International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 7827-7839

https://doi.org/10.70517/ijhsa464674

Scientometric mapping and algorithmic optimization of physical education research frontiers in China and the United States under big data environments

Liping Lang[1,*] and Xiao Ma[2]
[1]Chengdu sport University, Chengdu 610041 , Sichuan, China
[2]Beijing Tianrongxin Network Security Technology Co., Ltd. Chengdu 610041 , Sichuan, China

Corresponding authors: (e-mail: ls20150925@163.com).

Abstract With the rapid advancement of multimedia technology and big data, the landscape of physical education (PE) research has undergone significant transformation. However, there remains a lack of quantitative and visual comparative analysis of PE research frontiers between China and the United States. This study adopts CiteSpace software to analyze 946 English-language publications from the Web of Science and 232 Chinese-language publications from the CSSCI database, constructing knowledge maps and clustering co-cited references to reveal research hotspots and trends over the past five years. Results indicate that US PE research primarily emphasizes health-oriented pedagogical models, teacher professional development, and evidence-based practices, whereas Chinese research focuses more on curriculum reform, teaching modes, and educational policy alignment. To enhance knowledge processing efficiency, an improved genetic algorithm combined with rough set theory (IGA+RS) is proposed for knowledge abbreviation. The algorithm introduces heuristic information on attribute significance into the genetic search process, integrates deletion, repair, and smoothing operators, and applies niche evolution to avoid premature convergence. Experimental results demonstrate that IGA+RS significantly reduces redundancy in decision tables while preserving classification accuracy, outperforming traditional rough set methods.

Index Terms physical education (PE) research, scientometric analysis, citespace, big data analytics, improved genetic algorithm (IGA+RS), rough set knowledge reduction

I. Introduction

In the era of rapid technological advancement, multimedia technology and big data have penetrated into almost every domain of human activity, including education, health, and scientific research [1], [2]. The integration of information technology with education has significantly reshaped teaching methods, learning processes, and research paradigms. Physical education (PE), as an essential component of higher education, is no exception. In recent years, the modernization of PE has become a critical subject in educational reform, driven by increasing societal demands for holistic student development and health promotion. However, despite these advancements, the research field of PE still faces substantial challenges in terms of systematically identifying its evolving trends, mapping its research frontiers, and efficiently processing the growing volume of scientific literature [3].

PE plays a crucial role in fostering not only physical fitness but also social skills, cognitive development, and lifelong wellness habits. In both China and the United States, PE has undergone profound reforms aimed at enhancing teaching quality, promoting physical activity, and aligning curricula with broader educational goals [4], [5]. While the United States has emphasized health-based pedagogical models, teacher professionalization, and evidence-driven policies, China has focused on curriculum innovation, reform-oriented teaching practices, and the integration of traditional sports with modern education frameworks. Understanding the similarities and differences between these two leading countries is essential for advancing global PE development, facilitating knowledge exchange, and optimizing pedagogical strategies [6].

In the context of the big data era, the sheer scale of digital information generated in the field of PE research has increased exponentially. This vast volume of data holds valuable insights into research trends and emerging themes, yet it also poses challenges for researchers attempting to extract meaningful patterns. Traditional literature reviews are no longer sufficient to handle this information overload, underscoring the necessity for advanced quantitative methods, particularly scientometric analysis and data visualization techniques, to identify hotspots, track knowledge evolution, and support evidence-based

decision-making in education [7], [8].

Despite the growing importance of PE, several challenges persist in the current research landscape.First, there is a lack of comprehensive, data-driven comparative analysis between different countries, particularly between China and the United States, which represent two distinct educational and cultural contexts. Most existing studies focus on individual components such as curriculum design, teaching methods, or policy impacts, rather than presenting an integrated understanding of research frontiers and knowledge structures [9], [10]. Second, the identification of research trends and hotspots remains limited due to the reliance on traditional content analysis and manual reviews. Without advanced analytical tools, it is difficult to capture the dynamic nature of knowledge production and the complex relationships between research topics. Third, the efficient processing of large-scale scientometric data is hindered by algorithmic limitations. Traditional rough set (RS) methods, while useful for knowledge reduction and pattern discovery, often suffer from high computational costs and redundant information, leading to inefficiency in handling massive datasets. There is an urgent need for improved algorithms that can enhance the performance of rough set techniques while maintaining accuracy [11], [12].

Previous studies have explored various aspects of PE research from both domestic and international perspectives. In China, PE research has evolved over the past century, with early works focusing on teaching content and methods, followed by gradual incorporation of pedagogical theory and educational reforms. The implementation of the "new curriculum reform" has further encouraged innovative approaches, integrating multimedia technology and digital resources to improve student engagement and learning outcomes [13]. However, as noted by several scholars, most Chinese studies remain descriptive and lack quantitative scientometric analysis.

In contrast, the United States has developed a more diversified research landscape, with a strong emphasis on health-oriented PE models, teacher professional development, and evidence-based practices. Studies have highlighted the success of the US in embedding physical activity into broader educational and public health frameworks. Research by Curtner-Smith, Haerens, and others has emphasized the role of teachers, social learning, and cooperative teaching models in shaping PE outcomes. Scientometric techniques, such as co-citation analysis and knowledge mapping, have been increasingly employed in Western academia to track research evolution, yet comparative studies involving China are still scarce [14], [15].

Additionally, previous works on rough set theory and its applications in knowledge reduction have demonstrated its potential for managing complex datasets. However, traditional RS approaches have limitations in scalability and efficiency. Some researchers have introduced genetic algorithms (GAs) to optimize RS-based feature selection, but these methods often suffer from premature convergence and lack mechanisms to maintain population diversity, thus failing to achieve optimal results in high-dimensional scenarios [16].

Based on the literature, three main gaps can be identified: Insufficient comparative scientometric studies of PE research between China and the United States; Lack of integration between quantitative visualization methods and algorithmic optimization in the analysis of PE research data; Limitations in existing RS-based algorithms, which impede efficient knowledge processing and fail to handle large datasets effectively. These gaps highlight the necessity of developing an integrated framework that combines scientometric methods with advanced optimization algorithms to enhance both the accuracy and efficiency of PE research analysis.

To address the aforementioned challenges, this study proposes an integrated methodology that combines CiteSpace-based scientometric analysis with an improved genetic algorithm incorporating rough set theory (IGA+RS). Specifically: CiteSpace is utilized to construct co-citation networks, cluster references, and visualize knowledge structures, enabling the identification of research hotspots and trends in PE research across China and the United States.The IGA+RS algorithm is designed to enhance knowledge reduction by introducing heuristic information on attribute significance into the genetic search process, employing deletion, repair, and smoothing operators, and leveraging niche evolution strategies to avoid premature convergence. This approach improves both the computational efficiency and the accuracy of rough set knowledge reduction.

The major contributions of this paper are as follows:

A comprehensive scientometric comparison of PE research in China and the United States over the past five years, revealing differences in research focuses, trends, and knowledge evolution.

The development of an improved genetic algorithm (IGA+RS) that integrates rough set theory with heuristic attribute significance and advanced genetic operators, effectively reducing redundancy while maintaining classification accuracy.

The integration of big data analytics, information visualization, and algorithmic optimization into PE research, providing both theoretical insights and practical tools for scholars and policymakers.

II. Rough sets overview

II. A. Principle of genetic Algorithm

The flow chart of the basic principle of the genetic algorithm is shown in Figure 1.

The genetic algorithm (GA) is a population-based optimization technique inspired by the principles of biological evolution. It operates under the fundamental idea of "survival of the fittest", mimicking natural selection mechanisms to iteratively evolve better solutions to complex problems.

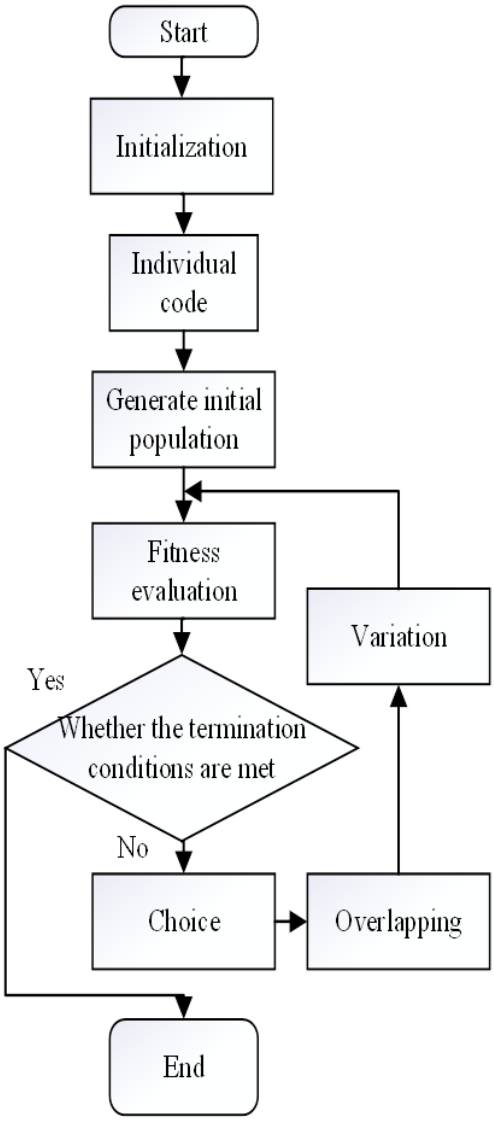

Figure 1: The fundamental flow diagram for the genetic algorithm concept

The algorithm begins by generating an initial population composed of randomly constructed candidate solutions. Each candidate, referred to as an *individual*, is assessed using a predefined fitness function that reflects its quality or effectiveness in solving the target problem.

Individuals with higher fitness scores are more likely to be selected for reproduction. During reproduction, genetic operators such as crossover (recombination of features between two parents) and mutation (random alteration of traits) are applied to produce new offspring. This process introduces both exploitation (preserving good traits) and exploration (introducing diversity) into the population [17].

Through repeated generations of selection, recombination, and mutation, the population gradually evolves toward regions of the solution space with higher overall fitness. While the genetic algorithm does not ensure discovery of the global optimum, it effectively guides the search process by discarding low-fitness individuals and favoring the propagation of advantageous traits [18], [19].

The strength of GA lies in its ability to perform global search without the need for gradient information or strict assumptions about the problem landscape, making it suitable for solving nonlinear, multidimensional, and multimodal optimization

problems.

II. A. 1) Initialization of the population and modeling of the environment

To assess the performance of genetic algorithms, two prerequisites must be met: population initialization and environmental modeling. Effective environmental modeling can effectively assess the relationship between the choice of rural revitalization path and actual development, as well as accurately depict the actual environment of rural industrial revitalization. Both the more diverse members of the original population and the optimization of the rural industrial revitalization path benefit from the relationship [20].

II. A. 2) Environment Modeling

A well-defined environmental model serves as the foundational layer for simulation, perception, and decision-making tasks in intelligent systems. In this study, the environment is abstracted as a structured two-dimensional grid, allowing for straightforward computation and spatial reasoning.

Let the environment $\mathcal{E} \subset \mathbb{R}^2$ be discretized into a uniform grid structure defined by:

$$\mathcal{G} = \{(x_i, y_j)x_i = i \cdot \Delta x, y_j = j \cdot \Delta y; i, j \in \mathbb{Z}, \tag{1}$$

where, $\Delta x, \Delta y$ represent the grid resolution in the x- and y-directions respectively, (x_i, y_j) denotes the center of the grid cell at row i and column j.

Each grid cell $g_{ij} \in \mathcal{G}$ can be assigned a state variable $s_{ij} \in \{0, 1, \ldots, k\}$ to indicate its status (e.g., obstacle, free space, target area). Thus, the environment can be represented as a state matrix:

$$S = [s_{ij}]_{m \times n}, \tag{2}$$

where mmm and n are the total number of grid points in the vertical and horizontal directions, respectively.

This modeling strategy enables:

1) Deterministic representation of environmental elements,
2) Efficient collision detection via grid state inspection,
3) Modular extension to dynamic environments or multi-agent interaction.

Although more complex methods such as topological maps or multifaceted semantic representations exist, the use of a 2D grid in this work provides sufficient structure for trajectory planning and algorithm validation, while ensuring computational tractability.

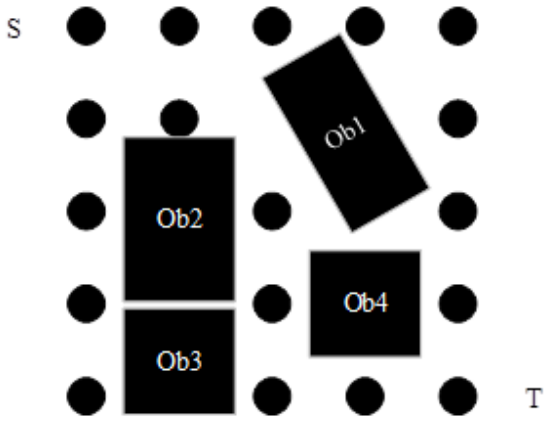

Figure 2: Working environment

The path obtained based on the improved genetic algorithm under multiple constraints is represented as:

$$P^* = \{p_1^*, p_2^*, p_3^*, ..., p_m^*\}, \ \ P^* \in P. \tag{3}$$

In the formula, m represents the number of paths obtained by the algorithm in this paper, and the path needs to satisfy

$$f(P^*) = \min\{f_1(P), f_2(P), f_3(P),\}, \tag{4}$$

where, $f_1(P), f_2(P), f_3(P)$ is defined by Eqs. (4), (5) and (6), respectively

Definition 1. The length index refers to the total length of the path, which can be expressed as:

$$f_1(P) = \sum_{i=1}^{n-1} |p_i p_{i+1}|, \tag{5}$$

where $|p_i p_{i+1}|$ is the Euclidean distance from the path node p_i to the path nodep_{i+1} . Finding the shortest path is one of the optimization objectives of the genetic algorithm under multiple constraints. The other two optimization objectives are defined as follows:

Definition 2. The sum of the angles of each neighboring vector line segment in the path is known as the smoothness index, and it may be written as follows:

$$f_2(P) = C_1 \times S + \frac{1}{N_I} \sum_{i=2}^{N_i-1} \theta, \theta(p_i p_{i+1}, p_{i+1} p_{i+2}), \tag{6}$$

where $\theta(p_i p_{i+1}, p_{i+1} p_{i+2})$ is the angle between adjacent vector segments $\xrightarrow[p_i p_{i+1}]{}$ and $\xrightarrow[p_{i+1} p_{i+2}]{}$ $(0 \leq \theta \leq \pi)$, C_1 is a positive integer, s is the number of segments in a path, and N_i is the number of points in the i th iteration path.

II. A. 3) Population initialization

The population initialization method adopted by the traditional genetic algorithm is random. Many studies generate points by randomly scattering them in free space or by considering all points in a grid map. These methods have to consider unnecessary points for the optimal path in the path generation stage, which is computationally expensive. To this end, some scholars have proposed some improved methods.

The specific steps of SPS algorithm to generate the initial population are as follows.

Step 1: In the two-dimensional grid point map, set the coordinates of the starting point and the target point.

Step 2: Rural industry revitalization moves towards the target point along a straight line. If the development is hindered, the SPS algorithm is used to generate a set of points around it.

Step 3: Use Dijkstra's algorithm to test whether the generated path points are feasible, if feasible, move on; if not, reduce the grid point width and repeat Step 2.

Step 4: Determine whether the target point is reached. If it has been reached, save the currently generated path. If not, repeat Step 2 and Step 3.

Step 5: Judge that the number of individuals reaches the population number of 2M, if it is reached, it will end; if not, it will be restarted. Repeat the above steps.

II. B. Chromosome coding and fitness function

The chromosome code and fitness function are the core elements that determine the performance of the genetic algorithm. A good chromosome code and fitness function can reduce the complexity of the performance of the genetic algorithm, and can select individuals with better fitness to inherit it to the next generation. It helps to further improve the performance of the algorithm [21].

II. B. 1) Chromosome coding

In this study, the spatial domain of rural industry revitalization is modeled using a uniformly distributed two-dimensional grid space, which provides a structured and discrete environment for planning, analysis, and simulation. Each spatial element corresponds to a grid point with unique coordinates, making it inherently suitable for two-dimensional encoding and algorithmic manipulation.

Let the environment be represented by a grid point set:

$$= \{(x_i, y_j) x_i = i \cdot \Delta x, y_j = j \cdot \Delta y; i = 0, 1, \ldots, m, j = 0, 1, \ldots, n\}. \tag{7}$$

Each point $p_{ij} = (x_i, y_j) \in$ can be uniquely encoded using a two-dimensional indexing function:

$$Code(p_{ij}) = (i, j). \tag{8}$$

Alternatively, for linear storage or identification, a flattened index can be defined as:

$$Code_{1D}(p_{ij}) = i \cdot n + j. \tag{9}$$

This encoding enables efficient mapping between spatial locations and data structures (e.g., matrices, arrays, or graphs), and is particularly useful in algorithms such as path planning, resource allocation, or regional optimization.

Each grid point can also be associated with a feature vector:

$$f_{ij} = \left[r_{ij}^{(1)}, r_{ij}^{(2)}, \ldots, r_{ij}^{(k)} \right], \tag{10}$$

where f_{ij} denotes the l-th socioeconomic or geographic attribute at location (x_i, y_j) , such as population density, land use type, industrial output, or accessibility index.

By discretizing the spatial domain in this way, the rural revitalization process can be modeled as a function $\mathcal{R} : \ \rightarrow \mathbb{R}^k$, mapping spatial positions to development attributes, thereby enabling spatial decision-making and policy optimization over the entire region. Figure **??** is an example of path encoding.

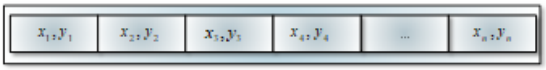

Figure 3: Chromosome representation of feasible paths

II. B. 2) Fitness function

In the context of genetic algorithm (GA) optimization, the fitness function serves as the core mechanism for evaluating and guiding the evolutionary process. It quantifies the quality of each individual I_i in the population , and is derived from one or more objective functions that reflect specific evaluation criteria.

1. Multi-Criteria Evaluation Structure

Let each individual $I_i \in$ correspond to a feasible solution (e.g., a candidate path), and define a multi-objective function vector:

$$\vec{F}(I_i) = [f_1(I_i), f_2(I_i), f_3(I_i)], \tag{11}$$

where, $f_1(I_i)$: Path length (to be minimized), $f_2(I_i)$: Path safety score (to be maximized), $f_3(I_i)$: Path smoothness (to be minimized).

To reflect the relative importance of criteria:

1) f_1 and f_2 are treated as primary objectives,
2) f_3 is a secondary objective, incorporated when solutions are otherwise comparable.

2. SPEA2-Based Fitness Assignment

This study adopts the Strength Pareto Evolutionary Algorithm 2 (SPEA2) framework for multi-objective fitness evaluation. For each individual IiI_iIi, the raw fitness is computed based on its dominance relation:

$$R(I_i) = \sum_{I_j \succ I_i} S(I_j), \tag{12}$$

where, $I_j \succ I_i$ indicates that individual I_j dominates I_i in objective space (i.e., better in all objectives), $S\,(Ij)$ is the strength of I_j , defined as the number of individuals it dominates:

$$S(I_j) = |\{I_k \in I_j \succ I_k\}|. \tag{13}$$

Thus, the total fitness of I_i is composed of both dominance count and density estimation. A density function $D\,(I_i)$ is calculated as:

$$D(I_i) = \frac{1}{\sigma_k(I_i) + 2}, \tag{14}$$

where $\sigma_k(I_i)$ is the distance between I_i and its k-th nearest neighbor in objective space. The final fitness is given by:

$$Fitness(I_i) = R(I_i) + D(I_i). \tag{15}$$

Lower fitness values indicate better individuals under this scheme, consistent with minimization.

II. B. 3) Genetic operators

To address the limitations of traditional genetic algorithms (GAs) in complex path planning tasks, this study introduces a set of enhanced genetic operators aimed at accelerating convergence and improving solution quality. Building upon the classical framework—which comprises selection, crossover, and mutation—three additional operators are incorporated: deletion, repair, and smoothing. Together, these extensions enable the algorithm to better handle constraints and produce feasible, high-quality paths in rural spatial environments [22], [23].

1. Selection Strategy

Instead of using proportional or roulette wheel selection, this paper employs a tournament-based selection mechanism. In each round, a fixed number of individuals are randomly sampled from the population, and the one with the highest fitness is selected to propagate to the next generation. This process is repeated until the desired population size is achieved. Furthermore, an elitism mechanism is applied: a specified fraction of top-performing individuals are directly carried over to the offspring population, thereby preserving high-quality solutions across generations without modification.

2. Crossover Mechanism

A single-point crossover approach is adopted with domain-specific refinements. The method operates as follows:

1) Two parent individuals are randomly chosen.
2) If their respective paths share common waypoints, one such point is selected as the crossover node, ensuring that the resulting offspring inherit a structurally valid and continuous path.
3) In the absence of shared waypoints, two random positions from each path are used as crossover points. If this leads to discontinuity, a corrective step is invoked:
   a) The algorithm identifies the endpoints of the segmented paths.
   b) A connection patch is constructed using adjacent nodes from the grid space, avoiding obstacles and maintaining path validity.

This strategy guarantees that all offspring paths remain navigable, even after structural modifications during crossover.

3. Additional Operators for Enhanced Performance

To further refine the evolutionary process and reduce unnecessary computational cost, the algorithm integrates the following auxiliary operators:

1) Deletion Operator: Removes redundant or cyclic path segments to simplify the overall trajectory and reduce length.
2) Repair Operator: Automatically resolves infeasible segments caused by crossover or mutation, ensuring connectivity and constraint satisfaction.
3) Smoothing Operator: Optimizes the geometry of the path by eliminating abrupt directional changes, promoting natural and realistic movement patterns.

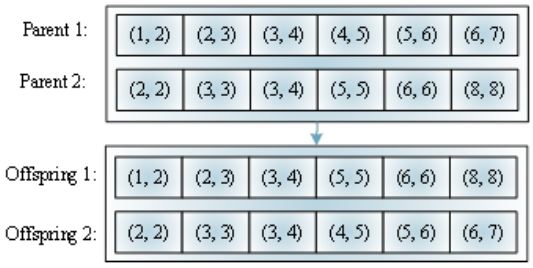

Figure 4: Crossover Process

To improve adaptability and efficiency in path evolution, this study incorporates refined strategies for both mutation and deletion, tailored to the feasibility status of individuals during the genetic process.

1. Adaptive Mutation Strategy

The mutation operator dynamically adjusts its behavior based on the validity of the path. Specifically:

1) For feasible solutions—those that comply with spatial constraints and obstacle avoidance—the mutation is applied conservatively, with a low probability. Local perturbations are introduced by slightly shifting selected waypoints within a limited neighborhood, maintaining path integrity while exploring nearby alternatives for optimization.
2) For infeasible solutions—particularly those intersecting with obstacles—the mutation is executed with a higher probability, allowing for broader adjustments to problematic nodes. This increases the likelihood of escaping constraint violations and moving toward valid regions in the search space.

Formally, the mutation probability P_{mut} is adaptively defined as:

$$P_{mut}(I_i) = \begin{cases} \varepsilon_1, & if I_i \in \mathcal{F}_{valid} \\ \varepsilon_2, & if I_i \in \mathcal{F}_{invalid}, \quad \varepsilon_2 > \varepsilon_1 \end{cases} \tag{16}$$

where $\mathcal{F}_{valid}$ and $\mathcal{F}_{invalid}$ represent the feasible and infeasible individual sets, and $\varepsilon_1, \varepsilon_2 \in (0, 1)$ are user-defined constants.

2. Deletion-Based Path Simplification

To address redundancy and reduce convergence time, a deletion operator is introduced as a structural optimization tool. Without this mechanism, evolved paths often retain unnecessary segments—such as zigzags or local loops—that require multiple generations to resolve through crossover and mutation alone.

The deletion operator systematically scans the path for non-contributing waypoints, particularly those forming nearly collinear triplets or cycles. When identified, these nodes are pruned, thereby:

1) Shortening the path length
2) Improving smoothness
3) Accelerating convergence

This operation is especially effective in eliminating the scenario illustrated in Figure 5(a), where excessive bends hinder optimization and lead to prolonged evolution cycles.

Together, the adaptive mutation and strategic deletion mechanisms significantly enhance the algorithm's ability to maintain feasible, efficient, and navigable paths throughout the evolutionary process.

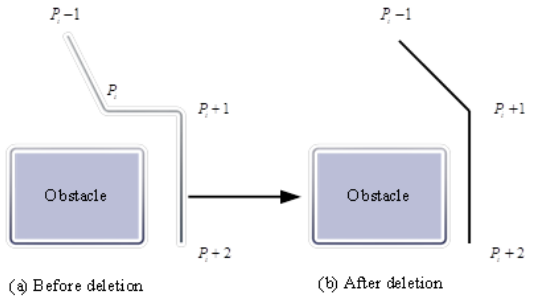

Figure 5: Removal process

The points surrounding the created obstruction can be connected sequentially by the repair operator if a path segment connects with an obstacle, as seen in Figure 6(a) (Figure 6(b)).

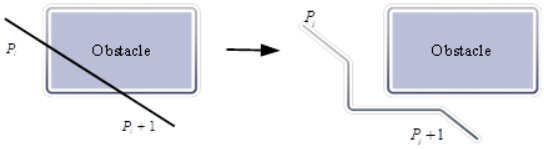

Figure 6: Repair process

To further refine the path structure and avoid premature convergence during the evolutionary process, this study introduces two key mechanisms: a smoothing operator inspired by particle swarm dynamics, and a niche-based parallel evolution strategy for maintaining population diversity.

1. Smoothing Operator Based on Particle Dynamics

The smoothing process treats each waypoint in a candidate path as an analogue to a particle in Particle Swarm Optimization (PSO). By leveraging the particle movement model, the position of each point is iteratively adjusted based on its local neighborhood.

Let a waypoint $P_t \in \mathbb{R}^2$ be located at step t in the path. The velocity vector v_t for this point is computed based on its adjacent points P_{t-1} *and* P_{t+1} . The update rules follow a modified PSO model:

$$v_t^{(k+1)} = \omega \cdot v_t^{(k)} + c_1 \cdot r_1 \cdot (P_{t-1} - P_t) + c_2 \cdot r_2 \cdot (P_{t+1} - P_t), \tag{17}$$

$$P_t^{(k+1)} = P_t^{(k)} + v_t^{(k+1)}, \tag{18}$$

where, ω is the inertia weight, determining how much of the previous velocity is retained, c_1, c_2 are acceleration coefficients controlling the influence of neighbor positions, $r_1, r_2 \in [0, 1]$ are random scalars introducing stochasticity.

This dynamic adjustment encourages the path to evolve into a smoother and more natural trajectory over multiple iterations, especially useful in reducing sharp turns or redundant detours.

The smoothing process is illustrated in Figure 7.

2. Niche Strategy for Parallel Evolution and Diversity Preservation

To mitigate issues such as premature convergence and stagnation in local optima, this study implements a niche-based population structure, effectively simulating parallel sub-evolutions.

The total population is divided into N niches, where each niche contains two individuals with high structural similarity. Similarity is quantified using the Hamming distance between encoded individuals:

$$Hamming(I_i, I_j) = \sum_{k=1}^{L} \delta(I_i^k, I_j^k), \quad \delta(a, b) = \begin{cases} 0 & if a = b, \\ 1 & if a \neq b. \end{cases} \tag{19}$$

A smaller Hamming distance indicates higher similarity, making individuals suitable for niche pairing.

Each niche undergoes independent reproduction and selection, producing new offspring in parallel. After evolution within niches, top-performing individuals from each niche are selected to form the next global generation, ensuring both:

1) Population diversity is retained,
2) Algorithmic parallelism is achieved, improving convergence efficiency.

This approach enhances the exploration capability of the genetic algorithm while maintaining solution quality and accelerating runtime.

II. B. 4) Termination Conditions

In order to enable the improved genetic algorithm in this paper to find the optimal or sub-optimal path in a short time, this paper sets three termination conditions: first, the optimal individual fitness value obtained after multiple iterations satisfies the preset value The algorithm can be terminated when the threshold is reached; second, the overall fitness value of the population does not change much after many iterations, and the algorithm can be terminated; third, when the number of iterations of the algorithm reaches the preset algebra, the algorithm It can be terminated. This paper sets the iteration for 100 times.

II. C. Implementation of the IGA+RS Algorithm

In this study, a hybrid optimization framework named IGA+RS (Improved Genetic Algorithm with Rough Set heuristics) is developed. The key innovation lies in integrating attribute dependency measures—specifically, the *support* and *significance* of condition attributes with respect to decision attributes—into the genetic algorithm to provide heuristic guidance during evolution.

The overall procedural framework of the IGA+RS algorithm is outlined in Figure 1.

1. Integration of Rough Set Heuristics

To enhance the search efficiency and guide the population evolution toward more meaningful solutions, rough set theory is utilized to compute two key indicators for each attribute$x \in C$, where C denotes the set of condition attributes and D the set of decision attributes:

1) Support is computed based on Eq. (19), quantifying how strongly x contributes to distinguishing decision classes.
2) Importance is evaluated using the MMPC12 criterion (Eq. (20)), which reflects the degree to which removing x affects classification performance.

These metrics act as fitness modifiers within the genetic algorithm, influencing selection and crossover in a goal-directed manner.

2. Attribute Core Judgment and Termination Criterion

Once the support and importance values are obtained, the algorithm checks whether the core condition set can be identified:

Let

$$core(C) = \{x \in CSupport(x) > \tau_1 \wedge MMPC12(x) > \tau_2\}. \tag{20}$$

If the core attribute set satisfies the sufficiency condition for decision classification, i.e., it constitutes a minimal reduct of C with respect to D, then the search is terminated and the current attribute subset is considered optimal.

Otherwise, the algorithm proceeds to the IGA stage, where the improved genetic algorithm is activated to further refine the attribute selection and search for a globally optimal reduction.

The detailed evolutionary process, incorporating crossover, mutation, and the extended operator set, is illustrated in Figure 7.

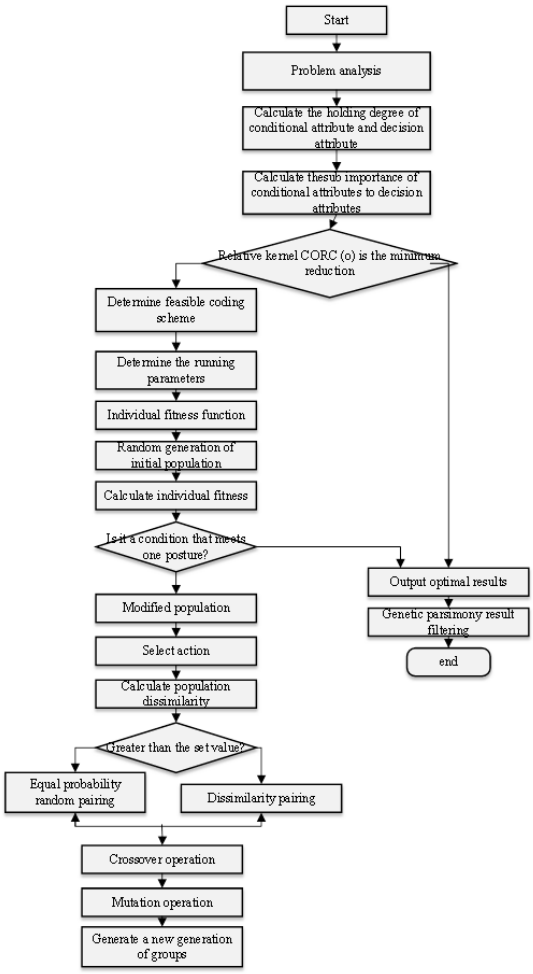

Figure 7: Flow chart for the IGA+RS algorithm

III. Visual analysis and comparative study on the research frontiers of PE teaching in China and the United States

III. A. Visualization of research frontiers in U.S. physical education over the past five years

To explore the knowledge structure and evolving trends in the field of physical education (PE) research in the United States, this study conducts a scientometric analysis using CiteSpace V (version 5.1.R8). A total of 946 scholarly articles published between 2012 and 2016 were selected as the data source. The parameter configuration includes:

Time Slicing: 2012–2016 with 1 year per slice

Top N: 50 (selecting the top 50 most cited items per slice)

Visualization Settings: Cluster View – Static; Show Merged Network enabled

Thresholds: set as (2,2,20), (4,3,20), and (4,3,20) for different pruning schemes

Upon executing the analysis, a document co-citation network was generated, comprising 256 nodes (representing cited references) and 699 edges (indicating co-citation links). This network was further processed through clustering algorithms and automatic cluster labeling based on article titles, resulting in a detailed knowledge domain visualization, as illustrated in Figure 8.

The clustering process revealed a total of 39 thematic groups, representing various subfields and trending topics within U.S. PE research. Among these, the eight most significant clusters—based on their modularity value and spatial size—were identified as clusters #0, #1, #2, #3, #4, #5, #9, and #10. These prominent clusters are summarized in Table 1 and collectively represent the core intellectual frontiers shaping recent discourse and innovation in American physical education research.

Each cluster highlights specific focus areas, such as curriculum development, physical literacy, health promotion, and pedagogical reform, offering a comprehensive overview of current research priorities and directions in the field.

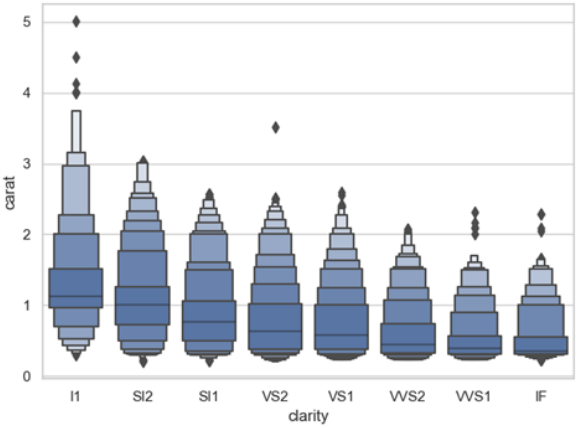

Figure 8: Using document co-citation analysis, clustering mapping the knowledge domains of the PE Teaching Study of America research front

Table 1: Cluster information of co-cited references in research fronts of PE teaching studies in the United States (2012–2016)

Cluster ID	Cluster Label (Title Term)	Size (No. of Nodes)	Silhouette Value	Mean Year	Research Focus
#0	Physical Literacy	42	0.923	2014	Core competencies and physical literacy in curriculum frameworks
#1	School-Based Intervention	37	0.887	2013	Effectiveness of school-based physical activity interventions
#2	PE Teacher Education	31	0.915	2015	Pre-service training and professional identity of PE teachers
#3	Physical Activity Promotion	29	0.901	2013	Community and institutional strategies to promote youth activity
#4	Inclusive Education	27	0.876	2014	Equity and inclusion for students with disabilities in PE settings
#5	Health-Oriented Curriculum	24	0.892	2015	Integration of health education in PE curriculum design
#9	Motor Skill Development	21	0.867	2012	Early childhood movement skill acquisition and assessment
#10	Technology Integration	18	0.854	2016	Use of digital tools and wearable devices in PE instruction

As shown in Figure 2, the cluster names of 11 clusters that reflect the main research frontiers in the field of PE teaching in the United States are: "#0: models-based practice", "#1: physical activity program", "#2: adventure-PE lesson", "#3: quantitative findings", "#4: prospective cross-domain investigation" cross-cutting survey)", “#5: initial teacher education”, “#9: teachers support” and “#10: self-reported achievement goal”.

The main research front area 1: 39 cited literatures of the knowledge base contained in cluster #0, the year span is 2006-2015, among which there are 8 main literatures in the literature represented by the node, as shown in Table 2, including PE teachers The professional development, the PE model, the reform of PE, the method of PE, and the cooperative learning in sports [24]. There are 19 citing documents corresponding to the documents, and the year span is 2012-2016. Among them, there are 11 documents with a citation activity of ≥ 0.05, as shown in Table 3, and the citation activity of the remaining documents is 0.03. From the titles of these 11 papers, it can be seen that the research topics of this cluster mainly focus on the practice-based PE teaching mode, the role of PE teachers, and the PE teaching mode, method and content. As a result of class naming, the research front of this cluster is dominated by practice-based models of PE, and also includes the role of PE teachers in education and public health, the role of PE in health and education, the effect of social learning on girls’ participation in sports, and the models of PE and methods [25].

Table 2: Citing documents with citation activity ≥ 0.05 in cluster #0 (Practice-based PE teaching models)

No.	Title of Citing Document	First Author	Year	Citation Activity	Research Focus
1	Practice-Based Professional Development for Physical Education Teachers	Parker, M.	2013	0.08	Teacher professional growth and practice-based learning
2	The Role of Physical Education in Public Health	McKenzie, T.L.	2014	0.07	PE's contribution to student health and behavior
3	Models-Based Practice in Physical Education: A Review of Literature	Casey, A.	2015	0.06	Pedagogical models and their classroom effectiveness
4	Cooperative Learning in Physical Education and Physical Activity	Dyson, B.	2012	0.05	Social learning and inclusion in sports education
5	Physical Education for the 21st Century: Toward a Practice-Oriented Curriculum	Kirk, D.	2014	0.06	Curriculum reform and practical teaching frameworks
6	Exploring Girls' Participation in Sport and PE: A Social Learning Perspective	Smith, J.	2013	0.05	Gender inclusion and motivation in PE
7	Implementing Models-Based Practice in Secondary Schools	Ward, G.	2015	0.05	Application of PE models in real classroom settings
8	Pedagogical Approaches in Physical Education: A Shift Toward Learner-Centered Methods	Harvey, S.	2013	0.05	Transition to student-focused instructional practices
9	Physical Education Teachers and Health Promotion Responsibilities	Jones, R.	2014	0.06	Expanding the role of PE educators into health and wellness promotion
10	The Impact of Model-Based Instruction on Students' Physical Activity	Johnson, K.	2015	0.05	Outcome-based evaluation of teaching models

III. B. Algorithm simulation and result analysis

Table 3: The information function $\int$ is shown in the decision table of Table 1

Object ID	A1 (Teaching Method)	A2 (Teacher Experience)	A3 (Class Size)	A4 (Student Participation)	Decision D (Teaching Effectiveness)
1	Model-Based	High	Small	High	Effective
2	Traditional	Low	Large	Low	Ineffective
3	Cooperative	Medium	Small	High	Effective
4	Traditional	Medium	Medium	Low	Ineffective
5	Model-Based	High	Medium	Medium	Effective
6	Cooperative	Low	Large	Medium	Ineffective
7	Traditional	High	Medium	Low	Ineffective
8	Model-Based	Medium	Small	High	Effective

IV. Conclusion

This study presented a comprehensive comparison of physical education research frontiers in China and the United States by integrating scientometric visualization and algorithmic optimization. CiteSpace-based analysis revealed distinct research focuses: while the United States advances health-oriented, evidence-based, and teacher-centered pedagogical innovations, China emphasizes curriculum reform and policy-driven teaching models. These differences reflect varied educational priorities and sociocultural contexts. To address the challenge of processing large-scale scientometric data, the proposed IGA+RS algorithm effectively improved the efficiency of rough set knowledge reduction by combining heuristic attribute significance with enhanced genetic operations. Experimental validation confirmed its superiority in maintaining accuracy while reducing redundancy. The findings highlight the potential of combining big data analytics with intelligent algorithms to uncover research dynamics and optimize knowledge management in education. Furthermore, this study provides a theoretical and methodological reference for future PE research, offering a framework to monitor evolving trends, identify collaboration opportunities, and guide evidence-based decision-making.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

Funding Statement

This paper has no fund support.

References

[1] Kuang, Y., Zhu, Y., & Niu, Q. (2024). Visual analysis and comparison of frontiers and hotspots of physical education teaching research in china and america based on improved genetic algorithm rough sets. *J. COMBIN. MATH. COMBIN. COMPUT*, *123*(75), 85.

[2] Sun, F. (2024). Frontiers and hotspots of high-intensity interval exercise in children and adolescents: text mining and knowledge domain visualization. *Frontiers in Physiology*, *15*, 1330578.

[3] Wang, Q., Li, R., & Zhan, L. (2021). Blockchain technology in the energy sector: From basic research to real world applications. Computer Science Review, 39, 100362.

[4] Li, F., Zhang, H., Chen, B., Zhou, Q., & Li, X. (2024, August). Visualization of Hotspots and Frontiers in Learning Analytics under Big Data Environment—Based on Citespace Knowledge Map Analysis. In *Proceedings of the 2024 8th International Conference on Digital Technology in Education (ICDTE)* (pp. 221-231).

[5] Yang, R., Yuan, Q., Zhang, W., Cai, H., & Wu, Y. (2024). Application of Artificial Intelligence in rehabilitation science: A scientometric investigation Utilizing Citespace. *SLAS technology*, *29*(4), 100162.

[6] Zhao, B., & Liu, S. (2021). Basketball shooting technology based on acceleration sensor fusion motion capture technology. EURASIP Journal on Advances in Signal Processing, 2021(1), 1-14.

[7] Lyu, T., Wang, P. S., Gao, Y., & Wang, Y. (2021). Research on the big data of traditional taxi and online car-hailing: A systematic review. Journal of Traffic and Transportation Engineering (English Edition), 8(1), 1-34.

[8] Xu, Y., Li, W., Tai, J., & Zhang, C. (2022). A Bibliometric-Based Analytical Framework for the Study of Smart City Lifeforms in China. International Journal of Environmental Research and Public Health, 19(22), 14762.

[9] Jiang, J., Guo, Y., Bi, Z., Huang, Z., Yu, G., & Wang, J. (2022). Segmentation of prostate ultrasound images: the state of the art and the future directions of segmentation algorithms. Artificial Intelligence Review, 1-37.

[10] Zhu, X., & Peng, X. (2024). Strategic assessment model of smart stadiums based on genetic algorithms and literature visualization analysis: A case study from Chengdu, China. *Heliyon*, *10*(11).

[11] Wang, Q., Li, R., & Zhan, L. (2021). Blockchain technology in the energy sector: From basic research to real world applications. Computer Science Review, 39, 100362.

[12] Ma, L., Wang, Y., Wang, Y., Li, N., Fung, S. F., Zhang, L., & Zheng, Q. (2021). The Hotspots of Sports Science and the Effects of Knowledge Network on Scientific Performance Based on Bibliometrics and Social Network Analysis. Complexity, 2021.

[13] Zhang, M., & Meng, X. (2025). School built environment and children's health: a scientometric analysis. *Reviews on Environmental Health*, *40*(2), 465-480.

[14] Song, D., Meng, W., Dong, M., Yang, J., Wang, J., Chen, X., & Huang, L. (2022). A critical survey of integrated energy system: Summaries, methodologies and analysis. Energy Conversion and Management, 266, 115863.

[15] Cheng, C., Xue, J., Gou, W., & Xie, M. (2025). Quantitative analysis and evaluation of research on the application of computer vision in sports since the 21st century. *Frontiers in Sports and Active Living*, *7*, 1604232.

[16] Yan, L., & Du, Y. (2025). Exploring Trends and Clusters in Human Posture Recognition Research: An Analysis Using CiteSpace. *Sensors*, *25*(3), 632.

[17] Cheng, S., Zhang, J., Wang, G., Zhou, Z., Du, J., Wang, L., ... & Wang, J. (2024). Cartography and neural networks: a scientometric analysis based on CiteSpace. *ISPRS International Journal of Geo-Information*, *13*(6), 178.

[18] Shen, S., Qi, W., Li, S., Zeng, J., Liu, X., Zhu, X., ... & Cao, S. (2025). Mapping the landscape of machine learning in chronic disease management: A comprehensive bibliometric study. *Digital health*, *11*, 20552076251361614.

[19] Gómez-Domínguez, V., Navarro-Mateu, D., Prado-Gascó, V. J., & Gómez-Domínguez, T. (2022). How Much Do We Care about Teacher Burnout during the Pandemic: A Bibliometric Review. International Journal of Environmental Research and Public Health, 19(12), 7134.

[20] Liu, L., Guo, F., Zou, Z., & Duffy, V. G. (2022). Application, Development and Future Opportunities of Collaborative Robots (Cobots) in Manufacturing: A Literature Review. International Journal of Human–Computer Interaction, 1-18.

[21] Xiong, Y., Liu, T., Qin, Y., & Chen, H. (2024). A scientometric examination on performance-driven optimization in urban block design research: State of the Art and future perspectives. *Buildings*, *14*(2), 403.

[22] Liu, M., Fang, M., Liu, M., Jin, S., Liu, B., Wu, L., & Li, Z. (2024). Knowledge mapping and research trends of brain-computer interface technology in rehabilitation: a bibliometric analysis. *Frontiers in Human Neuroscience*, *18*, 1486167.

[23] Zhang, X., & Wu, Y. (2025). A bibliometric analysis and science mapping of home telemonitoring for older adults. *Educational Gerontology*, 1-19.

[24] Panackal, M. B., Mathew, P., Sunny, S., Joseph, J., & Jose, J. (2025). Mapping Research on mHealth and Wearable Technologies in Sports and Gaming: A Bibliometric and Visualization Approach (2005–2025). *Informatica*, *49*(23).

[25] Yanmin Xu, Yitao Tao, Chunjiong Zhang, Mingxing Xie, Wengang Li, Jianjiang Tai, "Review of Digital Economy Research in China: A Framework Analysis Based on Bibliometrics", Computational Intelligence and Neuroscience, vol. 2022, Article ID 2427034, 11 pages, 2022. https://doi.org/10.1155/2022/2427034