Technology mining for emerging S&T trends and developments: dynamic term clustering and semantic analysis

Pavel Bakhtin* and Ozcan Saritas*

*{pbakhtin, osaritas}@hse.ru
Institute for Statistical Studies and Economics of Knowledge,
National Research University, Higher School of Economics (Russian Federation)

Introduction

In the world of rapidly developing Science and Technology (S&T), with increasing volumes of S&T-related data and greater interdisciplinary and collaborative research, technology mining (TM) helps to acquire intelligence about emerging trends and future S&T developments. The task is becoming crucial not only for high-tech startups and large organizations, but also for venture capitalists and other companies, which make decisions about S&T investments. Governments and Public Research Institutions are also among the main stakeholders and potential users of TM to set up R&D priorities, plans and programs according to the current and future state of S&T development.

Existing methods and tools in the field of TM provide opportunities for the collection and extraction of metadata (e.g. list of inventors, assignees, technology domains or research areas, publication dates, keywords etc.) from patents, publications and other S&T sources, as well as visual representation of the data gathered, further term cluster analysis and a wide variety of other outputs depending on the purpose of the activity (Porter & Cunningham, 2004; Yoon B., 2008). At the same time developments in the area of natural-language processing (NLP) made the extraction of text structure, dependencies between tokens and identification of part of speech, dependency types and other relevant linguistic information (Manning, et al., 2014) more sophisticated and applicable to various fields, including TM. NLP is exploited to process patent abstracts and other textual data to extract technology properties and functions (Yoon J. & Kim, 2012) and even predict possible future trends based on TRIZ[1] classification (Park, et al., 2013).

Term clusters built by TM and bibliometric tools based on co-occurrence of authors' keywords or terms processed from titles and abstracts of scientific documents combine totally different types of objects: research fields, major problems and challenges, methods, inventions, products, technologies and etc. Specific expertise in the field may allow a researcher to identify key objects of the study. However, objects themselves and their frequency dynamics over the time period alone do not fully indicate S&T developments and emerging trends in the area.

In order to improve the process of the identification of emerging S&T trends and developments, the paper focuses on dynamic term clustering and suggests a systemic approach to combine TM, bibliometrics, NLP and semantic analysis as part of the unified analytical framework. The approach proposed utilizes existing clustering methods and tools along with the analysis of term linguistic dependencies (Marnefe & Manning, 2008) in order to study changes of objects over the time along with their semantic meanings.

Findings

Proposed systemic approach can be described as the process with several steps.

First, the scope of the S&T trends and developments research is identified in terms of keywords, patent technology domains, international patent classifications (IPCs), publication research areas and categories. Then, big data samples sorted by the most highly cited publications and patents for each year of the

[1] TRIZ ("theory of inventive problem solving") – formal forecasting approach to identification of possible development paths of inventions developed by Genrich Altschuller

researched time period are processed with high performance TM and bliometric tools like VantagePoint and VOSviewer. As a result, terms clusters are identified for each year of the time period.

Second, clusters from consecutive years are cross-analyzed in order to identify the most relevant terms undergoing S&T development. Such terms along with their relationships are grouped into dynamic term clusters. Publications, patents and other S&T documents that contain relevant information to these clusters are processed by Stanford CoreNLP toolkit. Extracted sentence structures along with term linguistic dependencies (adjective modifiers with nouns, nouns with verbs and etc) are analyzed by the software developed by the authors of the paper identifying main subjects, objects and their properties and functions.

Then the terms from dynamic term clusters are mapped as subjects and objects of various relationships along with properties and functions. Such semantic relationships are visualized as network chains that can be assessed by an expert to validate and formulate either S&T trends or developments in the area.

Such utilization of various methods and tools makes it possible to integrate quick big data processing tools with less accuracy (based on co-occurrence of terms) but effective enough to later form dynamic term clusters and more sophisticated tools that process small volumes of data and produce semantic context. Hence, that leads to overall optimization of the identification process of emerging S&T trends and developments.

The overall work will aim to provide a detailed description of proposed approach and software developed along with case examples that demonstrate the application of the methodology on the collection of publication and patent documents, resultant analytical outputs and visualization and interpretation of results.

References

Manning, C. et al. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (pp. 55-60).

Marnefe, M., Manning, C. (2008). *Stanford typed dependencies manual.* Stanford University.

Park, H. et al. (2013). Identification of promising patents for technology transfers using TRIZ. *Expert Systems with Applications 40*, 736–743.

Porter, A., Cunningham, S. (2004). *Tech Mining: Exploiting New Technologies for Competitive Advantage.* John Wiley & Sons, Inc.

Yoon, B. (2008). On the development of a technology intelligence tool for identifying technology opportunity. *Expert Systems with Applications 35*, 124–135.

Yoon, J., Kim, K. (2012). TrendPerceptor: A property–function based technology intelligence system for identifying technology trends from patents. *Expert Systems with Applications 39*, 2927–2938.