中国科学院
CHINESE ACADEMY OF SCIENCES

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

Hu Z.Y., Wei L., Dong K., Xu H.Y.
Chengdu Library and Information Center, Chinese Academy of Sciences (CLAS, CAS)
Pang H.S., Qin X.C., Song Y.B.
Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences (GIBH, CAS)
Atlanta, GA, USA
Oct. 09, 2017

Introduction

- Intellectual Property (**IP**) Information Center, CLAS, CAS
  - IP Information Portal: IP database, IP analysis tools, IP training, IP assessment, IP trading & transforming, etc.
  - IP Information products: IP information journal, IP analysis reports, IP consulting reports, Patents Tech Mining, etc.
  - IP Information Services: custom data, intelligence products, consulting, training services for researchers, IP managers, IP policymakers, etc.
  - Groups : 24 researchers including 3 professors, 4 associated professors and 10 Ph. D. and Master Students, etc.
- Information & Intelligence Center, GIBH, CAS
  - GIBH is a leading research institute on Stem Cell Biology and Regenerative Medicine (**RM**).
  - Specialized RM Information & Intelligence Center.
  - StemCell Cloud of CAS: more than 200 GB RM and StemCell S&T Literature and Scientific Data, Supercomputing Environment, Stem Cell Bulletin.

Some Products and Outputs

Research Background

Why Regenerative Medicine?

- RM deals with the "process of replacing, engineering or regenerating human cells, tissues or organs to restore or establish normal function" (Wikipedia). Since 1999, the research achievements of stem cells and RM have been selected 11 times world's ten largest technology breakthrough in Science magazine.
- In 2012, Prof. Yamanaka of Kyoto University and Prof. John Gordon of Cambridge University won the Nobel prize in physiology and medicine for the research about stem cells and RM.
- In 2011, Chinese Academy of Sciences launched a “Stem Cells and Regenerative Medicine Research” Strategic Leading Science & Technology Program.
- In 2016, “Organ Repair and Reconstruction” (applications of RM) was chosen as one of 60 major breakthroughs of Chinese Academy of Sciences’ 13th Five-Year Plan.

What is Research Front?

- Research front distinguishes the sciences from the rest of scholarship and characterizes the research field whose recent literature are cited so much faster by scientists. (Price. D. D., 1965)
- Research areas in science, particularly those at the cutting edge of their fields, are characterized by patterns of intense communication between scientists, including "Emerging Research Fronts ,Fast Moving Fronts". (Essential Science Indicators)
- Similar Concepts: Emerging Trend, Emerging Research Domains, Emerging Research Area, Emerging Fields, Emerging Knowledge Domains, Emerging Topics, etc.
- In this paper: new and emerging **research topic** that gets scientists' high concerns for a certain period of time in a specific domain, with following features: **novelty**, **timeliness**, **clusterization**.

Present State of Research

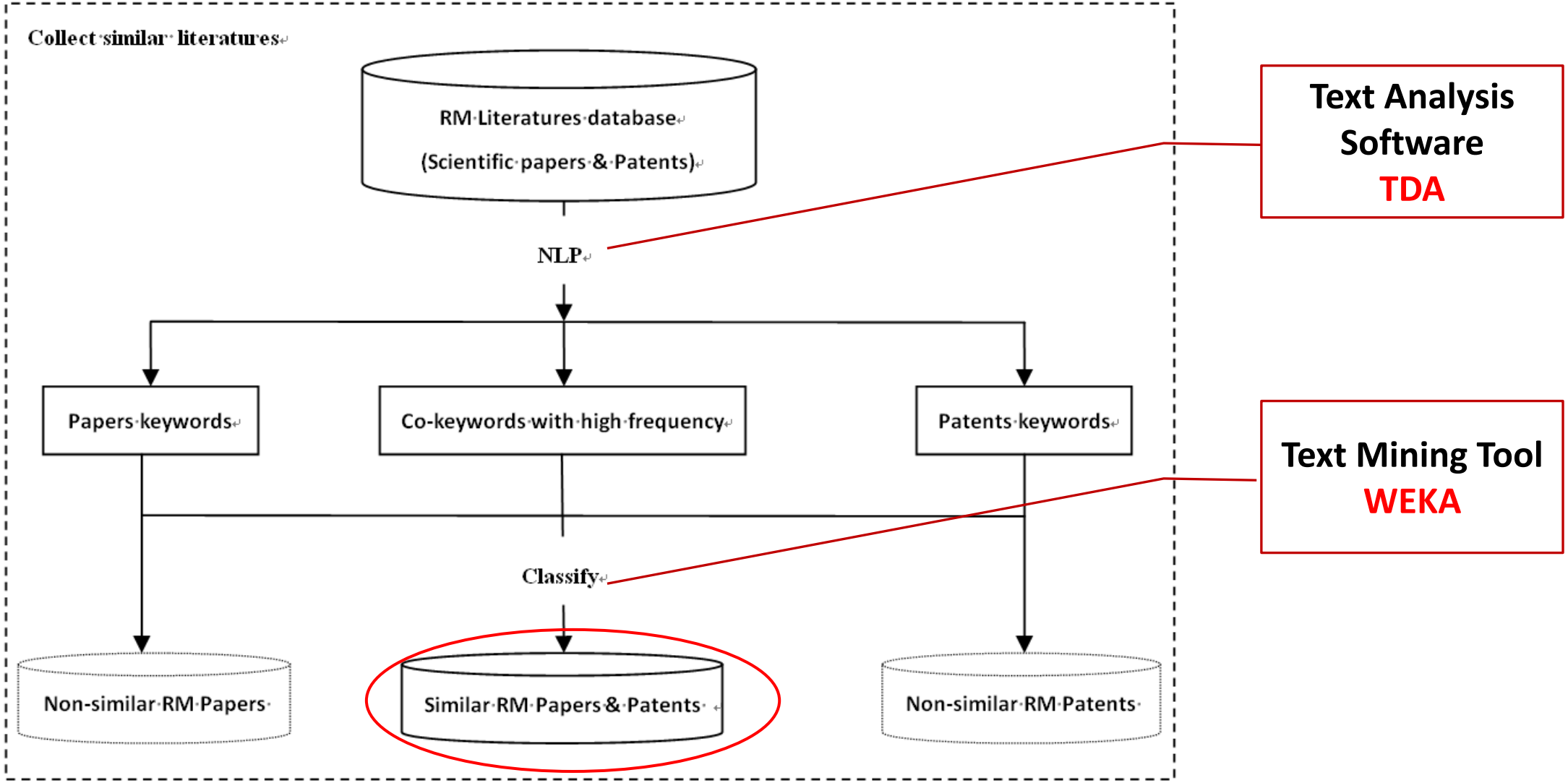
- Scientific Literatures
  - Mainly focuses on scientific papers ("Science"), rarely involves patents ("Technology").
  - ESI: Highly Cited Paper or Hot Papers
- Method
  - Citation: Co-Citation Analysis, Bibliographic Coupling, Direct Citation Link, etc.
  - Content: Words Frequency, Co-Words Analysis, etc.
  - Combined: keywords & Co-Citation Analysis, etc.
- Limitations
  - Single view and incomplete data: Generally, scientific papers generally only reflect the basic research achievements, while patents reflect the application research achievements.
  - Citation Analysis is popular and accurate, but a larger amount of literatures is needed. It usually used for a macro level research fronts analysis (ESI), not a meso-level or micro level analysis.
  - Content and text analysis usually used for meso-level or micro level analysis, but the existing researches focus on keywords analysis and text mining methods rarely involved.

Research Method

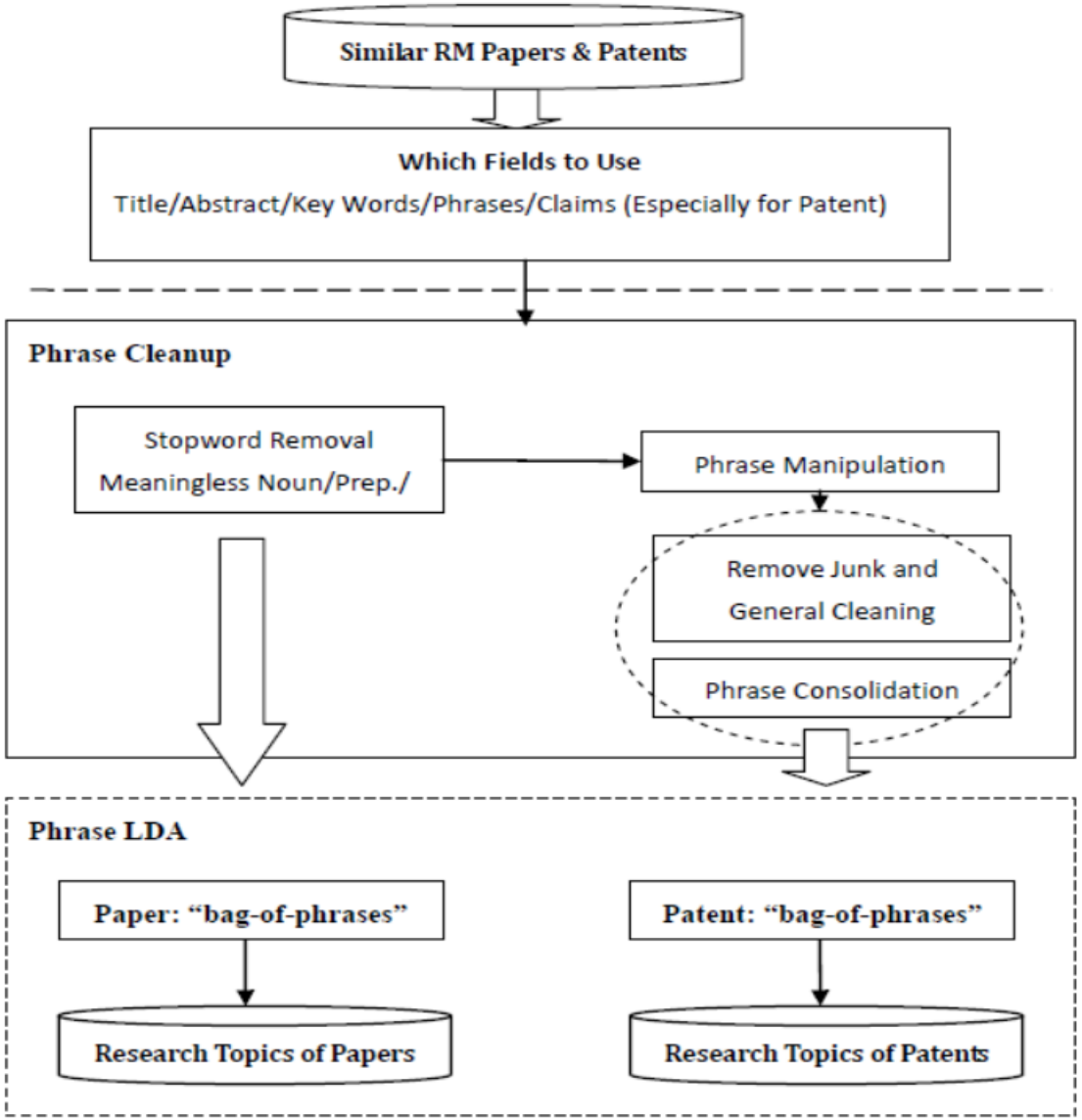
Overview

- Objective
  - Identifying research fronts that have been used at application research in RM .
- Method
  - Scientific Literatures: both scientific papers and patents.
  - Process : text mining and expert verification
    - Collecting papers and patents literatures with Co-keywords
    - Generating *separate* research topics from papers and patents
    - Mining *common* research topics
    - Identifying *applied* research fronts

Step 1: Collecting Papers & Patents with Co-keywords



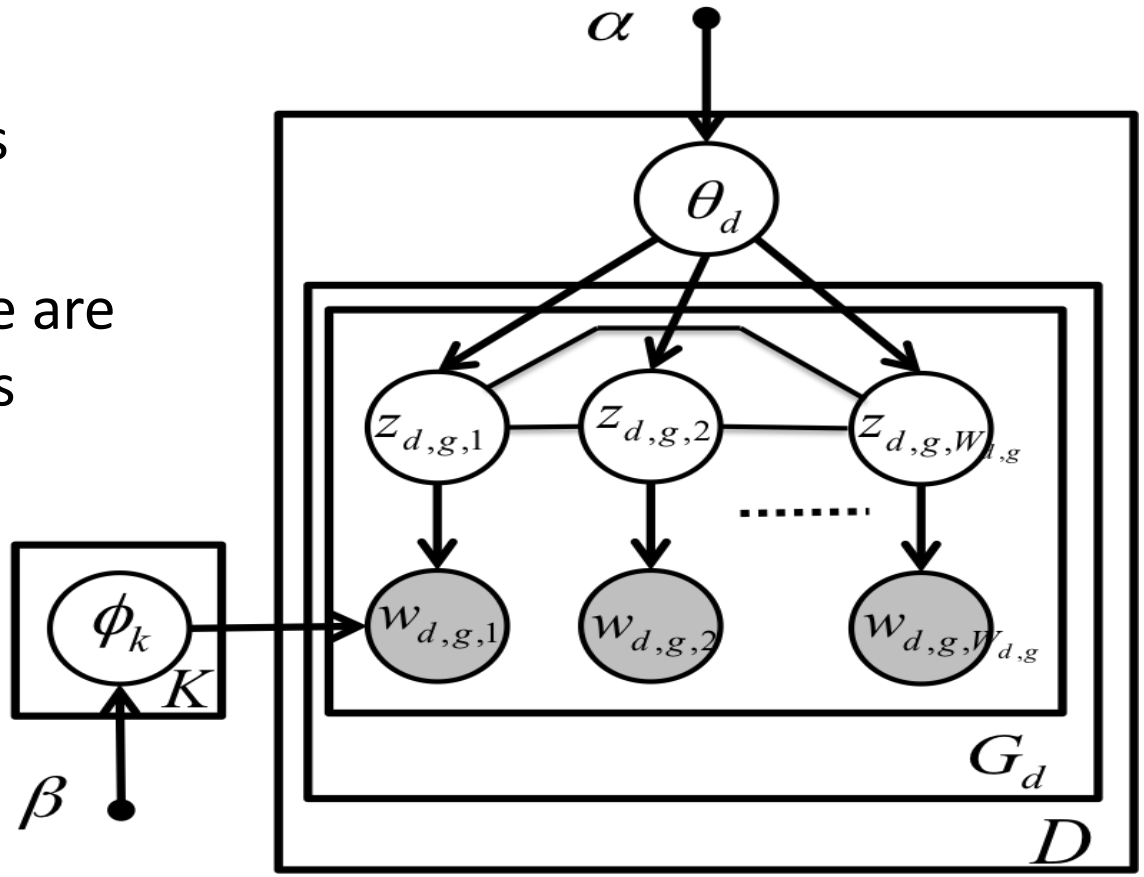
Step 2: Generating Separate Research Topics from Papers & Patents



- **First Phrase Mining then Topic Modeling**
  - Using NLP of TDA to extract candidate phrases and cleanup them
  - Perform agglomerative merging by thesaurus —This segments each paper or patent into a ***"bag-of-phrases"***
  - The newly formed bag-of-phrases are passed as input to PhraseLDA, an extension of LDA, that constrains all words in a phrase to each sharing the same latent topic (MALLET)

Phrase LDA (Constrained Topic Modeling)

- The generative model for PhraseLDA is the same as LDA
- Difference: the model incorporates constraints obtained from the **"bag-of-phrases"** input
  - Chain-graph shows that all words in a phrase are constrained to take on the same topic values



[knowledge discovery] using [least squares] [support vector machine] [classifiers] …

Topic model inference with phrase constraints

By Jiawei Han, et al. 2015

Step 3: Mining Common Research Topics

Cosine Similarity Algorithm

- Common Research Topics: simultaneously appear in scientific papers and patent documents with high similarities
- Topics Representation(LDA):

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

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

- Cosine Similarities of Topics:

$$sim(topic_i, topic_j) = \frac{\sum_{k=1}^{l} p_{ik} * p_{jk}}{\sqrt{(\sum_{k=1}^{l} p_{ik}^2)(\sum_{k=1}^{l} p_{jk}^2)}}$$

Statistical Computing Tool R

- Merge Topics: topics from scientific papers and patent of whose similarities are higher than threshold are merged and chosen as candidates for research fronts.

Step 4: Identifying Applied Research Fronts

- RTA and NRTA Indicators
  - Research Topic Age (RTA): $$RTA(topic_i)=\sum_{i=1}^{n} Y_{kw} * n_i / N$$

$$Y_{kw}=\sum_{i=1}^{n}(Year_{cur}-Year_i) * tfidf_i / (\sum_{j=1}^{n} tfidf_j)$$

Statistical Computing Tool R

RTA reflects time span of research topics, and the larger RTA value is, the wider the time span of distribution of topics

  - Number of Research Topic Authors (NRTA): $$NRTA(topic_i)=n_i / N*100\%$$

NRTA reflects academic attentiveness, the larger NRTA value is, the hotter the topics are

  - Research Topics with smaller RTA and larger NRTA can be considered as Research Fronts.

Case Study

Step 1: Collecting Papers & Patents with Co-keywords

- RM Literature Database
  - Scientific papers: 13,000+ ("Regenerative Medicine[MeSH Major Topic]) OR Regenerative Medicine[Title/Abstract]" from PubMed)
  - Patent: 1,200+ ("Regenerative Medicine in the title or abstract" from EPO)
- Co-Keywords with high frequency
  - Extract keywords from "Title/Abstract/Keywords" of papers and patent documents
  - Threshold of frequency: 2
  - Co-Keywords: 3000+
- Papers & Patents with high Co-keywords :
  - Scientific Papers: 9,655
  - Patent Documents: 1,044

Step 2: Generating Separate Research Topics from Papers & Patents

- Time Span: 2001-2005, 2006-2010,2011-2016
- Phrase LDA for Papers
  - LDA Param: number of iteration =1000; α = 1， β = 0.01
  - Topics: 100,100,100
- Phrase LDA for Patents
  - LDA Param: number of topics = 50; number of iteration =1 000; α = 1， β = 0.01
  - Topics: 50,50,50

Time Span	Topic Number(Science)	Phrase
2001-2005	S1-2	gene therapy, regenerative medicine, reimbursement; cardiovascular disease; mindfulness-based cognitive therapy;…
	S1-36	stem cell differentiation; endothelial cell differentiation; human pluripotent stem cells; nanomatrix gel; stem cell therapy;…
	S1-78	embryonic stem cell, mesenchymal stem cells, MSCs, alanine aminotransferase; d-galactosamine;…

Time Span	Topic Number(Tech)	Phrase
2006-2010	T2-2	induced pluripotent stem cell; mammalian target of rapamycin;apoptosis; cardiac differentiation; Retinal pigment epithelium; RPE cytoskeleton…
	T2-15	DNA analysis; calcium phosphate cement; injectability; mesoporous silica; toxicity analysis; bioactivity…
	T2-28	gene express; gene modification; hyperpolarization-activated cyclic nucleotide-gated; expression analysis;…

Step 3: Mining Common Research Topics

- Merge Science & Technology Topics
  - Cosine Similarities threshold: 0.3
  - Common Research Topics: 68

Time Span	Science Topic	Tech Topic	ST Topic(Merged)
2001-2005	S1-2	T1-6	**ST2**
	S1-36	T1-16	**ST16**
…	…	…	**…**
2006-2010	S2-5	T2-2	**ST5**
	S2-42	T2-15	**ST47**
…	…	…	**…**
2011-2016	S3-1	T3-8	**ST5**
	S3-16	T3-36	**ST62**
…	…	…	**…**

Step 4: Identifying Applied Research Fronts

- Mining Research Fronts from Common S&T Topics
  - Thresholds: RTA is 3.0 ; NRTA is 5%
  - Research Fronts: 15

Time Span	Research Fronts (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%
	…		

Conclusion

- Advantage
  - The method can not only identify research fronts based on scientific papers or patents, but also analyses from the perspectives of science and technology simultaneously at a micro level.
  - Phrase LDA enhance efficiency at manipulating unstructured data and make the topics more understandable.
- Challenge
  - The policy of Papers & Patents with Co-keywords has excluded 30% papers and 10% patents.
  - It is a challenge to properly mine the “common” research fronts between papers and patents.
  - How to represent the research fronts more accurately and understandably? Topics, Phrase or Keywords?
- Future:
  - It can also be applied to track the evolution trends of research fronts.

Thank You!

Acknowledgement: The work was supported by Guangdong Province Science and Technology Program “Integrated Information Service for Regenerative Medicine and Tissue Engineering” (Grant no. 2016A040403098) .