Research Team Identification and Leading Team Extraction in Artificial Intelligence (Postprint)

Authors: Yu Houqiang, Bai Kuan, Zou Bentao, Wang Yuefen

Date: 2023-04-01T16:16:02+00:00

Abstract

[Purpose/Significance] This study aims to identify research teams in the field of artificial intelligence and extract leading research teams based on multi-dimensional indicators, thereby enriching the processes and methods of research team identification and providing a foundation for analyzing the context, frontiers, and themes of the artificial intelligence domain from a research team perspective.

[Method/Process] Using Web of Science as the data source, we collected data on all scientific papers in the artificial intelligence discipline from 2009 to 2018, and performed data cleaning through algorithmic design and manual verification. We constructed a global co-authorship network based on fractional counting, and employed community detection algorithms with dynamic parameter tuning to identify research teams. Subsequently, we extracted leading teams based on multi-dimensional indicators and conducted comparative analyses.

[Results/Conclusions] We constructed rules for cleaning artificial intelligence scientific paper data from a practical perspective, established a process system for identifying artificial intelligence research teams based on co-authorship relationships, and proposed an approach of screening co-authorship networks by eliminating edge nodes and then using known teams as references for parameter adjustment. We systematically and accurately identified global artificial intelligence research teams and extracted leading research teams based on six-dimensional indicators including publication count, citation count, h-index, betweenness centrality, closeness centrality, and weighted degree centrality. Additionally, we provided exemplary analyses of each leading team by integrating paper data with empirical investigation.

Full Text

Identification and Extraction of Leading Research Teams in the Artificial Intelligence Field

Yu Houqiang, Bai Kuan, Zou Bentao, Wang Yuefen
School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094

Abstract:
[Purpose/Significance] This study identifies research teams in the artificial intelligence field and extracts leading teams based on multi-dimensional indicators, aiming to enrich the processes and methods of research team identification and provide a basis for analyzing the context, frontiers, and themes of AI from a research team perspective. [Method/Process] Using Web of Science as the data source, we collected all scientific papers in the AI discipline from 2009-2018, performed data cleaning through algorithm design and manual verification, constructed a global co-authorship network based on fractional counting, identified research teams using community detection algorithms with dynamic parameter tuning, and extracted leading teams for comparative analysis based on multi-dimensional indicators. [Result/Conclusion] We constructed data cleaning rules for AI scientific papers from a practical perspective, built a process system for identifying AI research teams based on co-authorship relationships, proposed an approach for screening co-authorship networks by eliminating edge nodes and then using known teams as references for parameter adjustment, systematically and accurately identified global AI research teams, and extracted leading teams based on six dimensions: publication count, citation count, h-index, betweenness centrality, closeness centrality, and weighted degree centrality, with exemplary analyses provided for each leading team using both paper data and empirical investigation.

Keywords: artificial intelligence; co-authorship network; research team; leading team; data analysis
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2020.20.001

Introduction

In the era of big science, research collaboration is regarded as a crucial means to improve research efficiency. A primary manifestation of research collaboration is the formation of research teams, which constitute an important component of the scientific community. Research teams are not only the backbone of scientific research, reflecting the degree of human resource concentration in a discipline, but also lead the trends and frontiers of scientific development. Therefore, research teams must be considered in studies of research development characteristics and patterns, and they require close attention in science policy formulation and adjustment. Moreover, in science and technology evaluation, particularly

regarding funding support and talent recruitment in universities, assessment and evaluation often need to be conducted in conjunction with research teams.

Research on team identification has primarily developed along two dimensions: (1) identifying research teams in different fields, and (2) improving algorithms for research team identification. Due to the importance of research teams in modern scientific research, various disciplines have addressed the issue, including management science, epidemiology, information science, and oncology. The most basic identification algorithm is the clique detection method in social network analysis. Subsequent developments have employed vector space models, introduced the FP-Growth algorithm from association rule mining, utilized factor analysis based on original data matrices, and implemented co-authorship network weighting methods for empirical research on team identification.

This study has two main starting points: First, the growing importance of AI has not yet been matched by systematic research specifically targeting research teams in this field. The rapid development of AI technology has attracted widespread global attention, not only drawing scholars from many disciplines but also deepening and expanding as countries and regions vigorously layout AI industries. Therefore, analyzing the AI field from a research team perspective has significant decision-support implications. Second, existing research team identification studies are primarily based on small-scale data, with few empirical analyses using large-scale datasets. Previous studies have typically selected a limited number of journals as data sources, identifying anywhere from a few to several dozen teams. Moreover, these methods mainly first identify team leaders and then expand to obtain team members. While this avoids the problem of indeterminate team size, the presence of a few extremely large teams means many other teams cannot be displayed when ranking by centrality.

This paper aims to address the following research questions: (1) How can research teams in AI be identified based on large-scale scientific literature data? This requires not only solving the large-scale cleaning of institutional and author data but also determining appropriate granularity for team segmentation through experimentation. (2) Which leading teams in AI can be extracted from different perspectives based on the identified research teams? Specifically, what leading teams emerge when different indicators are selected?

Research Design

2.1 Overall Process Design

The process of identifying research teams from large-scale datasets is essentially a data analysis process involving collection, processing, mining, and utilization. Therefore, this study adopts a data-driven approach, embedding the main content into operational steps and designing an overall process for research team identification, as shown in Figure 1 [Figure 1: see original paper].

2.2 Data Collection

This study uses Web of Science (WoS) data for analysis. WoS is one of the most authoritative databases for scientometric analysis, with its indexed journals typically considered core journals in their fields due to expert selection, ensuring high quality. Most WoS-indexed journals also have strong international orientation, facilitating analysis and comparison of global scientific publication patterns. Thus, WoS data offers good reliability and credibility.

AI is a complex emerging field. Keyword-based retrieval would face insufficient recall since AI involves numerous sub-topics, while using too many keywords would create precision problems due to ambiguous or overly broad terms. However, WoS's subject classification includes an AI subcategory under the Computer Science major category, covering all journals closely related to AI. Although WoS's subject classification is journal-based and has limitations, it is based on expert peer review and thus has good credibility. After comparison, using this subject classification for retrieval proved most effective. Therefore, we used the search query "WC = 'COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE' " to retrieve papers from 2009-2018, collecting 421,148 AI papers on January 16, 2019. The temporal distribution of these papers over the decade is shown in Figure 2 [Figure 2: see original paper].

2.3 Data Cleaning and Rule Construction

2.3.1 Institutional Data Cleaning Author disambiguation, a prerequisite for research team identification, requires combining author data with institutional data. Therefore, we first cleaned institutional data before proceeding to author data cleaning. Based on existing research and practice, we established the following institutional data cleaning process and rules:

(1) **Differentiation by country**: Extract paper institutions and country names. If institution names are identical but country names differ, they are treated as different institutions.

(2) **Iterative accumulation method**: Design cleaning rules to address four specific issues: First, institutions with different name orders are actually the same (e.g., "Washington Univ" and "Univ Washington"). Second, different laboratories or sub-institutions under the same organization are merged (e.g., "NICTA Canberra Lab" and "NICTA Queensland Lab" both belong to "NICTA"). Third, for domestic universities, English names and pinyin names are unified under the pinyin version (e.g., "Beijing Univ Aeronaut & Astronaut" becomes "Beihang Univ"). Fourth, if an institution has both abbreviated and full names, all abbreviations are unified to the full name (e.g., "EPFL, Switzerland" and "EPFL IC" become "Ecole Polytech Fed Lausanne, Switzerland").

(3) **Manual verification**: Using fractional counting to calculate publication counts for each institution and sorting them in descending order, we iden-

tified 17,511 institutions. We manually verified the top 1% of institutions by publication volume, merging cases where the same institution had different expressions not detected by the above rules, using the name with higher publication volume as the standard.

2.3.2 Author Data Cleaning Author name ambiguity is a widespread issue in scientometric analysis, generally categorized into "different forms, same meaning" (e.g., full vs. abbreviated names) and "same form, different meaning" (homonymy), which is particularly severe among Asian scholars. To ensure data quality, we established the following author name disambiguation process based on the cleaned institutional data:

(1) **Data format conversion**: Convert WoS-format raw data into a processable format.

(2) **Extract author institutions and co-authors**: Use the previously cleaned institutional information.

(3) **Author disambiguation based on institution and co-author information**: First, authors with the same name from the same institution are treated as the same author. Second, same-name authors from different institutions are judged to be the same author entity if they share co-authors. Third, same-name authors not meeting the above conditions are determined to be different authors and distinguished using "underscore + number" notation.

Before disambiguation, there were 530,000 authors. After disambiguation, we obtained 650,000 authors due to name splitting, with results stored in structured forms.

2.4 Research Team Identification

2.4.1 Construction of the Global Co-authorship Network After author name disambiguation, we constructed a global co-authorship network from all co-authored papers in the dataset, involving 656,668 nodes (authors) and 2,042,924 edges, with edge weights representing collaboration intensity calculated by total co-authorship frequency. Research shows that fractional counting better identifies clusters in networks and correlates more closely with actual research output than the commonly used full counting method. Therefore, this study uses fractional counting, where if a paper has n authors, the collaboration frequency between any two authors is 1/n. Due to the network's massive scale, visualization was impossible, and all subsequent operations were completed programmatically.

For the initial network, we used the Louvain method in Pajek with "Multi-Level Coarsening + Multi-Level Refinement" and default parameters, detecting 94,347 communities. These communities were tightly connected internally yet separated from others, thus regarded as research teams formed through

co-authorship relationships. Descriptive statistical analysis revealed the largest team contained 1,553 authors, and among the top 10 teams by size, only one had fewer than 1,000 members (996). These results failed to meet our fine-grained analysis needs, and actual research team sizes differed from our identification results, likely because numerous edge nodes with single collaborations and low importance were mistakenly identified as team members. Therefore, we needed to extract from the original co-authorship network.

In this study, we chose to remove author nodes with only one publication and fewer than 100 citations, as these authors had very weak influence in AI research and might not be true AI scholars (e.g., graduate students or technicians involved in peripheral work). After extraction, we obtained a new co-authorship network with 186,997 nodes and 543,351 connections—a substantial reduction of 469,671 nodes and 1,499,573 connections while preserving important nodes.

2.4.2 Selection of Identification Granularity To identify reasonably sized, analyzable research teams, we continuously adjusted parameters and evaluated the resulting teams. Due to the network's massive scale, each parameter adjustment required substantial computational resources. As parameters were tuned, identified teams gradually stabilized. We used the Louvain algorithm for clustering, where the Resolution parameter affects cluster size, and MaxLevel and MaxIteration parameters correspond to algorithmic iteration constraints. Using a known AI research team (Jiao Licheng's team from Xidian University) as a reference, we tuned parameters. When Resolution=290, MaxLevel=13, and MaxIteration=13, we achieved an appropriate granularity for observing internal team structures, yielding results matching our empirical investigation, with team sizes under 100 members. Overly large teams might have loose internal connections, while overly small teams might omit important connections. Note that no strict standard exists for team size granularity selection; it merely reveals different degrees of team cohesion.

During parameter tuning, we also validated the rationality of team origins to ensure identified teams derived from larger teams, as shown in Figure 3 [Figure 3: see original paper].

2.5 Extraction of Leading Teams

The AI field contains many research teams, but attention primarily focuses on leading teams. Therefore, after identifying research teams, we further extracted leading teams. We measured teams using six indicators from different perspectives: Number of Publications, Number of Citations, h-index, Weighted Degree Centrality, Betweenness Centrality, and Closeness Centrality. The first three measure team strength from node attributes, while the latter three measure strength from network structure. Publication count, citation count, and h-index values were calculated using custom Python programs based on their definitions, while centrality metrics were calculated using the large-scale social

network analysis tool Pajek. For each dimension, we selected the top 10 teams as leading teams.

Research Results and Analysis

3.1 Overall Situation of Research Teams

Based on the above process, we identified 23,423 AI research teams involving 186,997 authors. The team size distribution is shown in Figure 4 [Figure 4: see original paper]. Xu et al. argue that small teams (under 10 members) outperform large teams (over 25 members) in network density and collaboration intensity. Figure 4 shows that 89.4% of teams have 25 or fewer members, with 78% having fewer than 10 members, indicating a reasonable size distribution.

Second, we measured and ranked teams across six dimensions: publications, citations, h-index, betweenness centrality, closeness centrality, and weighted degree centrality. We defined the top 234 teams (top 1%) as leading teams and selected the top 10 from each indicator for display and analysis, involving 47 teams total (10 teams appeared in the top 10 of two or more indicators). Teams #205, #342, and #207 ranked in the top 10 for three indicators. Due to space limitations, we provide concise analyses only for the top three teams in each dimension and visualizations for the top-ranked team.

3.2 Leading Teams Based on Publication Count

Leading teams by publication count are shown in Table 1 . High-publication teams typically have over 50 members. For example, team #448 (ranked first) has 52 closely collaborating authors, as shown in Figure 5 [Figure 5: see original paper], with leading scholar Pedrycz Witold_2 (the number suffix distinguishes homonymous authors in our analysis). This team's research frontiers focus on time series, fuzzy cognitive maps, and data mining. Team #1927 (second) has 54 members led by Zhang Mengjie_2, focusing on edge detection and machine learning. Team #1064 (third) has 58 members led by Castillo Oscar, focusing on dynamic parameter adaptation based on fuzzy logic.

3.3 Leading Teams Based on Citation Count

Leading teams by citation count are shown in Table 2 . High-citation teams show greater variation in size, ranging from 23 members (team #5997) to 80 members (team #843). Team #2096 (ranked first) has 36 closely collaborating authors centered on Lin Chin-Jen, as shown in Figure 6 [Figure 6: see original paper]. This leading team focuses on AI-related algorithm research, particularly classification algorithms and large-scale linear classification. Team #5330 (second) centers on Sun Jian_{14}, focusing on new models and methods for image classification. Team #1959 (third) centers on Ma, Yi_4, focusing on computer vision issues such as image recognition under severe damage and object detection models, approached through matrix methods.

3.4 Leading Teams Based on h-index

Leading teams by h-index are shown in Table 3 . High h-index teams tend to have more members, with the largest (team #207) having 98 members. Team #594 (ranked first) has 77 closely collaborating authors centered on Wang Zidong from Brunel University, UK, as shown in Figure 7 [Figure 7: see original paper]. Their research frontiers focus on synchronization control, many-objective optimization, and exponential stability of neural networks under time-varying delays. Team #205 (second) centers on Herrera Francisco from the University of Granada, Spain, focusing on group decision-making, base classifiers, and classification systems. Team #108 (third) has 80 members focusing on logistics, healthcare, and warehousing applications combining fuzzy association rule mining with fuzzy logic.

3.5 Leading Teams Based on Betweenness Centrality

Leading teams by betweenness centrality are shown in Table 4 . High betweenness centrality teams typically have under 40 members. Team #698 (ranked first) has 44 closely collaborating authors centered on Zhang, wei_{27} from Chongqing University, as shown in Figure 8 [Figure 8: see original paper]. Their research frontiers focus on graphics recognition technology, with recent emphasis on improving image matching accuracy through line segment matching. Team #1348 (second) centers on Willmann, T, focusing on AI-related algorithm research, particularly learning vector quantization (LVQ) algorithms.

3.6 Leading Teams Based on Closeness Centrality

Leading teams by closeness centrality are shown in Table 5 . High closeness centrality teams generally have over 80 members. Team #2352 (ranked first) has 100 closely collaborating authors centered on Hassanien, Aboul Ella from Cairo University, Egypt, as shown in Figure 9 [Figure 9: see original paper]. Their research frontiers focus on quantum-behaved particle swarm optimization for SVM parameter optimization, Bayesian optimization approaches, multi-objective optimization algorithms, and mobile damped wave algorithms for global optimization problems. Team #2733 (second) centers on D'Mello, Sidney from the University of Notre Dame, focusing on educational applications of computer vision technology to capture and analyze learners' facial expressions for intelligent tutoring systems, with achievements in speech recognition and emotion analysis. Team #1063 (third) centers on Bajo, Javier from the Technical University of Madrid, focusing on adaptive fault-tolerant tracking control algorithms for IoT systems, nonlinear adaptive closed-loop control systems for blockchain management efficiency, and distributed continuous-time fault estimation control for IoT multi-devices.

3.7 Leading Teams Based on Weighted Degree Centrality

Leading teams by weighted degree centrality are shown in Table 6 . The distribution of team sizes is relatively uniform. Team #3127 (ranked first) has 22 closely collaborating authors centered on Perny, Patrice from Paris VI University, France, as shown in Figure 10 [Figure 10: see original paper]. Their research frontiers focus on decision theory applications in AI, including public facility location, electricity market trade negotiations, multilateral negotiation strategies, carbon emission assessment and trading, and multi-criteria decision-making. Team #2056 (second) centers on Xu, Yang_7, focusing on lattice-valued logic, lattice implication algebra, SAT problems, and fuzzy logic. Team #2242 (third) has 38 members centered on Ramirez, J. and Gorriz, J., focusing on pattern recognition and AI methods applied to biology and medicine, particularly Alzheimer's disease research in recent years.

3.8 Comparative Analysis of Leading Teams Across Six Dimensions

Comparing leading teams identified across the six dimensions (Table 7), teams #205 and #342 ranked in the top 10 across all three research dimensions (publications, citations, and h-index), while team #207 ranked in the top 10 for h-index, closeness centrality, and weighted degree centrality. Additionally, teams #196, #203, #2242, #2733, #3127, #594, and #795 appeared in the top 10 for two dimensions. These results demonstrate that different dimensions reveal different connotations of leadership advantages, while some leading teams show advantages across multiple dimensions. Notably, teams excelling in h-index are more likely to also demonstrate advantages in other dimensions.

Research Conclusions

Using WoS AI discipline data from 2009-2018, this study constructed a complete data analysis process from data cleaning through network construction, research team identification, and leading team extraction. The main contributions are threefold:

(1) **Iterative accumulation-based data cleaning rules**: We developed an AI research institution alias correspondence table and proposed a large-scale author disambiguation method based on institutional names and co-authors, empirically validated on the AI paper dataset. This simple and feasible approach yields good disambiguation results and can be applied to other disciplines.

(2) **Process system for AI research team identification**: Using fractional counting to construct global co-authorship networks, we extracted networks by eliminating edge nodes and dynamically adjusted parameters using known teams as references, identifying appropriately granular research teams. This approach is applicable to team segmentation in other disciplines where objective standards are lacking.

(3) **Leading team extraction across six dimensions**: We identified leading teams from publications, citations, h-index, betweenness centrality, closeness centrality, and weighted degree centrality, analyzing team compositions and research themes. While these six indicators are well-recognized in scientific evaluation and applicable to other fields, other dimensions could be explored in future research.

References

[1] Chen Chunhua, Yang Yingshan. Research on scientific research management based on team operation mode [J]. Science & Technology Progress and Policy, 2002(4): 79-81.

[2] ACEDO F J, BARROSO C, CASANUEVA C, et al. Co-authorship in management and organizational studies: an empirical and network analysis [J]. Journal of management studies, 2006, 43(5): 957-983.

[3] GREGORIO G, JINSEO P, CHARLES H, et al. Scientific authorships and collaboration network analysis on chagas disease: papers indexed in pubmed (1940-2009) [J]. Journal of the institute of tropical medicine in sao paulo, 2012, 54(4): 219-228.

[4] Li Liang, Zhu Qinghua. Empirical study of social network analysis in co-authorship analysis [J]. Information Science, 2008, 26(4): 549-555.

[5] Li Gang, Li Chunya, Li Xiang. Research on discovering scientific research teams based on social network analysis [J]. Library and Information Service, 2014, 58(7): 63-70, 82.

[6] Shen Gengyu, Huang Shuiqing, Wang Dongbo. Research on discovering scientific research teams using author collaboration co-occurrence as source data [J]. Data Analysis and Knowledge Discovery, 2013, 29(1): 57-61.

[7] Lü Lucheng, Zhao Yajuan, Wang Xuezhao, et al. R&D team identification method based on association rule mining [J]. Science and Technology Management Research, 2016, 36(17): 148-152, 189.

[8] ANTONIO P, CARLOS O, FÉLIX M. Detecting, identifying and visualizing research groups in co-authorship networks [J]. Scientometrics, 2010, 82(2): 307-319.

[9] Ren Ni, Zhou Jiannong. Discovery and evaluation of research teams in weighted co-authorship network mode [J]. New Technology of Library and Information Service, 2015, 31(9): 68-75.

[10] Fan Lipeng, Yu Houqiang, Jiang Yuxing, et al. Identification and analysis of AI research frontiers: based on high-productivity institution comparative research [J]. Information Studies: Theory & Application, 2019, 42(9): 16-21.

[11] TRAN H N, HUYHN T, DO T. Author name disambiguation using deep neural network [C]//Asian conference on intelligent information and database

systems. Phuket, Thailand: ACIIDS, 2014: 123-132.

[12] GLÄNZEL W. National characteristics in international scientific co-authorship relations [J]. Scientometrics, 2001, 51(1): 69-115.

[13] PRITYCHENKO B. Fractional authorship in nuclear physics [J]. Scientometrics, 2016, 106(1): 461-468.

[14] Xu Zhi, Chen Liyu, Wang Sihui. Research on influencing factors of university scientific research team cooperation degree [J]. Science Research Management, 2015, 36(5): 149-161.

Author Contributions:
Yu Houqiang: Data collection, research design, methodology, paper writing;
Bai Kuan: Data processing, paper writing;
Zou Bentao: Data acquisition, data analysis, paper revision;
Wang Yuefen: Overall paper structure, revision suggestions, final approval.

Abstract: [Purpose/significance] This paper identifies the research team in the artificial intelligence field and extracts the leading research team from multi-dimensional indicators, aiming to enrich the process and method of identification of the research team, and provide the basis for analyzing the context, frontier and theme of the artificial intelligence field from the perspective of the research team. [Method/process] This paper was based on the publication data of the Web of Science category Computer Science, Artificial Intelligence from 2009 to 2018, and did data cleaning via programming and manual check. Global co-author network is constructed based on the fractional counting method, and the Louvain algorithm was used to dynamically tune and identify the research teams. Moreover, the leading research team was extracted based on different indicators with parameter adjustment. [Result/conclusion] From practical view, the study has constructed a set of rules for cleaning publication data of artificial intelligence field. The process of identifying artificial intelligence research teams based on co-authorship is constructed. The study proposes the method of tuning the parameter by eliminating edge nodes in the collaboration network and further taking the known research teams as baseline. The worldwide research teams of artificial intelligence field are systematically and accurately identified. The leading research teams are further extracted based on indicators of six dimensions, i.e. number of publications, number of citations, h-index, weighted degree centrality, betweenness centrality, closeness centrality. Exemplary analysis is conducted on leading research teams of each dimension by combining the publication data and web information survey.

Keywords: artificial intelligence; co-authorship network; research team; leading team; data analysis

Note: Figure translations are in progress. See original paper for figures.