



An Analysis Framework of Research Frontiers Based on The Large-scale Open Academic Graph

Xiaoguang Wang, Hongyu Wang, Han Huang

School of Information Management

Wuhan University, Hubei, China

wanghongyu@whu.edu.cn

JCDL 2019 Workshop 4, June 6

<https://github.com/wanghongyu94/JCDL2019>

CONTENT

01

INTRODUCTION

02

LITERATURE REVIEW

03

FRAMEWORK CONSTRUCTION

04

CONCLUSION

01

PART ONE

INTRODUCTION



1. Introduction——Background

Multiple Cooperation

Interdisciplinary Integration

Emerging Disciplines

A comprehensive, rapid, and accurate detection and analysis of the research frontiers is necessitated

to grasp the situation of sci-tech innovation and to optimize the allocation of scientific research resources

Open Access

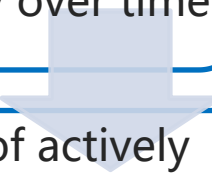
Open Data

With the further development of the **Open Science Movement**, the large-scale open academic graph has gradually become the **new dataset** for the analysis of research frontiers due to the integration of scientific papers and their relevant metadata




1. Introduction——Analysis of Research Problem

The research frontier is a clustering structure among the highly interactive papers, which reflects the topic that have gradually interested people and have been studied more frequently over time



Analysis of research frontiers refers to the process of actively discovering and analyzing emerging research topics and their development status and associated structure by means of expert judgment and scientometrics methods.



Commonly used scientometrics methods of research frontiers analysis include co-citation analysis, bibliographic coupling analysis, word frequency analysis and co-word analysis, et al

Time-delay in citation analysis, lack of semantic information for topic analysis, and inability to cross-integrate of data sources are the main existing shortcomings due to the influence by data sources and analysis principles*

*Ruhai Bai,et al. Research on the Comparison of the Main Methods and Development Trends of Frontier Exploration in Scientific Research[J]. Information Studies:Theory & Application,2017



1. Introduction——Paper Work



- ✓ Pointing out the research optimization direction of the research frontier analysis
- ✓ Investigating a specific academic graph
- ✓ Summarizing the thoughts and steps of the analysis of research frontiers based on the open academic graph
- ✓ Constructing an analysis framework of research frontiers based on the large-scale open academic graph

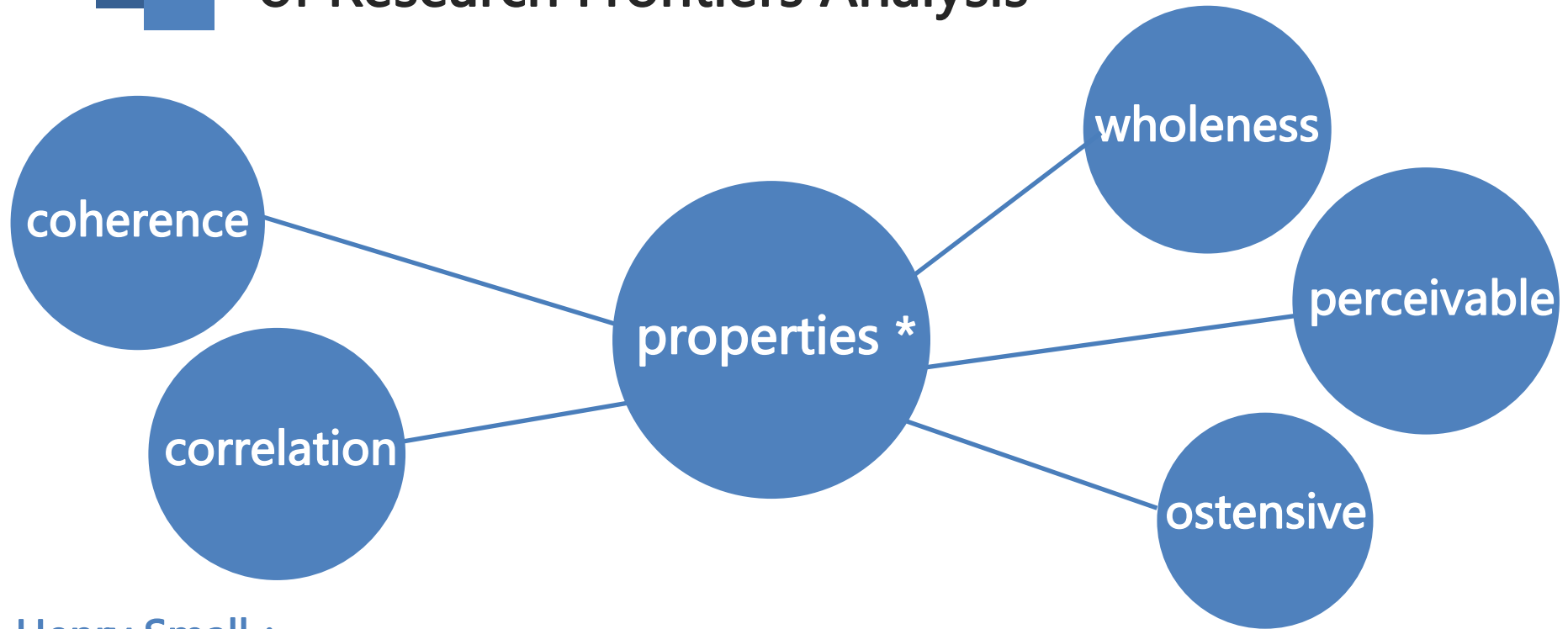
02

PART TWO

LITERATURE REVIEW



2. Literature Review——Conceptions and Methods of Research Frontiers Analysis



Henry Small :

(Small H , Boyack K W , Klavans R . Identifying emerging topics in science and technology[J]. Research Policy, 2014)

Novelty

Growth

Scientometrics methods: citation-based; word-based

Others: comprehensive evaluation based on experts; multi-source data-based and multi-dimensional-based #

*Goldstein J . Emergence as a Construct: History and Issues[J].E m e r g e n c e, 1999

#Rui Luo, et al. A Review of the Main Recognition Methods of Frontier Research[J]. Library and Information Service, 2018



2. Literature Review——Conceptions and Methods of Research Frontiers Analysis

The premise of research frontiers analyze is to **extract the subject topics** from data sources

Subject topics are word representation of the research content in a discipline

Extraction of subject topics: Candidate Keyword Generation, Feature Engineering, and Keyword Extraction *

Multi-dimensions :

TF-IDF

Co-occurrence

CiteTextRank

LDA/Word2Vec

After extraction, how to judge the state and change tendency of the subject topics is one of the difficulties in frontier analysis

Classic Methods: **Word Analysis** (burst detection and the co-word analysis) and **Topic Analysis** (topic model)

Novel Method: Extracting the core nodes in the **community of co-words** so as to analyze the **evolution process** of subject topic and its characteristics#



2. Literature Review——Research Optimization Direction of Frontier Analysis

Limited by the influence of **data source/data scale/analysis principle**, failed to achieve deep, relevant, and dynamic analysis of research frontiers in a **multidimensional, large-scale, and fine-grained manner** :

Strengthen Depth and Breadth

It is necessary to collect data from multiple disciplines for in-breadth and in-depth analysis.

Focus on Dynamic Changes

The scientometrics analysis should focus more on the dynamic evolution of citations and distribution of research topics.

Extended Data Sources

The analysis of research frontiers through multi-source data can obtain more comprehensive, accurate and objective results.

Pay Attention to Multi-factors

Exploring the correlative changes of multi-factors from the perspective of individual topic and group topic can reveal the development of research frontiers more profoundly.

Improving Intelligence

Under the support of academic big data, intelligent models of feature selection and weighting can be established through techniques such as machine learning, thereby reducing manual intervention.



2. Literature Review——The Construction and Application of Open Academic Graph

High-quality

Rich-semantics

Good-structure

Open Academic Graph(OAG): using semantic technologies to fulfill the representation and organization of sci-tech papers and their relevant information and knowledge, and allowing open access to the Internet.*

Typical open academic graphs :

SciGraph


MAG

Aminer

The key support of OAG is **knowledge graph**, which is a new technology of **rich semantic knowledge representation**. It realizes fine-grained and deep-level knowledge organization and representation. The main construction process of it includes **information extraction**, **knowledge fusion** and **knowledge processing**.#

*Lei Xu, et al. Semantic data of scientific publications and their applications[J]. Chinese Journal of Scientific and Technical Periodicals, 2018

#Qiao Liu, et al. Knowledge Graph Construction Techniques[J]. Journal of Computer Research and Development, 2016



2. Literature Review——The Construction and Application of Open Academic Graph

Constructing high-quality academic knowledge graph and exploring its application in sci-tech information analysis has gradually attracted the attention of scholars

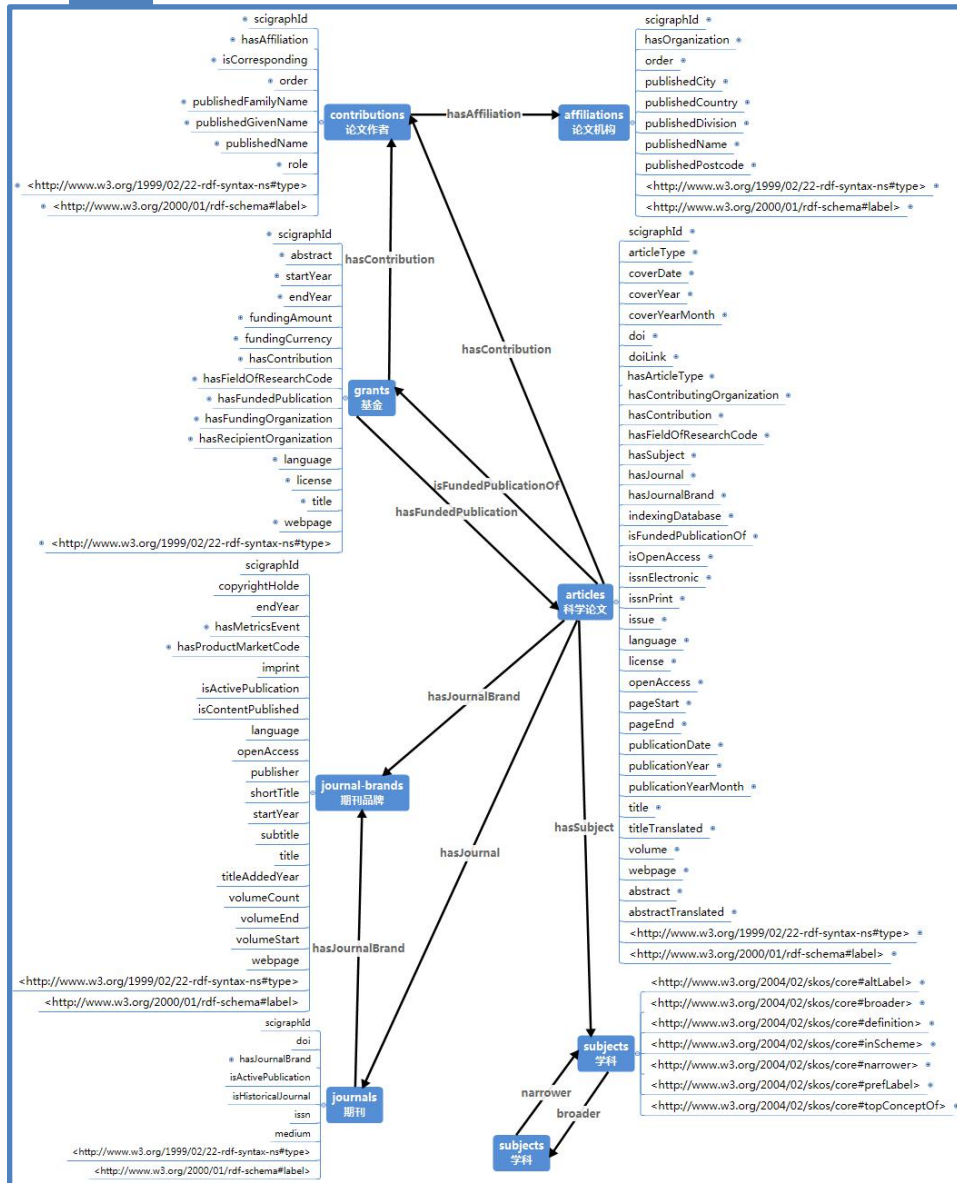
The unified modeling method and semantic information matching algorithm of heterogeneous objects in the scientific knowledge network was proposed

The feasibility of frontier analysis based on OAG was confirmed from the researcher portrait construction and paper impact analysis

03 PART THREE

• FRAMEWORK CONSTRUCTION •

3. Framework Construction——The Investigation of Open Academic Graph Data Structure



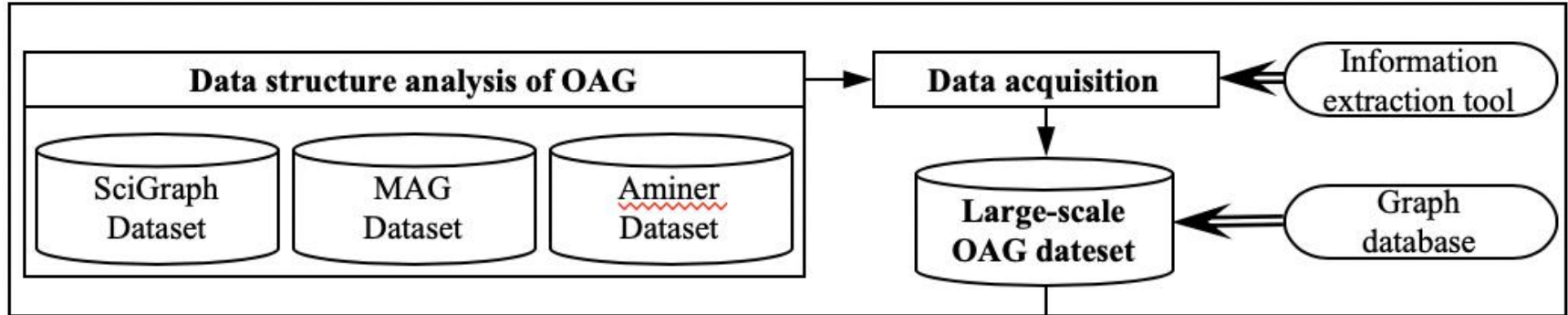
- ✓ The basic conditions for research frontier analysis using large-scale OAGs have been already available
- ✓ SciGraph is a graph stored in the form of triples, which integrates the data of papers and their relevant information
- ✓ By 2017, SciGraph had released metadata for more than 10 million papers, and nearly 1 billion triple data in 12 categories related to it.
- ✓ Discovering the evolution process of subject topics in multi-dimensions and exploring the development tendencies of research frontiers so as to realize the accurate analysis of research frontiers



3. Framework Construction——Overall Framework Construction

Open Data Acquisition :

In this framework, it was necessary to design unified mapping rules based on the data structure of various academic graphs first to complete standardized information acquisition and graphical storage of the multisource heterogeneous large-scale open academic graph.

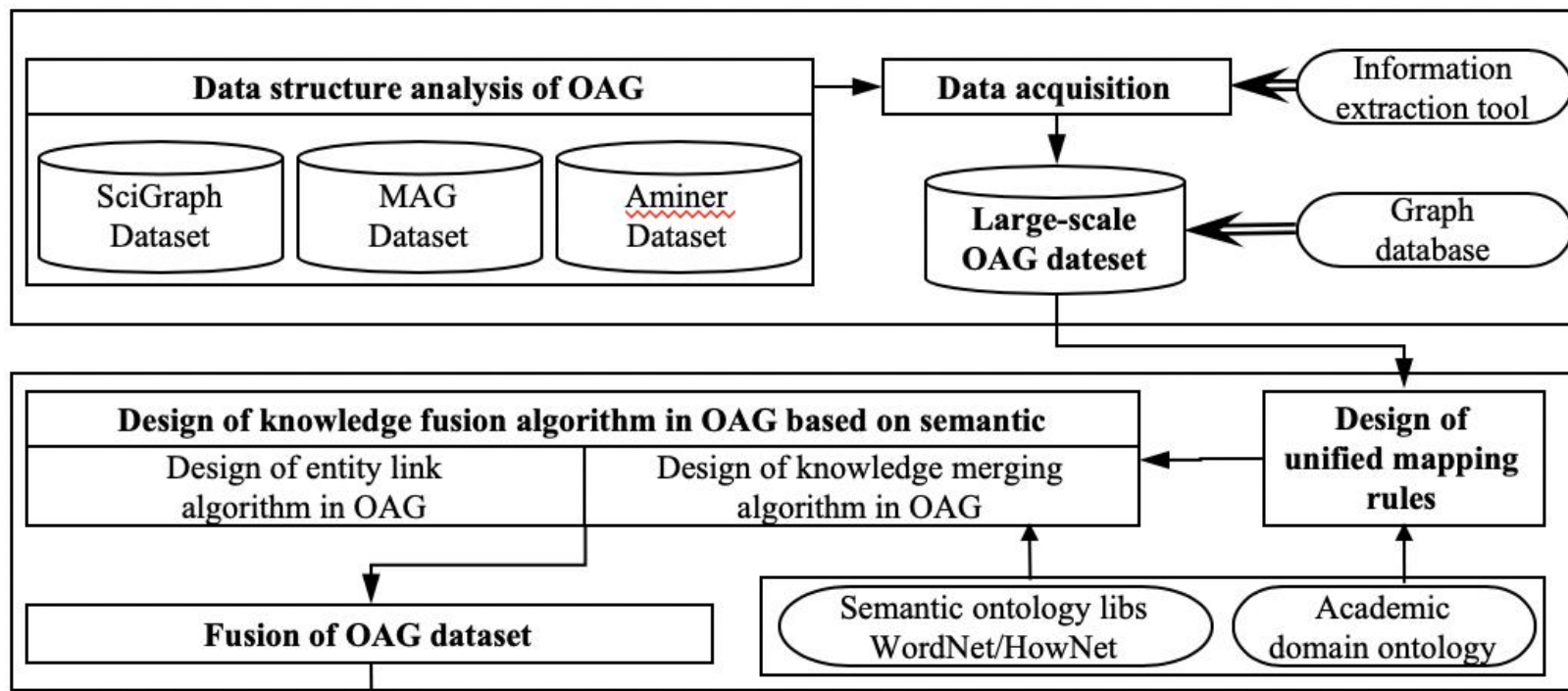




3. Framework Construction——Overall Framework Construction

Academic Graph Fusion :

The semantics-based knowledge fusion algorithm, which is used to complete the entity links, knowledge merge, and alignment between large-scale academic graphs, was designed with reference to the Aminer' s algorithm.

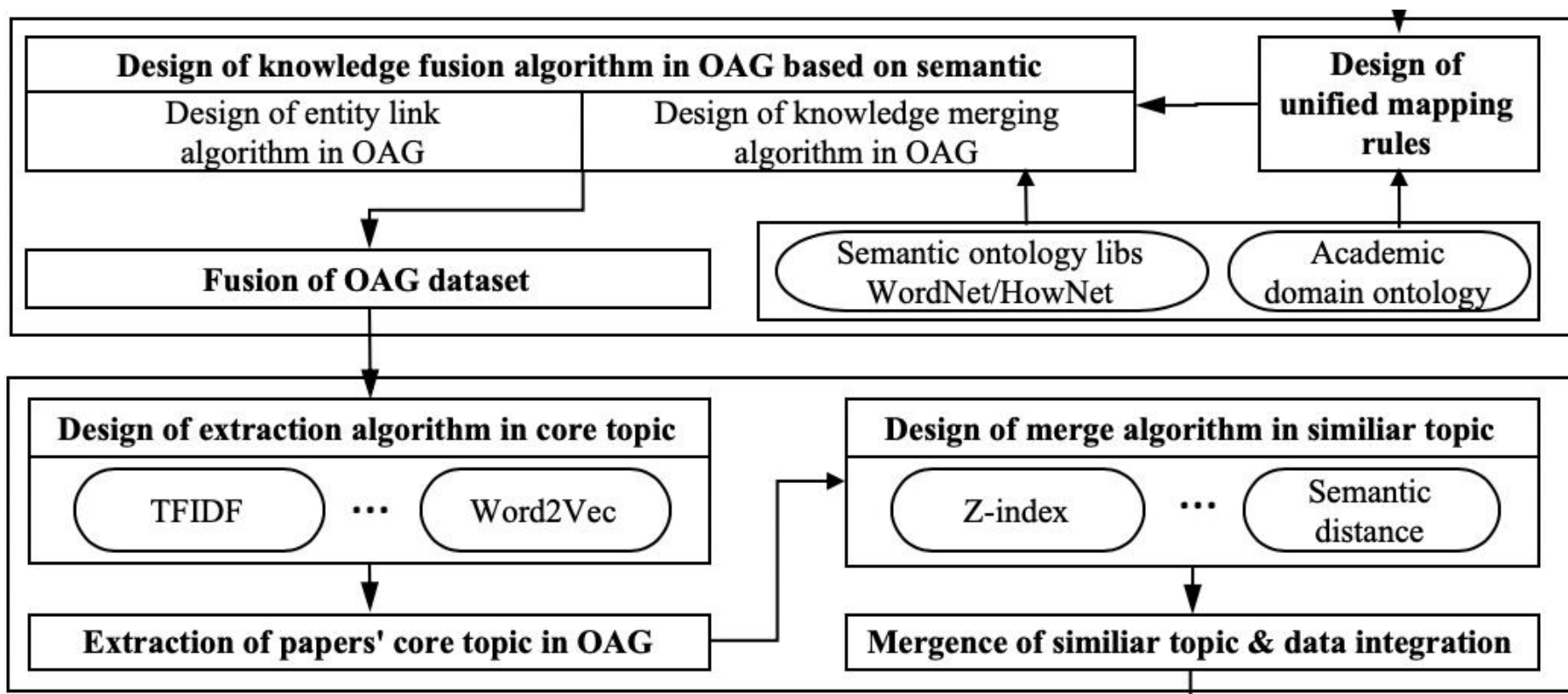




3. Framework Construction——Overall Framework Construction

Core Topic Extraction :

The core topics in the dataset were extracted according to multi-features, such as word frequency, word embedding vector, and topic model, and the similar topics were merged through the designed merge algorithm.

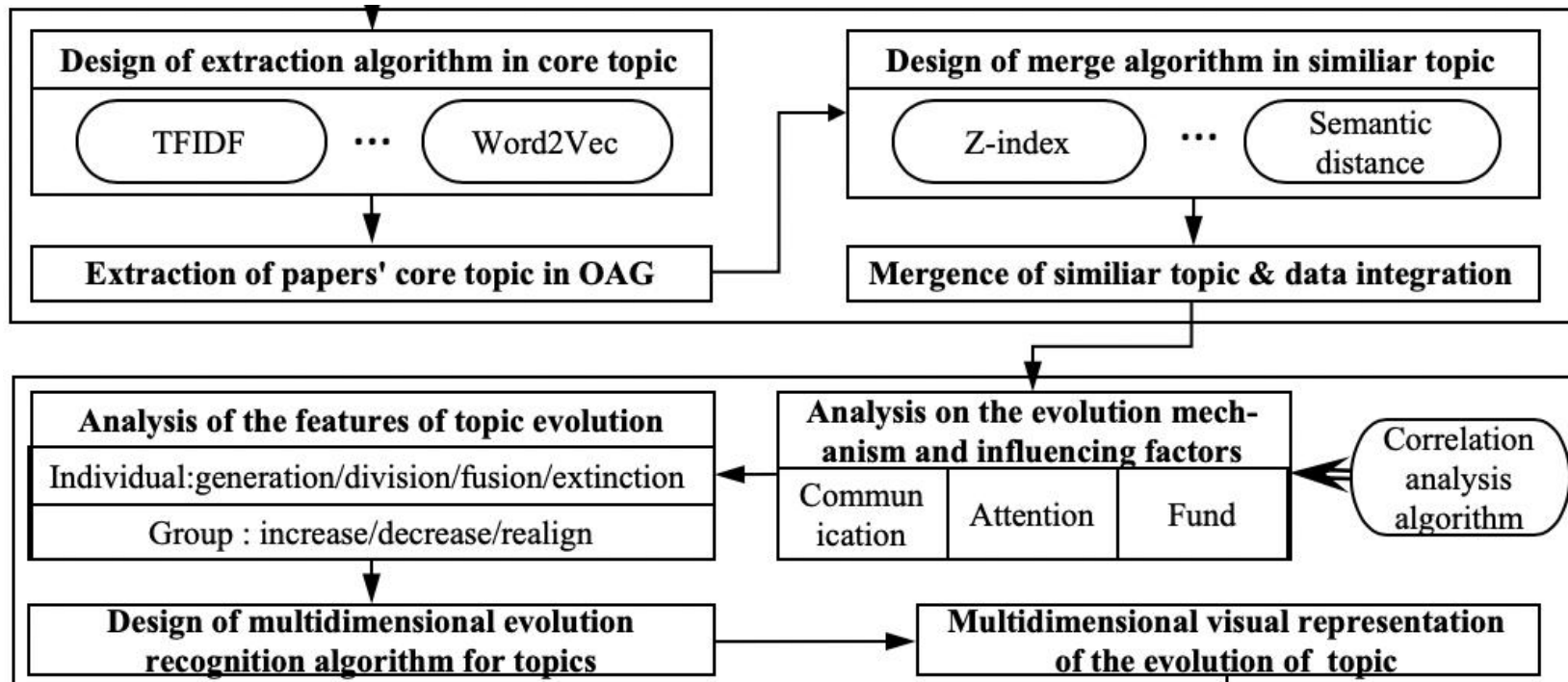




3. Framework Construction——Overall Framework Construction

Multi-dimensional Evolution Analysis :

The evolution mechanism and influencing factors of subject topics were analyzed using the correlation analysis algorithm from multi-dimensions. The expression of topics generation in the academic graphs were revealed to achieve multidimensional evolution analysis and visual representation of topics in the large-scale OAG.

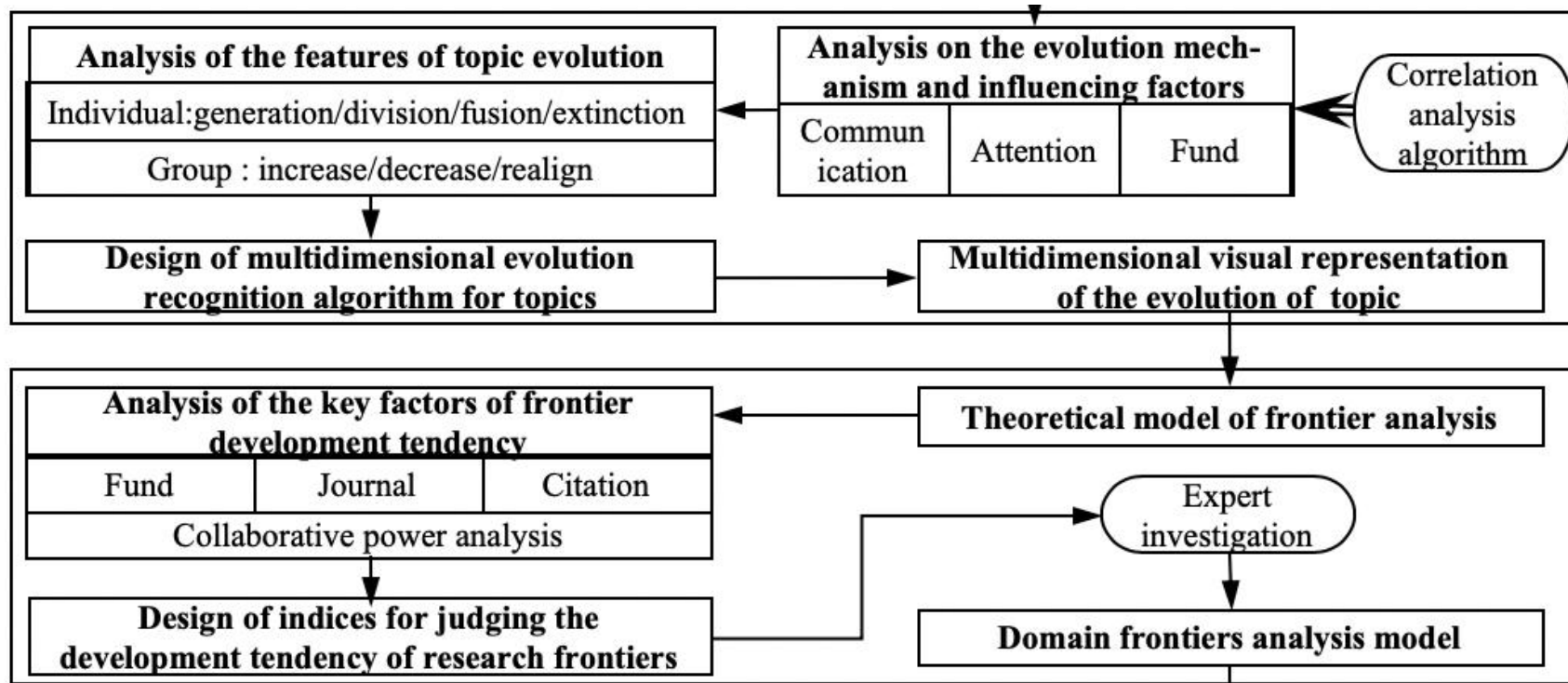




3. Framework Construction——Overall Framework Construction

Frontier Analysis :

A collaborative analysis of the evolution process of individuals and groups in multi-terms was completed. Then, the multivariate indices of the frontier development tendency were designed and the analyzing model of the research frontiers was constructed.

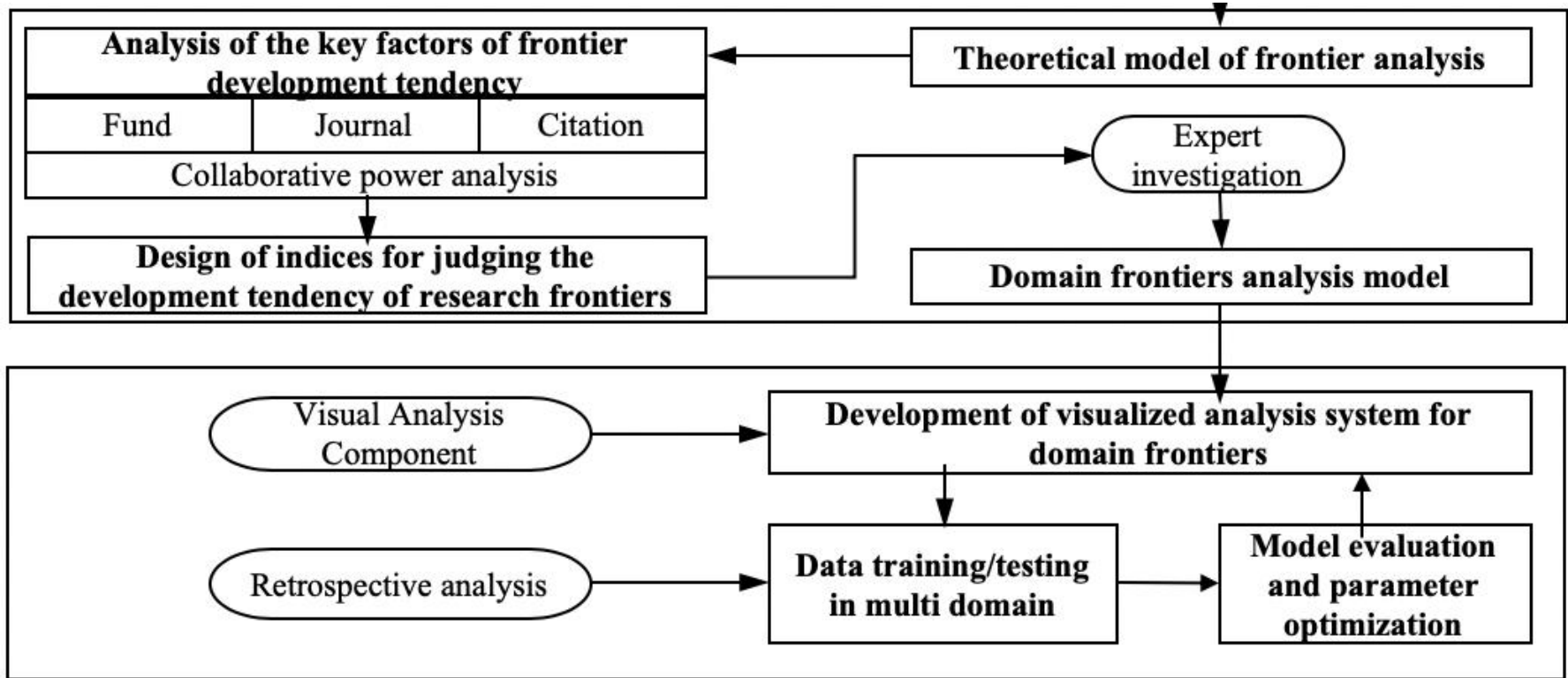




3. Framework Construction——Overall Framework Construction

System Development & Experiment :

A visualized analysis system of research frontiers was developed. Through the training data in various disciplines, the performance of the system was evaluated and optimized, and the comprehensive and accurate analysis of research frontiers was finally realized.



04

PART FOUR

CONCLUSION



4. Conclusion and Discussion

Conclusion

Analyzing the optimizing direction in the following frontier research, putting forward the research thoughts and steps of frontier analysis based on OAG, suggesting the main algorithms and tools to be used, and constructing an overall framework of frontier analysis based on OAG.

Expectation

OAG has created new research conditions for dynamic analysis across time and space, multiscale evolution analysis, multidimensional analysis under multi-factors, and automated intelligent analysis of research frontiers.

Discussion

There are still some problems, such as **the time and space complexity** of large-scale dataset analysis and calculation, the **intuitive visualization** of the evolution process of complex subject topics.



THANK YOU

For Your Listening !

W u h a n
University

Hongyu Wang