ICKM 2017

An Integration-based Conceptual Framework for Scientific Information Analysis

Wei Chen

chenw@whlib.ac.cn

Wuhan Branch of National Science Library, Chinese Academy of Sciences

October 26, 2017

International Council On Knowledge Management

Outline

- 1. Introduction**
- 2. Traditional Scientific Information Analysis Framework**
- 3. Opportunities in Big Data Analytics: Related Studies**
- 4. An Integration-based Conceptual Framework**
- 5. Summary and Future Work**

1. Introduction

- Chinese Academy of Sciences (CAS)
- National Science Library, CAS (NSLC)
- Scientific Information Analysis (SIA)

1.1 Chinese Academy of Sciences (CAS)

104 research institutes
3 universities
In 30+ cities

- China's largest comprehensive R&D organization, as well as the nation's foremost science and technology advisory body (China Top Think Tank).
- 55,000+ Researchers, and 50,000+ Graduate students.

1.2 National Science Library, CAS (NSLC)

Big-data-driven
Knowledge discovery enabler

国家科学图书馆到所服务示意图

- NSLC is proactively leading national efforts to build a shared National Scientific Knowledge Services Infrastructure.

1.3 Scientific Information Analysis (SIA)

- Develop a new SIA framework to expedite big data acquisition and processing, and to improve information service quality.
- A well-organized data management system integrated with a series of proper analytical methods may conduct the SIA tasks more effectively.

2. Traditional SIA Framework

- Traditional SIA Workflow
 - Planning
 - Information Collecting
 - Information Processing and Analysis
 - Results Documentation
 - Evaluation and Feedback
- Problems and Limitations

2.1 Traditional SIA Workflow

2.2 A Case Study

The case

User: The decision-making body of CAS.

Requirements: An investigation of global status of energy storage technologies, and recommendations for domestic development.

Time Limits: Within two months.

2.2 A Case Study (1): Planning

Questions-oriented Analytical Plan

Defined Questions	Analytical Strategies
1. What do major national governments do to promote energy storage technology development?	Describe energy storage technology policy landscape.
2. What are the status of energy storage technologies, and what are the trends of R&D?	Assess technological state-of-the-art and anticipated development trends.
3. Who are the main competitors and potential partners?	Conduct competitive analysis and benchmarking.
4. What R&D plans should China implement in energy storage areas to bridge the gap with leading nations?	Develop recommended strategies and practical solutions.

2.2 A Case Study (2): Information Collecting

2.2 A Case Study (3): Information Processing and Analysis

To process and analyze the research papers and patents

- Excel, TDA[®], and Thomson Innovation[®] as primary tools.
- Address the second and third questions.

To analyze policy and report documents

- Content analysis, Focus group discussion, Delphi method.
- Address the first and second questions

2.2 A Case Study (4): Results Documentation

Intelligence Products

Summary for Decision-makers

Contain a high-level description of main results and, more importantly, policy and technical recommendations for future development.

Complete Analysis Report

Include the detailed analysis process and findings.

Formal Presentation Slides

Focus on key findings illustrated concisely and clearly through graphics, tables, and narratives.

2.2 A Case Study (5): Evaluation and Feedback

2.3 Problems and Limitations

2.3 Problems and Limitations

“In the era of big data, it requires well-organized knowledge management and proper analytical methods designed to deliver the right intelligence at the right time to produce effective decision-making.”

Lee et al., 2012

3. Opportunities in Big Data Analytics: Related Studies

- Data Integration
- Data Analytics

Andrew Kusiak, 2015

3.1 Data Integration

Commercial

Open Source

	Tools	Organizations
Industry	Kettle	Pentaho
	Talend Open Studio	Talend
	Jaspersoft ETL	TIBCO Software Inc.
	CloverETL	Javlin, Inc.
Academia	DeepDive	Stanford Univ.
	DIBBs	Univ. of Arizona
	OBDM	Sapienza Univ. of Rome
	C-DBLP	Renmin Univ. of China

3.2 Data Analytics

Techniques	Data Sources	Area of Interests	Open Source Tools
Structured data analytics	Statistical data, Scientific data	Cluster analysis Factor analysis Correlation analysis Regression analysis Statistical analysis	Excel SPSS Rapidminer KNIME Weka
Text analytics	Documents, Reports, Text content of literature, patents	Document representation Natural language processing Information extraction Topic modeling Clustering	NLTK OpenNLP LingPipe Weka Rapidminer
Web analytics	Webpages	Web content mining Web structure mining Web usage mining	KXEN LIONsolver Dataiku
Multimedia analytics	Images, Audio, Videos	Summarization Annotation Indexing and retrieval Recommendation Event detection	OpenIMAJ LIRE ImageTerrier
Network analytics	Social networks Bibliometric	Link prediction Community detection Social network evolution	Cytoscape UciNet Pajek

4. An Integration-based Conceptual Framework

- Design principle
- System Architecture
 - Data Acquisition
 - Data extraction and integration
 - Data storage
 - Data retrieval
 - Data Analytics

4.1 Design Principle

From Human-driven to Data-driven SIA

collect a wide variety of R&D data and link data across datasets

offer more accurate analysis and visualization of new patterns and trends

different resource specifications, data models, and algorithms can be adapted

enable to quickly access required information and conduct in-depth intelligence analysis

4.2 System architecture

4.2 System architecture (1): Data Acquisition

4.2 System architecture (2): Data extraction and integration

4.2 System architecture (3): Data Storage

4.2 System architecture (4): Data Retrieval

4.2 System architecture (5): Data Analytics

5. Summary and Future Work

- Summary
- Future Research

5.1 Summary

This study proposed an integration-based conceptual framework for SIA, with the development of KRIS as digital infrastructure that can access, store, retrieve, organize, process, and visualize heterogeneous data.

The KRIS is expected to enable us to acquire and manage relevant data, and conduct in-depth intelligence analysis through a variety of analytical methods.

5.2 Future Work

Refine KRIS design by identifying differentiated use cases and needs related to analyzing a great amount of heterogeneous data.

Assess existing data integration and analytics techniques and open source tools so that KRIS could learn from their strengths and avoid the weaknesses.

Explore automatic approaches to construct domain-specific knowledge graphs aim at improving the efficiency and quality of knowledge services.

International Council On Knowledge Management

Thank You!

chenw@whlib.ac.cn