Identifying the latent technology opportunities based on a perspective of coupling publications with patents

Munan LI 1,2     Wenshu Wang 1     Keyu Zhou 1

1.  School of Business Administration, South China University of Technology

1.  Guangdong Province Key Laboratory on Decision & Innovation Management

2019/10/2
Within a specific topic, the publications and patents are the outputs of scientific research and technology transformation respectively, and also they are often utilized to evaluate the levels of R&D.
Under the traditional framework on identifying the TO (technology opportunity), the patents are the usually data source. And, the classic and emerging analytical methods on patents are the mainstream of current studies.
• **Research questions**

• **RQ1:** Within some topics or under the specific situations, the coupling relations between publications and patents can be conducted to identify the latent technology opportunities?

• **RQ2:** And how can we use these coupling relations between publications and patents to identify TO?

• **RQ3:** The new TO based on the proposed coupling methods can be identified by the other methods or can be compared with the other methods including traditional or emerging?
• The driving force of technology progress has been generally discussed since 1950s, and it should be endogenous from the view of productivity development (Solow, 1957). In consequence, to forecast the evolution of technology, or technology foresight has become one of critical approaches for strategic planning and industrial-policy making (Martin, 1995; Hasan and Tucci, 2010; Miles, 2010; Li et al., 2018).

• In general, the data of publications is often utilized to evaluate the scientific output and academic performance, efficiency or collaboration between different institutes or scientists, and so forth (Guenter et al., 2007; Kyle et al., 2015).

• Some latest studies have integrated social media data, e.g. Twitter, Facebook or WeChat into the framework of technological forecasting and TOA (Chen, 2018).
• Regarding the relevant studies on TOA (technology opportunity analysis), technology forecasting, emerging technologies, TCE(technological capability evaluation) and tech mining etc., the patent data play the key role in those processing (Breitzman & Mogee, 2002; Tseng et al., 2007; Youtie et al, 2008; Lee et al, 2009; Porter & Newman, 2011; Zhang et al, 2016; Wang et al., 2017).

• Also, the technique on combing bibliometrics with patent analysis has been conducted to forecast those emerging technologies or R&D evaluation (Hullmann and Meyer, 2003; Daim et al., 2006).
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientific evaluation on output, efficiency, performance and international collaborations and so forth.</td>
<td>TOA (Technology Opportunity Analysis), TCE (Technology Capability Evaluation), Technology Evolution Analysis and so on.</td>
<td>Forecasting emerging technologies, Evaluating the performance or efficiency of R&amp;D</td>
<td>Social medias data (Twitter, Facebook, WeChat) Internet data (homepages, search results and so on)</td>
</tr>
</tbody>
</table>

Basically, the general computing-framework based on the coupling relations between publications and patents is still insufficient, especially for those multi-disciplinary or trans-disciplinary topics, for example, artificial intelligence, smart cities and Industrial Internet etc.
Methodology

Here, an hypothesized framework is proposed to forecast the latent technology opportunities based on the coupling relations between scientific articles and corresponding patents within a specific topic.

<table>
<thead>
<tr>
<th>Publications</th>
<th>Patents</th>
<th>Approximate classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>More</td>
<td>More</td>
<td>High R(Research) &amp; High D(Development)</td>
</tr>
<tr>
<td>More</td>
<td>Less</td>
<td>High R(Research) &amp; Low D(Development)</td>
</tr>
<tr>
<td>Less</td>
<td>More</td>
<td>Low R(Research) &amp; High D(Development)</td>
</tr>
<tr>
<td>Less</td>
<td>Less</td>
<td>Low R(Research) &amp; Low D(Development)</td>
</tr>
</tbody>
</table>
Corresponding concepts and equations for computing:

**Perplexity:** *In natural language processing, it is used to measure the quality of the trained language model and evaluate the generalization ability of the model (Blei et al., 2003).*

\[
\text{Perplexity} = \exp \left\{ - \left( \sum_{m=1}^{M} \sum_{n=1}^{N_m} \log \left( \sum_{k=1}^{K} p(\omega_n | z_k) p(z_k | d_m) \right) \right) \right/ \left( \sum_{m=1}^{M} N_m \right) \right\}
\]

**Coherence:** *The coherence of topics can be measured by calculating the degree of correlation between the features with higher scores in the topics, which is helpful to classify the topics into understandable categories (Newman et al., 2010; Stevens, 2012).*

\[
\text{Coherence} = \sum_{i<j} \text{score}(\omega_i, \omega_j, \epsilon)
\]
For a specific topic, the coupling relationships can be described as such two aspects:

- For a specific term, the amount of publications and the amount of patents can present some useful information.
- Based on the topic modeling, some sub-topics can be extracted from the retrieved publications or patents, and the similarities between these sub-topics can be computed and conducted to evaluate the strength of coupling relations.
Methodology

**SDS (Similarity between Different Sub-topics):** the Euclidean Distance between two sub-topics, which also needs some techniques for computation.

\[ d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \]

Two new concepts are proposed:

- **TI (Transformation Intensity):** we just utilized two factors: (1) the probability of a term belonging to a specific sub-topic in publications; (2) this term has been transformed into how many sub-topics of the corresponding patents.

- **TE (Transforming Efficiency):** here, the concept of HLP (half-life period) is conducted to calculate the TE.

\[
\rho_{sz, j} = \sum_{k=1}^{n} p_{\omega_k} \cdot m_k
\]

\[
= \sum_{r=1}^{a} p_{\omega_r} \cdot m_a + \sum_{s=1}^{b} p_{\omega_s} \cdot m_b \cdot \left(\frac{1}{2}\right)^1 + \sum_{t=1}^{c} p_{\omega_t} \cdot m_c \cdot \left(\frac{1}{2}\right)^2
\]
March, 2016, the milestone event about Go game, **AlphaGo vs Sedol Lee** (4:1)

A mazing hand in the center from Sedol Lee
In 1956, the Dartmouth conference is thought as the mark event about the origins of artificial intelligence.

In August 1956, at Dartmouth college, John McCarthy, Marvin Minsky, Claude Shannon, Alan Newell, Herbert Simon and other scientists are together, talking about a complete out-of-touch theme: Using machines to mimic human learning and other aspects of intelligence.

The meeting lasted for two months, and although there was no general consensus, a name was given to the discussion: artificial intelligence. Thus, 1956 became the first year of artificial intelligence.
In the 1960s, artificial intelligence evolved different theories and branches such as computational complexity, cognitive computing and computational linguistics. With the promotion and application of relevant models and algorithms of artificial intelligence, research related to human labor union intelligence has spread to different fields and disciplines.

Considering that there are many researches on artificial intelligence and the scope of knowledge diffusion is wide, that is, it is more or less involved in most research fields, the author considers topic modeling to conduct more in-depth content mining and presentation. LDA（Latent Dirichlet Allocation）LDA is a popular and mature topic mining algorithm at present. In essence, LDA is a Bayesian model including subject, document and topic, which is completely based on bayesian reasoning mechanism and has good knowledge interpretation ability (Blei 2003; 2012).
Case study: Artificial intelligence

1. Collect the related publications to AI
2. Select the data columns as the source for topic mining
3. Topic modeling in each phases (python3.7)
4. Calculate the distance between topics
5. Evaluate the branched topics
6. Output report

Return to Step #2?
Case study: Artificial intelligence
Patent applications surged 7,351 in 2017, of which 5,655 were from China, accounting for 76.93 percent of the total growth in 2017.
## Case study: Artificial intelligence

<table>
<thead>
<tr>
<th>Topic code</th>
<th>Publications</th>
<th>Topic tag</th>
<th>Topic code</th>
<th>Publications</th>
<th>Topic tag</th>
<th>Topic code</th>
<th>Publications</th>
<th>Topic tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_{1-6}</td>
<td>20957</td>
<td>Classification and classifier</td>
<td>L_{2-13}</td>
<td>57945</td>
<td>Artificial neural network prediction model</td>
<td>L_{3-7}</td>
<td>79750</td>
<td>Fuzzy model and fuzzy system</td>
</tr>
<tr>
<td>L_{1-7}</td>
<td>19345</td>
<td>Associative memory and pattern recognition</td>
<td>L_{2-12}</td>
<td>51028</td>
<td>Dynamic neural network stability</td>
<td>L_{3-18}</td>
<td>67900</td>
<td>Neural network model</td>
</tr>
<tr>
<td>L_{1-10}</td>
<td>13698</td>
<td>Nonlinear dynamic simulation</td>
<td>L_{2-8}</td>
<td>48661</td>
<td>Optimization design of genetic algorithm</td>
<td>L_{3-5}</td>
<td>64369</td>
<td>Parametric prediction model</td>
</tr>
<tr>
<td>L_{1-14}</td>
<td>13200</td>
<td>Clinical medicine</td>
<td>L_{2-2}</td>
<td>48181</td>
<td>Protein structure sequence prediction</td>
<td>L_{3-10}</td>
<td>63070</td>
<td>Support vector machine classification</td>
</tr>
<tr>
<td>L_{1-9}</td>
<td>12593</td>
<td>Adaptive fuzzy control</td>
<td>L_{2-3}</td>
<td>45952</td>
<td>Spectral analysis of cancer</td>
<td>L_{3-15}</td>
<td>60883</td>
<td>Optimization design of genetic algorithm</td>
</tr>
</tbody>
</table>
### Case study: Artificial intelligence

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic tag</strong></td>
<td><strong>Patents</strong></td>
<td><strong>Topic code</strong></td>
<td><strong>Patents</strong></td>
</tr>
<tr>
<td>P_1-4</td>
<td>2001</td>
<td>Neural network</td>
<td>P_2-5</td>
</tr>
<tr>
<td>P_1-1</td>
<td>1731</td>
<td>Image processing</td>
<td>P_2-3</td>
</tr>
<tr>
<td>P_1-6</td>
<td>1686</td>
<td>Signal sensor</td>
<td>P_2-6</td>
</tr>
<tr>
<td>P_1-5</td>
<td>1369</td>
<td>Computer system</td>
<td>P_2-4</td>
</tr>
<tr>
<td>P_1-2</td>
<td>1318</td>
<td>System and parameters of neural network</td>
<td>P_2-1</td>
</tr>
</tbody>
</table>
Case study: Artificial intelligence

The sub-topics evolution of publications on artificial intelligence.

The thickness of the line between two topics is expressed as follows: the smaller the distance (the smaller the difference) is, the thicker the line, indicating the closeness of the relationship and the degree of similarity.
Case study: Artificial intelligence

The sub-topics evolution of patents on artificial intelligence
Case study: Artificial intelligence

The TI (Transformation Intensity) on those sub-topics on artificial intelligence
Case study: Artificial intelligence

The TE (Transformation Efficiency) on those sub-topics on artificial intelligence.
A framework was proposed to forecast the latent TO (technology opportunity) based on the coupling relations between publications and patents, and the case study on artificial intelligence was conducted.

In case study, this paper partially addresses some interesting phenomenon on artificial intelligence.

However, the generalization capability of this proposed framework still needs more cases and verifications in the following studies.
Thank you!