# Postprint: Research on Knowledge Fusion Technology System in Big Data Environments

**Authors:** Chen Mo, Guangjian Li, Li Guangjian

**Date:** 2022-11-25T00:00:00+00:00

## Abstract

This research investigates key technologies for multi-source knowledge fusion in big data environments, constructing a comprehensive technical system by integrating the characteristics of multi-source knowledge objects across different domains, thereby providing technical support and practical solutions for the implementation of knowledge fusion. Employing qualitative analysis methods to examine existing research, combined with inductive and deductive reasoning, and utilizing literature analysis, this study systematically identifies the fundamental problems in knowledge fusion, categorizes the task types involved, and delineates the workflows and specific technologies required for each task, thus forming a knowledge fusion technical system. By comprehensively considering the inherent properties of various technologies, their applicable knowledge objects, and their levels of abstraction, a three-layered architecture comprising a computational layer, functional layer, and task layer is established. These three layers are intricately interconnected, mutually influential, and interlocking: they can be abstracted upward to align with specific knowledge fusion challenges (tasks), while being operationalized downward to yield concrete, computable technical solutions for addressing these challenges.

## Full Text

## Preamble

**Volume 66, Issue 20, October 2022**
*ChinaXiv Cooperative Journal*

## Research on the Knowledge Fusion Technology System in Big Data Environments

**Chen Mo, Li Guangjian**
Department of Information Management, Peking University, Beijing 100081

## Abstract

[**Purpose/Significance**] This study investigates the key technologies for multi-source knowledge fusion in big data environments, constructing a comprehensive technical system tailored to the characteristics of multi-source knowledge objects across different domains, thereby providing technical support and solutions for the practical implementation of knowledge fusion. [**Method/Process**] Using qualitative analysis to examine existing research, combined with inductive and deductive reasoning, and employing literature analysis to identify the problems to be solved in knowledge fusion, this paper summarizes the types of knowledge fusion tasks and the workflows and specific technologies required to implement various tasks, forming a knowledge fusion technology system. [**Result/Conclusion**] By comprehensively considering the inherent characteristics of various technologies, their applicable knowledge objects, and their levels of application abstraction, a three-level technical architecture comprising the computing layer, function layer, and task layer is established. These three layers are interconnected, mutually influential, and interlocking. They can be abstracted upward to associate with specific problems (tasks) of knowledge fusion, and can be concretized downward to identify operable and computable technical methods for solving specific knowledge fusion problems.

**Keywords:** knowledge fusion; technology system; big data
**Classification Number:** G250
**DOI:** 10.13266/j.issn.0252-3116.2022.20.003

## 1. Introduction

The rapid development of science and technology has dramatically increased humanity' s total knowledge volume, with vast amounts of knowledge distributed across different data sources worldwide and presented in various forms. In the information society, knowledge has become a strategic element driving economic growth, and the degree of technological leadership and innovation capacity directly influences the formation of knowledge advantages. Therefore, how to achieve the fusion of knowledge objects at the technical level has become a fundamental and priority issue for knowledge fusion in big data environments, involving technologies and methods spanning resource collection and storage, computational analysis, and access applications.

The term "knowledge fusion technology in big data environments" as used in this paper encompasses two layers of meaning:

First, it refers to technologies developed to solve knowledge fusion problems. Specifically, it denotes a cluster of technologies for addressing various knowledge fusion challenges—not a single technology, but a modular technology combination with recursive structure. Through the configuration of various basic technical components or units, it can ultimately achieve task-oriented, function-specific, and multi-scale knowledge fusion. In the big data era, the characteristics and application processes of knowledge fusion technologies follow their own patterns, reflecting the inherent requirements of knowledge fusion as a special phenomenon. Simultaneously, as research on knowledge fusion deepens and demands for it increase, various information technologies are being improved and enhanced to address the specific features of knowledge fusion, and even specialized implementation technologies need to be developed specifically for knowledge fusion. Here, knowledge fusion technology includes both information technologies that can be applied (directly or after adaptive modification) to the knowledge fusion process and implementation technologies specifically developed for knowledge fusion.

Second, it possesses the characteristics of big data technology itself. In big data environments, knowledge fusion faces data that is massive in volume, low in density, highly real-time, structurally diverse, and from multiple sources—challenges that traditional knowledge fusion technologies and methods struggle to handle, necessitating reliance on new computing architectures. To address massive data volumes, big data knowledge fusion technologies must support the partitioning of ultra-large datasets, distributing them across multiple machines to leverage their storage and computational capabilities. For low data density and strong real-time requirements, big data knowledge fusion technologies have shifted from pure "dataset" processing to "data stream" processing capabilities, enabling real-time processing of large-scale streaming data during its continuous movement. For diverse data structures and multiple sources, big data knowledge fusion technologies can process various data types, including numerical, textual, graphical, image, audio, and video data. More specifically, most commonly mentioned big data technologies belong to the Hadoop ecosystem, so knowledge fusion technologies in big data environments can also be understood as those within the Hadoop ecosystem that can solve knowledge fusion problems. Their origins are twofold: one is knowledge fusion technologies upgraded from traditional ones to conform to Hadoop ecosystem specifications, and the other is knowledge fusion technologies that have emerged within the Hadoop ecosystem itself.

Existing research on knowledge fusion technology methods, whether traditional or big data-oriented, primarily provides underlying technical support for knowledge fusion activities but is generally fragmented and strongly targeted at specific operational levels. Even studies involving knowledge fusion technology systems are basically simple classifications of key technologies, lacking holistic research on the knowledge fusion technology system. For example, X. Yu and Q. Lin [2] classified knowledge fusion methods into four types: ontology-based, rule-based, statistical learning-based, and context-based. Similarly, Qiu Junping and Yu Houqiang [3] divided knowledge fusion technologies into four categories based on implementation paths: semantic rule-based, Bayesian network-based, D-S theory-based, and knowledge mining-based. Gou Jin [4] summarized a classification system of knowledge fusion methods based on genetic algorithms, fusion rules, improved algorithms, and demand-driven algorithms.

In big data environments, knowledge fusion emphasizes the fusion of multi-source knowledge and the integration of knowledge with diverse structures and representations. Even without considering the characteristics of big data ecosystem specifications, the application of knowledge fusion technologies faces new requirements, necessitating the comprehensive utilization of multiple existing knowledge fusion technologies and integrating these underlying technical methods with specific fusion practices. Depending on knowledge fusion requirements and tasks, one may either adjust parameters based on certain underlying technical methods or "combine" and "integrate" multiple underlying technical methods to make them cooperate in solving fusion problems, or even create entirely new technical methods in knowledge fusion practice. Therefore, we need a complete understanding of the knowledge fusion technology system in big data environments that can accommodate various knowledge fusion technologies (including currently unknown ones) and clarify the roles that various technologies can play in knowledge fusion to adapt to the new characteristics of knowledge fusion in big data environments.

This paper aims to construct a knowledge fusion technology system for the big data era by combining the characteristics of multi-source knowledge objects in different fields, providing technical support and solutions for the practical implementation of knowledge fusion, and offering references for the technical theory and application of knowledge fusion. For clarity and simplicity, "big data environment" is omitted before "knowledge fusion technology" in most of the following text.

## 2. Construction of the Knowledge Fusion Technology System

### 2.1 Knowledge Fusion Technology System and Its Construction Approach

A system generally refers to a whole composed of similar things within a certain scope combined according to a certain order and internal connections, consisting of different subsystems. A technology system is the manifestation of technological integrity—a macro-level, socially-oriented overall technical structure [5], typically a unified entity formed by various organically connected technologies with specific functions. The knowledge fusion technology system in the big data era refers to the unified whole that reflects the various information technologies used for knowledge fusion in big data environments and their interrelationships.

As mentioned earlier, knowledge fusion technology does not refer to a single technology but a technology cluster. Literally, it is a collection or system of technologies formed based on their internal associations and differences. Each specific technology holds a different position and role in this collection or system, leading to different understandings of the term "technology" from various perspectives, including: technology as a tool or means; technology as a method or knowledge about methods; technology as human activity (process) or behavior; and technology as a combination or sum of skills, methods, means, tools, and knowledge [6]. Thus, when people use the term "knowledge fusion technology," they often refer to different levels—either the path to completing a knowledge fusion task (combining known means, methods, or tools according to task requirements, or even inventing new ones to form an overall problem-solving framework), or the operational process for processing fusion objects (data and knowledge), including operational steps, rules, and techniques.

Therefore, constructing a knowledge fusion technology system requires 梳理 (sorting out) knowledge fusion concepts at different levels, clarifying their positions and interrelationships within the technology cluster, and reflecting their complementary characteristics in solving knowledge fusion problems. Drawing on previous understandings of the concept of technology and combining them with the needs of this study, this paper categorizes knowledge fusion technologies into three layers: computing layer, function layer, and task layer. Computing layer technologies are the most fundamental operational technologies, oriented toward specific data and covering formulas, variables, models, indicators, algorithms, etc., used in knowledge fusion processes, embodying the notion that "technology is a means, skill, and method." The function layer emphasizes the role technologies play in the knowledge fusion process, referring to unit problems that must be solved to complete knowledge fusion tasks. Any knowledge fusion task can be decomposed into specific problems, and the combination of these problems constitutes the knowledge fusion task. Function layer technologies serve a connecting role: on one hand, they receive results from computing layer data processing (integrating multiple computing layer technologies); on the other hand, combinations of different technologies within the function layer fuse computing layer results to achieve specific knowledge fusion functions, thus roughly embodying the notion that "technology is knowledge about methods" or "human activity (process) or behavior." Task layer technologies mainly reflect the agency of knowledge fusion subjects in completing tasks, representing combinations or overlays of function layer technologies, and embodying the notion that technology is "a combination or sum of methods, means, tools, and knowledge."

To fully reveal the multifaceted characteristics of technologies at different levels in the knowledge fusion technology system and avoid classification based solely on a single (standard) criterion that might cause omissions, this paper employs a multi-dimensional classification method to construct the big data environment knowledge fusion system. Multi-dimensional classification categorizes things or objects from multiple attribute perspectives, using multiple features or facets to subdivide categories. It can perform multiple independent classifications and category divisions of things or objects based on several parallel attribute dimensions through corresponding attribute analysis, and can construct a network structure with multiple category combinations supported by computer technology, thereby achieving multi-angle understanding of things or objects. This approach can adjust attribute dimensions (facets) as things or objects increase or decrease without affecting the original system structure. Within each technology layer, different dimension (classification standard) tables can be established according to the characteristics of technologies at that layer, and each dimension table can be further divided into multiple levels based on relevance. Levels represent further refinement of dimensions, and only multi-dimensional, multi-level divisions can deeply analyze and express the specific content and features of knowledge fusion technologies. Methods at different dimensions and levels can appear multiple times, and this multi-dimensional classification system can reveal the content attributes and external features of knowledge fusion methods from multiple angles in a network-like manner.

### 2.2 Overall Framework of the Knowledge Fusion Technology System

Based on the above construction approach, this paper builds a knowledge fusion technology system for big data environments from three layers: task layer, function layer, and computing layer, as shown in Figure 1 [Figure 1: see original paper].

**Figure 1. Framework of Knowledge Fusion Technology System in the Big Data Era**

**2.2.1 Computing Layer Technologies** The computing layer is the most concrete bottom layer in the knowledge fusion technology system, primarily providing content at various technical implementation levels, such as algorithmic processes, implementation models, and formulas. Each specific technology at this layer typically only implements or completes a part of the knowledge fusion process. Therefore, computing layer knowledge fusion technologies are the technical foundation for achieving knowledge fusion. No matter how complex the knowledge fusion task or function, they ultimately depend on these implementable algorithmic processes, implementation models, and formulas. The achievement of any complex knowledge fusion function or task requires the organic combination and coordination of one or more (mainly multiple) computing layer knowledge fusion technologies. Computing layer knowledge fusion technologies can be further subdivided using different criteria.

**2.2.2 Function Layer Technologies** The function layer refers to various technologies used in different functional segments of the knowledge fusion process to complete knowledge fusion tasks. The combination and overlay of specific technologies from the computing layer achieve concrete knowledge fusion functions. Knowledge fusion functions have always been an important research focus, and studies on both knowledge fusion concepts and systems depend on understanding these functions. Knowledge fusion can be viewed as a complex system that operates on knowledge, and system functions refer to the utilities, efficacy, or uses it can provide in different aspects. Therefore, the system can be described through functional decomposition. Xu Min argued that human knowledge processing essentially solves problems of knowledge form abstraction and transformation, knowledge simplification, discovering similarities and differences in knowledge content, and identifying relationships within knowledge. Accordingly, knowledge fusion functions can be summarized as knowledge transformation, knowledge subdivision, knowledge comparison, and knowledge dependency [7]. This paper adopts this functional classification and discusses the specific implementation technologies required for each function, clarifying the correspondence between function layer and computing layer technologies.

**2.2.3 Task Layer Technologies** Task layer technologies examine knowledge fusion from the perspective of solving specific tasks, where tasks refer to various purposeful activities conducted by knowledge fusion to address concrete problems. Task layer technologies are organic combinations of function layer and computing layer technologies, with multiple functions combined to complete a task, representing the highest level of the technology system. As research objects of knowledge fusion continue to expand, multiple application branches highly related to disciplines or fields have emerged, with different tasks in different disciplines or fields. While it is impossible to enumerate all knowledge fusion tasks across specific disciplines or fields, these tasks share commonalities—that is, setting aside specific knowledge content in disciplines or fields, the activities undertaken to solve disciplinary or field problems are similar, and the technologies used are basically universal. Any uniqueness lies in adapting technologies according to the knowledge content of the discipline or field. Combining previous research [8-11], this paper coarsely divides knowledge fusion tasks into perception, interpretation, decision-making, and prediction. Since each task in knowledge fusion addresses different problems, the required core or main technologies also differ.

In the above knowledge fusion technology system, technologies in the computing layer, function layer, and task layer are not independent but form a progressive relationship that is interconnected, mutually influential, and interlocking. In practical applications, one can start from the topmost task layer to select abstract tasks, then concretize to function layer selections, and finally descend to choose specific algorithms from the computing layer, thereby providing technical support for actual knowledge fusion scenarios. Conversely, one can also start from specific technologies in the computing layer and abstract upward layer by layer to understand which functions and tasks specific technologies can help implement.

The following sections elaborate on the contents of the computing layer, function layer, and task layer in the knowledge fusion technology system.

## 3. Computing Layer Knowledge Fusion Technologies

Computing layer technologies emphasize operability, specifically referring to implementable algorithms, formulas, models, or indicators. These process various types of concrete data, requiring different technologies to handle different data characteristics. Therefore, using the type of processing object as the classification standard for computing layer knowledge fusion technologies is most appropriate. Accordingly, computing layer knowledge fusion technologies are divided into those oriented toward numerical data, textual data, image data, and speech data, as shown in Figure 2 [Figure 2: see original paper].

**Figure 2. Computing Layer Knowledge Fusion Technology System**

### 3.1 Knowledge Fusion Technologies for Numerical Data

The earliest processing objects of knowledge fusion were numerical data, with initial data fusion targeting sensor data and similar numerical information. Much knowledge fusion for other data types ultimately requires converting data into numerical form for further fusion and analysis. Thus, numerical data is an indispensable data type in knowledge fusion processes. Knowledge fusion technologies for numerical data identify effective, novel, and potentially useful data from large, heterogeneous, incomplete, noisy, fuzzy, and random numerical datasets through a series of operations to achieve fusion. Generally, these technologies can be divided into two categories: statistical analysis techniques and mining/prediction techniques.

Statistical analysis techniques characterize general properties of target data, organizing, calculating, and constructing indicator systems for data or information about certain subjects or phenomena to achieve overall understanding and identify patterns. Statistical analysis is a foundational technology for knowledge fusion, essential for implementing most knowledge fusion functions. Commonly used statistical analysis techniques include descriptive statistics, hypothesis testing, correlation analysis, regression analysis, and time series analysis, such as Bayesian methods [12-13] and Kalman filtering algorithms [14]. As G. Feng et al. [15] noted, integrated Bayesian networks are a highly promising technology for knowledge fusion.

Mining and prediction techniques extract, reveal, or induce valuable latent content from large amounts of data or information, constructing and validating relationships, categories, and associations to identify underlying patterns. Commonly used techniques include online analysis, classification, clustering, and association rules, such as ant colony mining classification technology [16], multi-source classification knowledge fusion based on extended concept lattices [17], and multi-source knowledge fusion methods based on supervised learning [18].

### 3.2 Knowledge Fusion Technologies for Textual Data

Knowledge fusion for textual data is currently the most common type across various domains. Textual data is mostly unstructured, and knowledge fusion for such data typically extracts, mines, and reasons about knowledge at the level of documents, paragraphs, phrases, or even entities within articles. For example, in hotspot identification research, keyword co-occurrence analysis is a classic method that establishes co-word matrices and sets thresholds to fuse documents from multiple sources such as scientific reports, patent literature, and academic papers, discovering the most frequently occurring keywords to identify research hotspots or frontiers.

Knowledge fusion technologies for textual data are primarily used for large-scale text mining and analysis, completing tasks such as text topic identification, knowledge system construction, knowledge comparison, and knowledge relationship identification. These technologies mainly include automatic word segmentation, text feature extraction, part-of-speech tagging, named entity recognition, relationship extraction, event extraction, topic modeling and text representation, and text mining and pattern discovery, such as association rule extraction [19], text classification [20], and knowledge fusion algorithms for distributed knowledge objects based on semantic relationships and rules [21].

### 3.3 Knowledge Fusion Technologies for Image Data

Computing layer technologies for image data involve multiple aspects of image processing, including image enhancement and restoration, image segmentation, image recognition, image description, and image compression and decompression. Image data knowledge fusion has two main application scenarios: one is fusion among multiple images, typically integrating images of the same object from different times or from two or more image acquisition devices at the same (or different) time to eliminate contradictions, redundancies, and inconsistencies, thereby obtaining new knowledge about the images [22-23]. The other scenario involves fusing images with text, speech, and other data to deepen understanding of research objects. For example, using text to describe an image allows mapping keywords in the text to corresponding regions in the image, enhancing understanding of both [24]. To achieve these goals, the main challenges are image recognition, image semantic annotation, and image semantic understanding, which constitute the primary technologies needed for image data knowledge fusion, such as high-pass filtering [25] and wavelet transform fusion enhancement [26].

### 3.4 Knowledge Fusion Technologies for Speech Data

In knowledge fusion for audio data, speech data is often combined with graphical and image data. For example, in soccer match videos, besides recognizing player actions and scene changes, analyzing the crowd’ s cheers during goals can fuse audio features like volume and pitch with video content to extract knowledge about goal-scoring segments. This requires speech feature extraction, speech recognition, and other technologies. In recent years, speech recognition and conversion research has achieved remarkable results, with many knowledge fusion studies converting speech to text and then using textual knowledge fusion technologies. Therefore, computing layer technologies for speech data knowledge fusion can be divided into two levels: speech processing and text content analysis. Speech processing includes common techniques such as speech feature extraction [27] and speech recognition [28], along with corresponding text processing technologies.

## 4. Functional Layer Knowledge Fusion Technologies

As previously stated, the essence of knowledge fusion is to meet human functional needs for knowledge processing or simulate human knowledge processing functions. Specifically, it achieves knowledge form abstraction and conversion, discovers similarities and differences in knowledge content, solves the problem of knowledge simplification, and identifies relationships within knowledge through computation. Therefore, functional layer knowledge fusion technologies can be divided into four dimensions: knowledge transformation-oriented, knowledge subdivision-oriented, knowledge comparison-oriented, and knowledge dependency-oriented, as shown in Figure 3 [Figure 3: see original paper].

**Figure 3. Functional Layer Knowledge Fusion Technology System**

### 4.1 Knowledge Transformation-Oriented Technologies

Knowledge transformation refers to the inevitable changes in form, structure, and elements during the process from original knowledge to fused knowledge. Different knowledge must be abstracted and mapped into a unified space to complete fusion, which is foundational work for knowledge fusion. These technologies can be further subdivided into:

#### 4.1.1 Structural Abstraction

Structural abstraction of knowledge represents the logical and positional relationships between different knowledge elements, abstracting and summarizing the physical and logical structures of knowledge to facilitate placing different knowledge under the same standard for various fusion operations. Common structural abstraction techniques include full syntactic parsing, shallow syntactic parsing, and dependency syntactic parsing, such as support vector machine-based syntactic parsing [29-30] and syntactic tree function libraries [31].

#### 4.1.2 Vector Mapping

Knowledge vector mapping converts features in raw data containing knowledge into vector space components. This mapping can mine deeper semantic information embedded in knowledge features and elements, particularly effective for knowledge with unclear structures. The effectiveness of knowledge vector mapping is proportional to data scale and semantic richness—the larger the data volume and the richer the knowledge meaning, the more valuable the fusion results. Common vector mapping techniques include vector space models, bag-of-words models, and probabilistic language models, such as multi-feature fusion based on support vector machines for text recognition and classification [32].

## 4.2 Knowledge Subdivision-Oriented Technologies

Real-world knowledge often has characteristics of integrity, systematicity, and multi-facetedness. From a subdivision perspective, knowledge fusion can be understood as extracting specific elements from different knowledge sources and recombining them, requiring the decomposition or granulation of original knowledge into finer unit knowledge. Therefore, knowledge subdivision mainly involves decomposing knowledge into elements or granulating it into knowledge particles.

### 4.2.1 Element Decomposition

Element decomposition divides knowledge into constituent elements based on attributes, characteristics, and structures at physical and logical levels, enabling humans or machines to select appropriate elements for operation. It is mainly applied in natural language processing, multi-type knowledge feature extraction, knowledge graph applications, and knowledge base construction and supplementation. Common decomposition approaches include word-based, phrase-based, sentence-based, and document-based element subdivision. Mainstream technologies include mathematical statistics, machine learning, and deep learning, such as term extraction methods combining automatically generated filtering dictionaries with lexical density impact factors [33].

### 4.2.2 Granular Computing

Granular computing essentially classifies units within knowledge, extracting and discovering knowledge particles hidden within boundaries by analyzing data structures in knowledge carriers. It can handle uncertain and multi-faceted knowledge in knowledge fusion with simple and efficient processes. Common knowledge particles include those based on fuzzy rules and multi-level granular spaces. Mainstream technologies include rough set-based granular computing, fuzzy set-based granular computing, and quotient space-based granular computing, such as decomposition and merging models for granularity size [34].

## 4.3 Knowledge Comparison-Oriented Technologies

Knowledge comparison identifies similarities and differences between knowledge through certain means. In knowledge fusion, knowledge either shares the same source or represents different sources within the same category, so there must be many similarities and differences. Knowledge comparison aims to discover these, including knowledge similarity and difference.

### 4.3.1 Similarity Comparison

Similarity comparison discovers implicit patterns in knowledge and the proximity between internal and external knowledge units from the perspective of feature similarity. It emphasizes semantic similarity rather than superficial attribute similarity, which is essential for deep knowledge mining and fusion. Common approaches include distance-based, probability-based, and structure-based similarity comparison. Mainstream techniques include distance calculation, path calculation, and semantic analysis, such as feature word distribution, LDA topic distribution, and similarity calculation for citation structure networks [35].

#### 4.3.2 Difference Identification

Difference identification identifies the most significant differences between knowledge through comparison of main features after dimensionality reduction, using these differences as criteria for knowledge distinction or anomaly detection. It is particularly suitable for knowledge with missing values, mutation characteristics, weak structure, or unclear objectives. Common applications include outlier knowledge detection and mutation knowledge detection. Mainstream techniques include principal component analysis and factor analysis, such as technology opportunity identification based on anomaly detection [36].

### 4.4 Knowledge Dependency-Oriented Technologies

Dependency refers to the coexistence and mutual attachment of two or more things. Since things in the world are universally related, knowledge from different sources in knowledge fusion processes usually has certain associations that can be exploited to discover relationships between knowledge pieces.

#### 4.4.1 Association Analysis

Association analysis seeks association or correlation characteristics in large knowledge collections, ultimately forming descriptions of patterns and rules for simultaneous occurrence of certain traits or attributes. Before processing knowledge, the potential patterns and rules are often unknown, making association analysis more of a data-driven knowledge fusion that can discover high-value or easily overlooked knowledge. Common approaches include rule-based, probability-based, and path-based association analysis. Mainstream techniques include association rule mining, graph models, and probabilistic statistical models, such as using association rule models to reveal deep relationships behind co-occurrence patterns and help detect hidden patterns in technology development [37].

#### 4.4.2 Relationship Discovery

Relationship discovery mines large numbers of explicit or implicit relationships within knowledge, starting from strong internal or external correlations. Its characteristic is having preliminary expectations about potential relationships or conditions before processing, making it more of a business-driven knowledge fusion. Common approaches include trigger condition-based and mathematical statistics-based relationship discovery. Mainstream techniques include logical languages, sequence labeling, and neural networks, such as non-taxonomic relationship extraction [38].

## 5. Task Layer Knowledge Fusion Technologies

Task layer knowledge fusion technologies are essentially comprehensive methods, primarily combinations of various computing layer technologies that achieve specific knowledge fusion functions to complete concrete tasks. These include perception-oriented, explanation-oriented, decision-making-oriented, and prediction-oriented knowledge fusion technologies, as shown in Figure 4 [Figure 4: see original paper].

**Figure 4. Task Layer Knowledge Fusion Technology System**

### 5.1 Perception-Oriented Knowledge Fusion Technologies

These technologies primarily address the discovery, collection, and acquisition of data and knowledge, including information perception, intelligent search, and knowledge discovery.

#### 5.1.1 Information Perception

Information perception integrates data or information about the same object across different times and spaces, using effective interconnection of multiple information networks as the foundation to achieve integrated and three-dimensional collection and discovery, thereby drawing relevant conclusions. Data and information are the most basic objects of knowledge fusion, and information perception has always been an important research topic. Knowledge fusion information perception technologies are typically based on linear algebra, probability theory, random processes and statistical computation, information extraction, semantic modeling, and automatic clustering. For example, in crime fighting, these technologies can already be comprehensively used to perceive and obtain valuable crime clues from massive Web information sources [39].

#### 5.1.2 Intelligent Search

Intelligent search is the product of combining search technology with knowledge fusion technology. It employs knowledge fusion technologies such as sentiment analysis and metaphor recognition to achieve deep understanding of information content and perception and recognition of information scenarios during the collection phase, thereby more accurately filtering valuable information directly related to research problems, reducing "noise" in collected information, avoiding information overload, and even consciously "optimizing" the direction and scope of information collection based on scenarios involved in already-collected information to actively discover seemingly unrelated or previously unconsidered information.

#### 5.1.3 Knowledge Discovery

Knowledge discovery is the process of obtaining knowledge from various information sources according to different needs. Its purpose is to shield users from tedious details of raw data and extract effective, novel, and potentially useful knowledge from raw data [40]. Knowledge discovery can be seen as an advanced form of information perception and intelligent search, producing high-quality integration results from data, information, or knowledge to knowledge. Knowledge discovery involves three types of technologies: knowledge base construction [41], knowledge representation and reasoning [42-44], and knowledge application [45].

### 5.2 Explanation-Oriented Knowledge Fusion Technologies

In big data environments, knowledge from multiple domains and data sources involves different disciplines and may be expressed in different forms, lacking common understandability. Therefore, mutual explanation of knowledge from different disciplines is necessary to achieve understanding during knowledge fusion.

#### 5.2.1 Numerical Explanation-Oriented Technologies

Data fusion is an important component of knowledge fusion and a basic stage that can be divided into two categories: non-semantic fusion and semantic-based fusion. Non-semantic fusion mainly uses Kalman filters, Bayesian networks, and rule-based technologies [46] to achieve numerical interpretation, while semantic-based fusion primarily uses metadata and ontology technologies to convert raw data into semantic description formats like XML and RDF to aid understanding of data meaning [47].

#### 5.2.2 Textual Explanation-Oriented Technologies

These technologies extract knowledge and relationships from large amounts of dispersed textual content and fuse them semantically according to existing prior knowledge and certain narrative logic, helping users understand causal relationships between knowledge pieces. Textual semantic explanation in knowledge fusion mainly involves rule-based, statistics-based, and machine learning-based natural language processing, information extraction, and causal inference technologies, with typical applications including cross-document entity relationship extraction and merging [48-49], sentence fusion and ranking [50], and document topic clustering and summarization [51-52].

#### 5.2.3 Image Explanation-Oriented Technologies

These technologies aim to reveal deep semantics hidden behind image content to help users correctly understand images. Generally, image explanation-oriented knowledge fusion is essentially semantic fusion of images—integrating multiple images of the same scene with complementary information and redundant features into a complete, information-rich, and clearly meaningful image. This semantic fusion is comprehensive work that can be divided into three aspects: image semantic segmentation, image semantic annotation, and semantic-based image fusion [53].

### 5.3 Decision-Making-Oriented Knowledge Fusion Technologies

These technologies integrate information and knowledge from multiple sources into new common knowledge that can be used for decision-making and problem-solving or provide better insights and understanding of situations. Decision-making can be classified from different perspectives; here we focus on knowledge fusion technologies in emergency decision-making, long-term decision-making, and intelligent decision-making tasks.

#### 5.3.1 Emergency Decision-Making-Oriented Technologies

Emergency decision-making requires decision-makers to use existing knowledge to judge and select satisfactory handling strategies for emergency events within limited time. In such tasks, decision-makers use existing knowledge for responding to and handling 突发事件 (emergencies), combined with current situational information, to understand and comprehend decision problems and form decisions. Emergency decision-making-oriented knowledge fusion typically uses context identification [54], fuzzy set theory [55], and scenario analysis [56].

#### 5.3.2 Long-Term Decision-Making-Oriented Technologies

Long-term decision-making addresses potential events over decades or longer [57]. Such knowledge fusion tasks extract and synthesize possible evaluation criteria from numerous decision-makers to select criteria for determining decision solutions. Current typical approaches use sentiment analysis, fuzzy sets, and entropy methods for decision criteria fusion [58-59].

#### 5.3.3 Intelligent Decision-Making-Oriented Technologies

Intelligent decision support systems are products combining artificial intelligence and decision support systems—interactive computer-based systems that assist decision-makers in decision-making, pursuing automated decision-making to reduce human dependency and avoid subjective interference. Statistical analysis, knowledge extraction, knowledge representation, knowledge comparison, and relationship discovery technologies play important roles in this process with successful applications [60].

### 5.4 Prediction-Oriented Knowledge Fusion Technologies

In big data environments, knowledge fusion directly participates in problem-solving and complex decision-making. Theoretically, scholars have proven that prediction driven by multi-source information and knowledge fusion and collective intelligence is highly effective [61-62]. Consequently, academia and industry increasingly emphasize using knowledge fusion to solve prediction problems.

#### 5.4.1 Technology Prediction-Oriented Technologies

These use knowledge fusion to predict technology development. For example, K. Heyel proposed a multi-source data fusion framework for emerging technology intelligence analysis [63], which comprehensively uses traditional scientific publications, Web academic data, Wikipedia data, etc., through latent semantic analysis and knowledge extraction to achieve technology analysis and prediction.

#### 5.4.2 Human-Machine Hybrid Prediction-Oriented Technologies

These organically involve analysts in intelligent interactive knowledge fusion processes during prediction, complementing machine behavior to collaboratively complete complex prediction tasks. DARPA and University of Southern California researchers have made beneficial attempts with good results [64-65].

#### 5.4.3 Crowdsourcing Prediction-Oriented Technologies

These leverage collective intelligence effects in prediction tasks to obtain relatively optimal results. The essence is aggregating analysts' wisdom, knowledge, and information through network platforms and forming optimized prediction results through market mechanisms. A typical representative is the methodology proposed by P.E. Tetlock et al. [66] for aggregating collective intelligence for prediction, emphasizing evidence collection from multiple sources, probabilistic thinking, teamwork, multi-factor analysis, and continuous improvement.

## 6. Conclusion

The development and application of information technology in the big data era have greatly enriched knowledge fusion methods and systems. Starting from the multi-source, heterogeneous data characteristics of big data era knowledge fusion and combining the big data technology ecosystem, this paper analyzes the connotation of knowledge fusion technology in big data environments, conducts in-depth analysis of knowledge fusion characteristics and key implementation technologies, and uses functions in the knowledge fusion implementation process as an intermediary. By comprehensively considering various technologies' characteristics, applicable knowledge objects, and application abstraction levels, it establishes a three-level technical architecture comprising computing layer, function layer, and task layer, and summarizes technologies at each layer.

Among them, computing layer knowledge fusion technologies are the most concrete, reflecting the computability, measurability, and verifiability of various technologies, manifested as algorithmic processes, implementation models, and formulas. Function layer knowledge fusion technologies focus on their roles in the knowledge fusion process. Any knowledge fusion work requires specific technologies to achieve certain functions, which are formal descriptions of how various data (knowledge) processing technologies can function. Computing layer technologies are "bottom-level" operational technologies that strictly process "data" rather than "knowledge," while function layer technologies elevate "data" processing to "knowledge" processing on the basis of computing layer results, laying the foundation for solving knowledge fusion tasks.

Building on previous research, this paper abstractly defines knowledge fusion tasks as perception, interpretation, decision-making, and prediction—a distinctive feature of this study. Task layer knowledge fusion technologies are comprehensive methods that combine various computing layer technologies and/or function layer technologies to achieve knowledge fusion functions and complete specific tasks, with clear specificity and purposefulness.

The knowledge fusion technology system constructed in this paper is a multi-layered, multi-dimensional, and three-dimensional system that compensates for previous research' s limitation of classifying technologies mainly by technical principles while paying insufficient attention to application scenarios. The three-layer system is interconnected, mutually influential, and interlocking. It can be abstracted upward to associate with specific knowledge fusion problems and tasks, and concretized downward to guide the selection of operable and computable technical methods for knowledge extraction, collection, organization, modeling, combination, reasoning, and integration. This system not only has epistemological significance for understanding technologies involved in knowledge fusion processes but also provides references for technology selection and guides practical knowledge fusion work.

## References

[1] Dong Xiaoying. The theoretical basis and strategic choice of knowledge advantage[J]. Journal of Peking University (Philosophy and Social Sciences Edition), 2004(4): 37-45. [2] Yu X, Lin Q. Knowledge fusion methods: a survey[C]//Proceedings of the 2017 2nd international conference on software, multimedia and communication engineering. Colorado: DEStech Publications, 2017: 300-304. [3] Qiu Junping, Yu Houqiang. Research progress and trends of international knowledge fusion from the perspective of knowledge science[J]. Library and Information Service, 2015, 59(8): 126-132, 148. [4] Gou Jin. Research on several key technologies in knowledge fusion[D]. Hangzhou: Zhejiang University, 2006. [5] Jiang Zhenhuan, Wu Mingtai, Wang Haishan, et al. Dictionary of Technology[M]. Shenyang: Liaoning Science and Technology Press, 1990. [6] Yang Kaicheng, Wang Bin. Viewing the nature of educational technology from the nature of technology[J]. China Educational Technology, 2007(9): 1-4. [7] Xu Min. Research on patterns and methods of knowledge fusion technology from the perspective of knowledge computing[D]. Beijing: Peking University, 2019. [8] Dong X L, Gabrilovich E, Heitz G, et al. From data fusion to knowledge fusion[J]. arXiv preprint arXiv:1503.00302, 2015. [9] Roemer M J, Kacprzynski G J, Orsagh R F. Assessment of data and knowledge fusion strategies for prognostics and health management[C]//2001 IEEE aerospace conference proceedings. New York: IEEE, 2001: 2979-2988. [10] Preece A, Hui K, Gray A, et al. KRAFT: an agent architecture for knowledge fusion[J]. International journal of cooperative information systems, 2001, 10(1/2): 171-195. [11] Tsukasa I, Takenaka T, Motomura Y. Customer behavior prediction system by large scale data fusion in retail service[J]. Transactions of the Japanese Society for Artificial Intelligence, 2011, 26(6): 670-681. [12] Feng G, Zhang J D, Liao S S. A novel method for combining Bayesian networks, theoretical analysis, and its applications[J]. Pattern recognition, 2014, 47(5): 2057-2069. [13] Zhou Fang, Han Liyan. Corporate failure discrimination method based on knowledge fusion[J]. Communication of Finance and Accounting, 2015(8): 61-63. [14] Saha R K, Chang K C. An efficient algorithm for multisensor track fusion[J]. IEEE transactions on aerospace and electronic systems, 1998, 34(1): 200-210. [15] Feng G, Zhang J D, Liao S S. A novel method for combining Bayesian networks, theoretical analysis, and its applications[J]. Pattern recognition, 2014, 47(5): 2057-2069. [16] Martens D, De Backer M, Haesen R, et al. Ant-based approach to the knowledge fusion problem[C]//International workshop on ant colony optimization and swarm intelligence. Berlin: Springer, 2006: 84-95. [17] Ma Feng. Research on multi-data source classification knowledge fusion problem based on extended concept lattice[D]. Hefei: Hefei University of Technology, 2006. [18] Dong

X, Gabrilovich E, Heitz G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion[C]//Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2014: 601-610. [19] Tai C H, Chang C T, Chang Y S. Hybrid knowledge fusion and inference on cloud environment[J]. Future generation computer systems, 2018, 87(1): 568-579. [20] Gruber T R. Ontolingua: a mechanism to support portable ontologies[R]. California: Knowledge Systems Laboratory, 1992. [21] Gou J, Jiang Y, Wu Y, et al. A new knowledge fusion method based on semantic rules[C]//International conference on signal processing. New York: IEEE, 2006: 1939-1942. [22] Zhang Lingkai, Yu Liang. A review of multi-source remote sensing data fusion research[J]. Urban Geography, 2017(2): 173. [23] Zhang Xiaofei, Wang Qiang, Wei Chunrong, et al. A review of medical image fusion technology research[J]. Guangxi Sciences, 2002(1): 64-68. [24] Karpathy A, Li F. Deep visual-semantic alignments for generating image descriptions[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 9(4): 664-676. [25] Chavez P S, Sides S C, Anderson J A. Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic[J]. Photogrammetric engineering & remote sensing, 1991, 57(3): 295-303. [26] Li H, Manjunath B S, Mitra S K. Multisensor image fusion using the wavelet transform[J]. Graphical models and image process, 1995, 57(3): 235-245. [27] Li Qi, Sun Guiling, Huang Cui, et al. Research on target recognition method based on multi-modal deep fusion in underwater acoustic environment space[J]. Journal of Ocean Technology, 2019, 38(6): 35-42. [28] Hu Yu. Artificial intelligence and speech recognition technology[J]. Electronic Engineering & Product World, 2016, 23(4): 23-25, 27. [29] Bravo À, Cases M, Queralt-Rosinach N, et al. A knowledge-driven approach to extract disease-related biomarkers from the literature[R]. BioMed research international, 2014: 1-11. [30] Bravo À, Piñero J, Queralt-Rosinach N, et al. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research[J]. BMC bioinformatics, 2015, 16(1): 1-17. [31] Chen Jianmei, Lin Hongfei. Construction of Chinese emotional common sense knowledge base[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(4): 492-498. [32] Gong An, Fei Fan. Sentiment analysis of review text based on multi-feature fusion[J]. Computer Technology and Development, 2018(8): 91-95. [33] Zeng Zhen, Lü Xueqiang, Li Zhuo. A domain term extraction method for patent abstracts[J]. Computer Applications and Software, 2016, 33(3): 48-51. [34] Hobbs J R. Granularity[C]//Proceedings of the ninth international joint conference on artificial intelligence. Los Angeles: Morgan Kaufmann, 1985: 432-435. [35] Ding Mengxiao, Bi Qiang, Xu Pengcheng, et al. Precision recommendation of knowledge discovery services based on user interest measurement[J]. Library and Information Service, 2019, 63(3): 21-29. [36] Zhai Dongsheng, Guo Cheng, Zhang Jie, et al. Research on technology opportunity identification method using anomaly detection[J]. New Technology of Library and Information Service, 2016(10): 81-90. [37] Zhou Lei, Yang Wei. Technology fusion detection based on weighted association rules[J]. Journal of Intelligence, 2019, 38(1):

67-72. [38] Zhu Hui, Wang Hao, Su Xinning, et al. Research on non-taxonomic relationship extraction method for Chinese domain terms[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(12): 1193-1203. [39] Kejriwal M, Szekely P, Knoblock C. Investigative knowledge discovery for combating illicit activities[J]. IEEE intelligent systems, 2018(1): 53-63. [40] Lu Xiongwen. Dictionary of Management[M]. Shanghai: Shanghai Lexicographical Publishing House, 2013. [41] Lenat D B, Prakash M, Shepherd M. CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks[J]. AI magazine, 1985, 6(4): 65-65. [42] Liu Jianwei, Yan Lufeng. Comparison of knowledge representation methods[J]. Computer Systems & Applications, 2011, 20(3): 242-246. [43] Huang Degen, Zhang Yunxia, Lin Hongmei, et al. Classification model based on rule reasoning network and its application on expert system for software fault diagnosis[J]. Journal of Software, 2020, 31(4): 1063-1078. [44] Chen Y W, Yang J B, Xu D L, et al. Inference analysis and adaptive training for belief rule-based systems[J]. Expert systems with applications, 2011, 38(10): 12845-12860. [45] Yan R, Li G, Liu B. Knowledge fusion based on DS theory and its application[J]. Expert systems with applications, 2011, 38(10): 12845-12860. [46] Partescano E, Brosich A, Lipizer M, et al. From heterogeneous marine sensors to sensor Web: (near) real-time open data access adopting OGC sensor Web enablement standards[J]. Open geospatial data, software and standards, 2017, 2(1): 1-9. [47] Sheth A, Henson C, Sahoo S S. Semantic sensor Web[J]. IEEE Internet computing, 2008, 12(4): 78-83. [48] Li X, Ma S, Zhou X. Large-scale Chinese cross-document entity disambiguation and information fusion[C]//Advancing big data benchmarks. Berlin: Springer, 2013: 105-119. [49] Bronselaer A, Van Britsom D, De Tre G. A framework for multiset merging[J]. Fuzzy sets and systems, 2012, 191(1): 1-13. [50] Yue L, Shi Z, Han J, et al. Multi-factors based sentence ordering for cross-document fusion from multimodal content[J]. Neurocomputing, 2017, 253(1): 6-14. [51] Witter S, Bergler S. Fuzzy clustering for topic analysis and summarization of document collections[C]//Conference of the canadians society for computational studies of intelligence. Berlin: Springer, 2007: 476-488. [52] Lebanoff L, Wang B, Feng Z, et al. Modeling endorsement for multi-document abstractive summarization[J]. arXiv preprint arXiv:2110.07844, 2021. [53] Ni Jingxiu. Research on key technologies of image semantic fusion[D]. Beijing: China University of Mining and Technology, 2018. [54] Smirnova V, Levashova T, Shilov N. Knowledge fusion in context-aware decision support systems[C]//Proceedings of the international joint conference on knowledge discovery, knowledge engineering and knowledge management. Rome: KEOD, 2014: 186-194. [55] Zhang Lei. Research on emergency decision knowledge fusion method with fuzzy uncertainty[D]. Dalian: Dalian University of Technology, 2019. [56] Zhang Zhixia, Hao Wenhui. Dynamic scenario model for sudden disaster accidents based on knowledge elements[J]. Oil & Gas Storage and Transportation, 2019, 38(9): 980-987, 995. [57] The RAND Corporation. Seeking examples of long-term decisions[EB/OL]. [2022-08-20]. https://www.rand.org/pard/LongTermDecisions/seeking.html. [58] Çaliş Y, Balaman Y. Improved decisions for marketing, supply and purchasing: mining

big data through an integration of sentiment analysis and intuitionistic fuzzy multicriteria assessment[J]. Computers & industrial engineering, 2019, 129(1): 315-332. [59] Chen X, Zhang W, Xu X, et al. A public and large-scale expert information fusion method and its application: mining public opinion via sentiment analysis and measuring public dynamic reliability[J]. Information fusion, 2022, 78(1): 71-85. [60] Morge M. The hedgehog and the fox[C]//International workshop on argumentation in multi-agent systems. Berlin: Springer, 2007: 114-131. [61] Charbonneau S, Fye S, Hay J, et al. A retrospective analysis of technology forecasting[C/OL]//AIAA SPACE 2013 conference and exposition. California: American Institute of Aeronautics and Astronautics, 2013[2022-06-14]. https://doi.org/10.2514/6.2013-5519. DOI: 10.2514/6.2013-5519. [62] Page S E. The difference: how the power of diversity creates better groups, firms, schools, and societies–new edition[M]. New Jersey: Princeton University Press, 2008[2022-06-14]. https://doi.org/10.2514/6.2013-5529. DOI: 10.2514/6.2013-5519. [63] Heyel K. Data-driven technology foresight: text analysis of emerging technologies[D]. Seoul: Seoul National University, 2018. [64] DARPA and data: a portfolio overview[EB/OL]. [2022-08-13]. https://www.nitrd.gov/nitrdgroups/images/3/31/DARPA-and-DATA.pdf.
[65] Morstatter F, Galstyan A, Satyukov G, et al. SAGE: a hybrid geopolitical event forecasting system[C]//Proceedings of the 28th international joint conference on artificial intelligence. California: AAAI Press, 2019: 6557-6559. [66] Tetlock P E, Gardner D. Superforecasting: the art and science of prediction[M]. New York: Random House, 2016.

---

**Author Contributions:**
Chen Mo: Collected, organized, and analyzed relevant literature; wrote and revised the paper.
Li Guangjian: Proposed the research question and outline; wrote and revised the paper.

---

## Research on the Knowledge Fusion Technology System in Big Data Environment

**Chen Mo, Li Guangjian**
Department of Information Management, Peking University, Beijing 100871

**Abstract:** [Purpose/Significance] This paper mainly studies the key technologies of multi-source knowledge fusion in the big data environment, and proposes a complete set of technology taxonomy based on the characteristics of multi-source knowledge objects in different fields to provide technical support and solutions for the realization of knowledge fusion. [Method/Process] The study utilized qualitative analysis method to analyze the existing related objects, the application of abstraction of knowledge, the paper establishes the calculating layer, function layer and mission layer-three levels of technical architecture. These three layers contact each other, influence each other and interlock each other. The upper layer can be abstracted and associated with specific problems (tasks) of knowledge fusion. The lower layer can be embodied, that is, to find operational and computable technical methods to solve specific problems of knowledge fusion. [Result/Conclusion] Considering all kinds of technology's own characteristics, applicable knowledge type and implement various tasks involved in the work process and its specific technology, and form a knowledge fusion technology system. [Result/Conclusion] Considering the characteristics of various technologies, applicable knowledge objects, and application abstraction levels, a three-level technical architecture with computing layer, function layer, and task layer is established. These three levels are interconnected, influence each other, and are interlinked. They can be abstracted upward and associated with specific problems (tasks) of knowledge fusion; they can be concretized downward, that is, to find operable and computable technical methods to solve specific problems of knowledge fusion.

**Keywords:** knowledge fusion; technology system; big data