Identifying Research Fronts Based on Scientific Papers and Patents using Topic Model: Case Study on Regenerative Medicine

Hu Zhengyin[1*] Pang Hongshen[2] Qin Xiaochu[3] Wei Ling[1] Dong Kun[1] Xu Haiyun[1] Song Yibing[3]
[1*] *huzy@clas.ac.cn*
[1]Chengdu Documentation and Information Center, Chinese Academy of Sciences, No.16, Nan'erduan, Yihuan Road, Chengdu, 610041 (China)
[2]Shenzhen University, Nanhai Avenue 3688, Shenzhen, 518060(China)
[3]Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, 190 Kai Yuan Avenue, Science Park, Guangzhou, 510530 (China)

Introduction

Generally, scientific papers reflect the basic research achievements-, while patents reflect the application research achievements. It is a new perspective to identify research fronts by combining scientific papers and patents (Kostoff & Schaller, 2001; Xiaoyang, Yanning & Zhihui. 2016). However, it is a challenge to properly mine the "common" research fronts between scientific papers and patent documents. For example classification codes are too broad and keywords are detailed to represent the research fronts (Xiaoyang, Yanning & Zhihui. 2016). Topic model is a series of algorithms to automatically learn topics from documents set based on statistical techniques, which represents document as probability distributions over topics and topic as probability distributions over keywords, and it has been widely applied in scientific literature mining (Blei, Ng. & Jordan. 2003). This paper focuses on how to identify research fronts using topic model based on scientific papers and patent documents, which can help to accurately identify research fronts from the perspectives of science and technology simultaneously.

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) 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).

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 (1), and the similarities sim(topic$_i$, topic$_j$) of topic$_i$ and topic$_j$ can be calculated by algorithm (2).

$$\mathrm{topic}_i = \sum_{k=1}^{l} term_k \cdot p(term_k \mid topic_i) \qquad (1)$$

$p(term_k \mid topic_i)$: weight in probability distribution of term$_k$ in topic$_i$

$$\mathrm{sim}(\mathrm{topic}_i,\mathrm{topic}_j)=\sum_{r=1}^{n}\sum_{k=1}^{m}\frac{p_{ir}\cdot \mathrm{sim}(term_{ir},term_{jk})}{m\cdot n} \qquad (2)$$

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$_\mathrm{i}$, term$_\mathrm{j}$) 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 merged and chosen as candidates for research fronts.

Identifying research fronts

Two indicators of “research topic age (RTA)” and “number of research topic authors (NRTA)” are chosen to identify research fronts. RTA reflects time span of research topics, and the larger RTA value is, the wider the time span of distribution of topics. NRTA reflects academic attentiveness, and the larger NRTA value is, the hotter the topics are. Therefore, research topics with smaller RTA and larger NRTA can be considered as research fronts. RTA and NRTA are defined as follows (Xiaoyang, Yanning & Zhihui. 2016):

$$\mathrm{RTA}(\mathrm{topic}_i)=\sum_{\mathrm{i}=1}^{n}\mathrm{Y}_{\mathrm{kw}}*n_i/N \qquad (3)$$

n$_\mathrm{i}$: number of terms in topic of the time span; N: total number of terms in all topics of the time span; Y$_\mathrm{kw}$: age of term

$$\mathrm{Y}_{\mathrm{kw}}=\sum_{\mathrm{i}=1}^{n}(Year_{cur}-Year_i)*tfidf_i/(\sum_{j=1}^{n}tfidf_{\mathrm{j}}) \qquad (4)$$

Year$_\mathrm{cur}$: last year of the time span; Year$_\mathrm{i}$: year of the time span; s in all topics; tfidf$_i$: TF/IDF value of term$_i$

$$\mathrm{NRTA}(\mathrm{topic}_i)=n_i/N*100\% \qquad (5)$$

n$_\mathrm{i}$: number of authors in topic$_i$ of the time span; N: total number of authors in all topics of the time span

Case Study

Regenerative Medicine (RM) was selected as a case study. We selected the database of WOS and DII as data sources and obtained 9655 papers and 1044 patents. Following the methodology mentioned above, 68 common research topics were gotten. We set the thresholds of RTA is 3.0 and NRTA is 5%. If the RTA of a common research topic is lower than 3.0 and the NRTA is higher than 5%, it can be considered as a research front, and some of them are stated in Table 1.

Tab.1. Research Fronts of RM (partial).

Time Span	Research Topic (Label)	RTA	NRTA
2001-2005	ST2 (gene therapy)	1.84	6.14%
	ST16 (stem cell differentiation)	2.15	5.12%
	ST47 (embryonic stem cell)	2.53	8.31%
	…		
2006-2010	ST5 (induced pluripotent stem cell)	2.92	9.43%
	ST47 (embryonic stem cell)	2.76	5.82%
	ST52 (DNA analysis)	2.63	5.31%
	….		
2011-2016	ST5 (induced pluripotent stem cell)	2.74	15.42%
	ST23 (mesenchymal stem cell)	1.95	5.08%
	ST62 (gene express)	2.68	11.21%
	…		

Conclusions

The result indicates that this method can not only identify research fronts based on scientific papers or patents, but also analyses from the perspectives of science and technology simultaneously, which makes is the results more accurate. Further, it can also be applied to track the evolution trends of research fronts.

Acknowledgments

The work described in this paper was partially supported by the Guangdong Science and Technology Program “Integrated Information Service for Regenerative Medicine and Tissue Engineering” (Grant no. 2016A040403098).

References

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

Kostoff, R. N., & Schaller, R. R. (2001). Science and technology roadmaps. Engineering Management IEEE Transactions on, 48(2), 132-143.

Xiaoyang,X. Yanning Z., & Zhihui L.(2016). Study on the Method of Identifying Research Fronts Based on Scientific Papers and Patents. Library and Information Service 60(24),97-106.

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.