ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202202_5(2).0008

Research on Multi-source Data Fusion Technology for University Scientific Research Management

Shuqin Li

North China University of Technology, Beijing, 100144, China

Abstract

Based on the data of university scientific research management system, multi-source data fusion technology is adopted to collect scientific research personnel information from multiple data sources, analyze the data sources and data characteristics of various characteristics, and carry out multi-source data fusion from the dimensions of scientific research personnel information, standards, scientific research awards, scientific research projects, paper achievements and patents, etc. To achieve objective evaluation and intuitive display of scientific research personnel information, provide scientific reference for scientific research management decision-making, and improve the efficiency and level of research management.

Keywords

Scientific Research Management; Multi-source Data Fusion; Scientific Decision-making.

1. Introduction

Researchers are the main body of scientific and technological activities, and scientific researchers in universities carry out scientific research activities and undertake a large number of scientific research projects, and the relevant information presents dynamic, massive and multi-source heterogeneous characteristics, which will generate a large amount of heterogeneous scientific research data [1]. At present, the following problems exist in the description of research personnel information: (i) The data description is single, and the research personnel are described singularly in terms of academic achievements, ignoring other information; (ii) There is no objective evaluation and intuitive presentation of research personnel information [2]. For a long time, traditional statistical analysis methods have been used to process and organize the relevant data, which is not only time-consuming and laborious, but also unable to effectively explore the hidden research value behind the data, resulting in a huge waste of research resources.

To address the above-mentioned bottlenecks and problems in scientific research management, we take the data of university scientific research management system as the basis, and scientific researchers, papers, monographs, projects, awards, patents and standards as the entry point to collect and fuse multi-source heterogeneous data, break data silos, eliminate data conflicts, inconsistencies and ambiguities, improve data quality, and form a relevant structured database through data fusion and integration. The multi-source data fusion effectively collects, manages and analyses information on researchers, reveals their individual characteristics and preferences, and realizes objective evaluation and intuitive display of information on researchers, thus providing scientific reference for research management decisions and improving the efficiency and level of research management.

2. Multi-source Data Fusion Technology

Multi-source data refers to data that is generated by different users and sources, which has multiple presentation forms. Data fusion is a comprehensive data processing technology and analysis method that integrates, integrates, shares and correlates data from multiple sources to ensure consistency and relevance of information while forming accurate, unified and useful descriptions. Multi-source data fusion uses heterogeneous data from multiple sources to obtain comprehensive and collaborative inferences. The aim of data fusion is to reduce uncertainty in decision making by aggregating evidence from multiple sources of information, thereby improving the quality of the final decision [3].

2.1. Basic Principles

By imitating the process of processing complex information in the human brain, multi-source data fusion optimizes the processing of complementary and redundant information in space and time according to certain combination rules through effective screening of information, and obtains understanding, cognition and more valuable information about the monitoring target after multi-faceted and multi-level data processing, which can achieve the purpose of improving the comprehensive utilization of information and the effectiveness of the whole system. Multi-source data fusion technology is able to synthesize different types of information and data at multiple levels, processing objects that can be attributes, data, and so on.

2.2. Data Fusion Levels

Multi-sensor data fusion systems are divided into three levels according to the level of data abstraction: data-level fusion, feature-level fusion and decision-level fusion. Decision-level fusion is the highest level of fusion, the results of this layer to judge the state of the target decision, to provide a reasonable and effective basis for command and control decisions in real time. Decision-level fusion allows the formation of a preliminary judgement on the measurement target, followed by a decision-level fusion judgement to obtain a comprehensive judgement result, which is used as a three-level fusion result and directly affects the level of decision-making [4]. Decision-level fusion is therefore the process of presenting concise and intuitive results about a target, using specific decision needs as a starting point. The main disadvantages of decision-level fusion: the effect of data information processing is more dependent on the performance of the pre-processing stage; the main advantages: low processing costs at the fusion center, low data interaction, better resistance to interference and good fault tolerance. Decision-level fusion often uses methods such as: D-S evidence theory, fuzzy inference theory and expert systems.

2.3. The Main Methods of Multi-source Data Fusion

(1)Weighted average method. This method is often used for data fusion and is simple and intuitive. The drawback is that the weighting of the data in the algorithm is artificially subjective.

(2)Kalman filtering. This allows for real-time data fusion of targets and is more tried and tested in dynamic environments. The Kalman filter is well suited to complex state estimation and data fusion problems [5]. The algorithm can be used to obtain a unique optimal estimate under the condition that the system fits the dynamics model and that the error and system satisfy the Gaussian white noise model.

(3)Fuzzy logic inference. Due to the uncertainty of the fusion process, fuzzy logic is modelled by inference to produce consistent fuzzy inference, which to some extent solves the problem that statistical methods cannot handle. The disadvantage is that fuzzy logic inference does not form a systematic theory and has an obvious subjective element.

(4)Neural network method. Neural networks have many advantages, such as self-learning, self-adaptive, self-organizing and fault-tolerant, and are capable of modelling complex non-linear mappings. These advantages are useful for data fusion techniques in the processing process. Neural networks can assign network weights to develop classification criteria based on similarity, and can well coordinate multiple input information relationships, which is suitable for multi-source data fusion [6].

3. Concrete Implementation

The multi-source data fusion for university research management consists of data source acquisition, data warehouse establishment, data analysis and mining. The data source collection is for data acquisition, cleaning and data sampling; data matching and fusion is for name disambiguation to achieve data fusion; data analysis and mining is for in-depth analysis of data to draw meaningful conclusions.

3.1. Data Collection

The university research management system has accumulated a large amount of data on researchers, including data on research projects, research achievements and research awards. The data sources are used to obtain information on research personnel, including: personnel attributes, research results, research behavioral preferences, research collaboration data, research social data and other types of structured, semi-structured and unstructured data. Personnel attributes data refers to the basic attributes of researchers, including name, gender, age, title, education background and other information. Data on research results is the result of each researcher's hard work, revealing his or her research ability and academic influence in various disciplines, including journal papers, conference papers, academic monographs, patents and so on. The main sources of scientific research results data are CNKI, Wanfang, Wipu, Web of Science and EI databases. The National Natural Science Foundation of China website and the National Social Science Foundation of China project database have researcher projects; the National Science and Technology Reporting Service provides thematic reports, progress reports, final reports and organizational management reports of research activities. Data on scientific research cooperation refers to data jointly generated by scientific research members in cooperation with each other.

3.2. Data Fusion and Integration

Data fusion is high-level knowledge organization, through the process of heterogeneous data integration, disambiguation, processing, inference validation and updating of data from different data sources, to realize the linkage and merging of data and unify multiple sources of heterogeneous source data into standard structured data for easy use in user profiling. The first step is to formatively extract heterogeneous information, mainly for named entity recognition of information in text or images. The named entities in this paper are mainly name entities, such as names of people, organizations and places, and time expressions, such as dates and times. Next the data is cleaned, missing values are filled in by filling in null values or special values, and duplicate data is removed.

Data integration allows data from multiple data sources to be integrated together, solving the problem of data redundancy caused by different data sources having different names for the same attribute. According to the data types and sources of research personnel, the research results database, academic expertise database, research collaboration database and basic member attributes database are used to store these data respectively, and unified fields are used to describe the information of research personnel.

3.3. Establish A Data Warehouse

From multiple sources of data, the full amount of data and continuous incremental data are extracted and stored through the Hadoop big data warehouse to establish the full amount of raw data warehouse; the raw data warehouse is cleaned and standardized, and analyzed and adapted to form data analysis tables and stored in the warehouse, and the traditional relational database, including semi-structured data such as XML, as well as unstructured data in the form of video, audio, text and other Unstructured data such as XML and unstructured data in the form of video, audio, text and other forms are processed, the result sets are stored and the cleaning results are recorded to form a standardized database warehouse and finally a model analysis subject data warehouse is created through modelling and analysis. According to the data types of researchers, four basic databases are established, namely the basic attribute database, the research result information database, the research preference database and the research relationship database, and each database is interrelated.

3.4. Data Integration, Modelling and Analysis

After constructing a unified data warehouse, a variety of operational processes such as data parsing and storage, aggregation and analysis are implemented to realize time window statistics and online data mining and analysis, and finally construct a researcher portrait model. In intelligent data mining, the overall framework of analysis and computation is described in three layers, the data layer, the algorithmic model layer, and the usage layer [7]. Using machine learning algorithms, the researcher data warehouse is analyzed in depth. The data analysis and mining aims to build a researcher research assessment model, transform research data into knowledge through analysis, prediction and mining of research data, uncover research interests of researchers, analyze the frontiers of research interests, as well as build a researcher influence model.

4. Conclusion

With the rapid development of mobile internet, big data and artificial intelligence, the effective collection, management and analysis of researcher information will help to grasp the current situation of research and accurately describe the characteristics of researchers, which will in turn fundamentally change the management and decision-making mode of traditional science and technology work. As the collection of university research management data is obtained from multiple sources, the data has problems such as conflict, inconsistency and ambiguity, which require the fusion of information on multiple data sources.

In this paper, based on data from university research management systems, we collect researcher information from multiple data sources, generate usable data through data pre-processing, store the data in a research data warehouse, and analyze the data sources and data characteristics of various types of features. The fusion of data from multiple sources in the dimensions of research personnel information, standards, research awards, research projects, thesis results and patents realizes objective evaluation and intuitive display of research personnel information, provides scientific reference for research management decisions, and improves the efficiency and level of research management.

Acknowledgments

The research results of this thesis were funded by “North China University of Technology Yuyou Talent Support Program Project (20XN213/011)”.

References

[1] Fan Xiaoyu, Dou Yongxiang, Zhao Pengwei, et al. Study for the Construction Method of Scientist Profile with Multi-Source Data Fusion[J]. Library and Information Service, 2018, 62(15): 31-40.

[2] Ji Zhenyan, Wu Mengdan, Yang Chun, et al. Scalable Recommendation Models Fusing Multi-Source Heterogeneous Data[J]. Journal of Beijing University of Posts and Telecommunications, 2021, 44 (03): 106-111.

[3] He Yaqi. Research and Applications on the Key Technology of Multi-Source Heterogeneous Data Fusion [D]. University of Electronic Science and Technology of China,2018.

[4] Qi Youjie, Wang Qi. Review of Multi-Source Data Fusion algorithm[J]. Aerospace Electronic Warfare,2017,33(06):37-41.

[5] Ji Zhenyan,Pi Huaiyu,Yao Weina. A Hybrid Recommendation Model Based on Fusion of Multi-Source Heterogeneous Data [J]. Journal of Beijing University of Posts and Telecommunications, 2019, 42(01): 126-132.

[6] Di Curzio Diego et al. Multi-source data fusion of big spatial-temporal data in soil, geo-engineering and environmental studies.[J]. The Science of the total environment, 2021, 788 : 147842-147842.

[7] Jiang Yiqi et al. A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks[J]. Journal of Cleaner Production, 2021, 318.

[8] Quan Xunzhong, Chen Jie. Multi-Source Data Fusion and Target Tracking of Heterogeneous Network Based on Data Mining[J]. Traitement du Signal, 2021, 38(3) : 663-671.