Exploration of a Science-technology Relationship Index and its Measurement Algorithm

Yan Qi qi.yan@imicams.ac.cn
Institute of Medical Information / Medical Library. CAMS & PUMC
Zhengyin Hu huzy@clas.ac.cn
Chengdu Documentation and Information Center, Chinese Academy of Sciences
Ziqiang Liu liuziqiang@mail.las.ac.cn
University of Chinese Academy of Sciences
Haiyun Xu xuhy@clas.ac.cn
Chengdu Documentation and Information Center, Chinese Academy of Sciences
Ran Zhang zhang.ran@imicams.ac.cn
Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences/Peking Union Medical College

Introduction

As early as the 1960s, Professor Price(Price,1965) pointed out that knowledge may flow from science to technology or from technology to science though they have unique knowledge accumulation structure. Later, many scholars(Verbeek,et al,2003; Garfield,1984; Narin, et al 1997; Meyer,2000) conducted much in-depth research and argumentation on the "Science-technology Relationship" from different perspectives. There are many ways to uncover science-technical knowledge linkages, for example the cooperative R&D activities of public scientific research institutes and enterprises (Garg,2001), their geographic association(Autant,2001),and the talents flow between science and technology departments. The most common method is the co-occurrence analysis or citation analysis between the academic papers and patents which are regarded as the manifestations of scientific research achievements and technological innovation results respectively. Based on the number or content characteristics of the papers and patents, this method can reveal the linkage between science and technology quantitatively and microscopically.

Narin et al.(Carpenter& Narin,1983; Narin& Noma,1985)pioneered the use of patent's essay citations to measure the interaction between science and technology. In 2000, the CHI Corporation established the scientific linkage index of technology(Science Linkage, SL), and then, many other index including non-patent citations per patent, the time lag and national distribution of the patents' paper citations are put forward.The citation of patents in academic papers also implies the correlation between basic research and technological innovation. Therefore, correspondingly, technology linkage (TL), time lag and national distribution of the paper s' patent citations are also applied by many scholars(Glänzel,2003; Huang, et al,2015). From the meaning and measurement methods of the indicators, it can be seen that existing studies (such as SL and TL indicators) mostly focus on one-way associations and base on the quantitative characteristics of the citation relationships, which is insufficient to reflect the bidirectional and content relationship between science and technology. This

paper proposes a new integrative index: Science-Technology Linkage (STL), which is based on the research topic analysis of thesis and patent and hence it can reflect the bidirectional and content relationship to some extent.

Index and Methodology

Index : STL

The new indicator is based on the topic analysis of the thesis and patent to measure the degree of association between science and technology represented by the corresponding collection of documents. The relationship between some topics of the two sets can partly represent the relationship between the two sets, and the relationship of all the topics of the two sets can represent the whole relationship. There can be many measurement ideas，and this article just focuses on the simplest idea——founding out the common topics between thesis set and patent set and then calculating their proportion with the number of all the topics of two sets, as follows:

$$\mathrm{STL} = \frac{\mathrm{Nct(patent + paper)}}{\mathrm{Nt(patent) + Nt(paper)}} \qquad (1)$$

Nct(patent+paper) indicates the number of common topics, Nt(patent) indicates the number of research topics of patent set, and Nt(paper) indicates the number of research topics of paper set.

Methodology: Generating research topics

After collecting scientific papers and patents, some national language processing (NLP) tools are used to extract keywords from the text fields, such as “Title” and “Abstract”, which are precise and meaningful for NLP. The input of topic model (e.g., LDA or PLDA) is a list of bag-of-words. Each document is represented as an exchangeable bag-of-words. The quality of these bag-of-words is very important to the result of topic model, and an inductive framework called “term clumping” is used to clean the bag-of-words (Yi, Alan & Zhengyin et al., 2014). Then LDA topic model is used to separately generate the research topics based on bag-of-words of scientific papers and patent documents. Each paper and patent document is represented as some topics with probability weight, and each topic as some keywords with probability weight (Blei, Ng. & Jordan. 2003). In 2009, Wang et al. proposed the PLDA model, which can effectively improve the analysis efficiency and precision of the traditional LDA model (Wang, et al,2009). We simultaneously used LDA and PLDA to generate themes for two collections at specific period.

Methodology: Mining common research topics

Common research topics means those simultaneously appear in scientific papers and patent documents with high similarities. According to the output of LDA, the research topics can be represented as algorithm (2), and the similarities sim(topic_i, topic_j) of topic_i and topic_j can be calculated by algorithm (3).

$$topic_i = \sum_{k=1}^{l} term_k \cdot p(term_k \mid topic_i) \quad (2)$$

$p(term_k \mid topic_i)$: weight in probability distribution of $term_k$ in $topic_i$

$$\text{sim}(\text{topic}_i,\text{topic}_j)=\sum_{r=1}^{n}\sum_{k=1}^{m}\frac{p_{ir}\cdot \text{sim}(term_{ir},term_{jk})}{m\cdot n} \quad (3)$$

n: number of terms in $topic_i$; m: number of terms in $topic_j$

We use cosine similarity analysis to calculate the similarities sim($term_i$, $term_j$) based on the co-occurrence matrix of terms in documents set. The topics from scientific papers and patent documents of which similarities are higher than a given threshold are regarded as common topic.

Case Study

Hepatitis C virus (HCV) research field was selected as a case study. We selected the database of WOS and DII as data sources and obtained 33524 papers and 6804 patents from 2008 to 2017, which are divided into five groups biennially. Following the methodology mentioned above, the topics of papers and patents per group are extracted by PLDA programming and then the value of STL is calculated out. All the relative data are stated in Table 1.

Tab.1. Number of topics and value of STL.

Group	Time Span	Science		Technology		STL
		Paper No.	Topics No.	Patent No.	Topics No.	
1	2008~2009	5734	65	1488	104	0.23
2	2010~2011	6388	72	1393	78	0.13
3	2012~2013	7074	77	1406	1228	0.20
4	2014~2015	7343	59	1542	89	0.24
5	2016~2017	6985	53	975	104	0.17

Some common research topics are listed in table2.

Tab.2. Common Research Topics of HCV (partial).

Time Span	Research Topic	similarity
2008~2009	polymerase inhibitor	0.7
	immunodeficiency	0.7
2010~2011	crystal structural elements	0.75
	supercharged protein for cell penetration	0.75
2012~2013	amplifying target nucleic acid molecule	0.8
	preventing or treating of fibrotic liver disease	0.8
2014~2015	virus replication and autophagy	0.8
	new or substituted compounds or derivatives for preventing, inhibiting or treating HC	0.8
2016~2017	extrahepatic manifestations of Hepatitis C	0.85
	targeted therapy	0.85

Conclusions

This paper puts forward the thought of measuring science-technology association based on the relationship analysis of themes that extracted from the papers and patents collections. The empirical test demonstrates the feasibility of this approach, as well as the possible advantages compared with the existing research, which can achieve both the quantitative measurement of the correlation tightness and the qualitative analysis of related topics. However, each step of the measurement process, such as domain search, text cleaning, calculation of similarity, and selection of thresholds, will affect the correctness and objectivity of the final result. This article designed a simplest formula, so it can only be used as a preliminary exploration. In the future, we need to think more deeply about each step, especially on the evaluation of similar topics. In addition, the definition of similar topics (examples in tab.2) also requires the assistance of domain experts.

Acknowledgments

The work described in this paper was supported by the Basic Scientific Research Project of Chinese Academy of Medical Sciences “Exploration of Measurement Method of Science-technology Correlation Degree based on Topic Analysis of Thesis and Patent Sets” (Grant no. 2017PT63008) and 2017 “Peking Union Youth Research Fund” project of Chinese Academy of Medical Sciences “Exploration of Selection Method of Collaborative Innovation Partners based on Innovation Chain Theory” (Grant no. 2017330008).

References

De Solla Price D J. Is technology historically independent of science? A study in statistical historiography[J]. Technology and Culture,1965,6(4):553-568.

Verbeek A, Debackere K, Luwel M. Science cited in patents: A geographic “flow” analysis of bibliographic citation patterns in patents[J]. cientometrics,2003,58(2):241-263.

Garfield E. Patent citation indexing and the notions of novelty, similarity and relevance[J]. Essays of An Information Scientist,1984,7(3):536-542.

Narin F, Hamilton K S, Olivastro D. The increasing linkage between US technology and science[J]. Research Policy,1997,26(3):317-330.

Meyer M. Does science push technology? Patents citing scientific literature[J]. Research Policy,2000,29(3):409-434.

Garg K C. A study of collaboration in laser science and technology[J].Scientometrics,2001, 51(2):415-427.

Autant B C. Science and knowledge flows: Evidence from the French case[J].Research Policy,2001,30(7):1069-1078.

Carpenter M P, Narin F. Validation study: patent citations as indicators of science and foreign dependence[J].World Patent Information,1983,5(3):180-185.

Narin F, Noma E. Is technology becoming science?[J]. Scientometrics,1985,7(3-6):369-381.

Glänzel W, Meyer M. Patents cited in the scientific literature: An exploratory study of' reverse' citation relations[J]. Scientometrics,2003,58(2):415-428.

Huang M H, Yang H W, Chen D Z. Increasing science and technology linkage in fuel cells: A cross citation analysis of papers and patents[J]. Journal of informetrics,2015,9(2):237-249.

Yi Zhang, Alan L. Porter, Zhengyin Hu, Ying Guo, & Nils C. Newman. (2014). “term clumping” for technical intelligence: a case study on dye-sensitized solar cells. Technological Forecasting & Social Change, 85, 26-39.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Wang Y, Bai H J, Stanton M, et al. PLDA: Parallel Latent Dirichlet Allocation for Large-scale Applications[C]//AAIM,2009:301-314