to commercialization and at the same time which ones are more emerging and with better prospects so as to illuminate how such correlational analysis can generate leading indicators for the future.

This study provides a quantitative aspect for researchers to monitor translational process in biomedical domain. Since this study is on preliminary stage, some results still need to be refined. To validate the result of this study, we will discuss the result with domain experts to ask for their comments for further improvement.