Pagerank Algorithms and Multi-Particle Environments Using Pettingzoo: Network Simulation and Validation

Yi Li *

Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

* Corresponding Author Email: yli488@sheffield.ac.uk

Abstract. This study introduces and investigates two novel algorithms, namely Topic-Specific Crawling PageRank and Motif-Based PageRank, along with their corresponding application scenarios. The former algorithm focuses on targeted web crawling, refining the collection of web pages relevant to specific topics. The latter algorithm addresses higher-order relationships in citation networks, enhancing insights into researcher influence. These algorithms are evaluated within a Multi-Particle Environment, showcasing their capability to simulate necessary scenarios and data for PageRank algorithms. However, the environment's realism is limited due to its use of virtual data. The environment offers controllability through the PettingZoo resource library, providing an efficient experimentation platform. While the Multi-Particle Environment displays potential and advantages, its deviation from real-world data should be carefully considered. The experimental methodology involves constructing a network of agents and their message interactions within a simulated environment. The resulting network undergoes the PageRank algorithm, and the resulting ranking aligns with the expected outcome, confirming the Multi-Particle Environment's compatibility with the algorithm. These experiments furnish empirical evidence for the Multi-Particle Environment's adaptability to algorithmic simulations and real-world scenarios. This research contributes to a deeper comprehension of algorithm behavior in diverse and dynamic settings, paving the way for further advancements in the field.

Keywords: PageRank; Multi-Partial Environment; Real-World Integration; Complex Network.

1. Introduction

Over the past seventy years, technological advancements have led to an unprecedented surge in inaccessible information, surpassing historical levels [1]. Virtually every aspect of people's lives has become inundated with a diverse array of information sources. While the internet has broken down traditional barriers in information dissemination, its expansive growth has also given rise to the problem of information overload [2]. The PageRank algorithm emerged as a response to combatting issues of information overload and low information quality on the internet. It was invented by Google's founders, Larry Page and Sergey Brin, in 1998. PageRank aimed to address the challenge of navigating through the vast volume of web pages. Traditional search engine algorithms often relied on rudimentary methods like keyword matching for ranking, struggling to accurately gauge the significance and quality of web pages. PageRank introduced a more sophisticated approach, analyzing link structures and assessing both the quantity and quality of incoming links to determine web page importance. This innovation mitigates information overload by guiding users to more pertinent and trustworthy content, while minimizing exposure to low-quality or misleading information [3].

The Multi-Particle Environments (MPE) experiment leverages the support of the PettingZoo framework. MPE within PettingZoo facilitates data extraction and algorithm evaluation by simulating essential information collection. PettingZoo, a Python library, is designed to aid the creation and testing of environments for multi-agent reinforcement learning. Its versatile range of intricate environments allows for the simulation of diverse interactive scenarios involving multiple agents. By providing various game-like environments, PettingZoo empowers researchers and developers to explore distinct dynamics, strategies, and agent interactions. This makes it an invaluable tool for algorithm evaluation, hypothesis testing, and the advancement of multi-agent reinforcement learning [4]. This paper aims to evaluate the appropriateness of the multi-particle framework for assessing the

effectiveness of PageRank algorithms. The obtained simulation results align with the initial predictions and expectations.

2. Related Work

The PageRank algorithm is rooted in graph theory, and its principles have significant applications in network analysis and web search. Initially designed to rank web pages based on their hyperlink structures, PageRank assesses a page's importance by considering the quantity and quality of links pointing to it. This iterative process assigns importance scores, enhancing search engine query results.

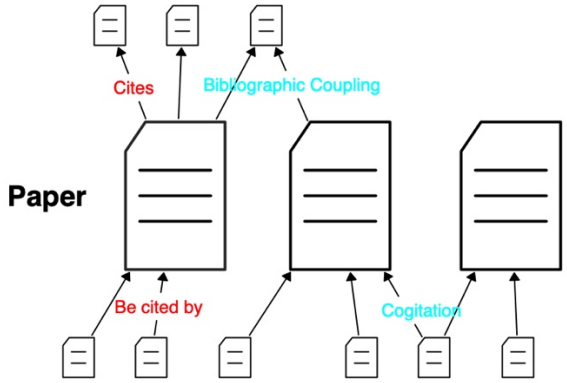

Fig. 1 PageRank in Citation Network

PageRank's influence transcends its original purpose in web search. In bibliometric studies, it's adapted to assess academic paper impact within citation networks, uncovering influential research and collaborations [5]. Fig.1 illustrates a reference relationship in a Citation Network. Additionally, PageRank aids social network analysis, identifying central nodes in online social networks. Fig.2 portrays a web page relationship. Its use in recommendation systems enables tailored content suggestions based on user interaction history, showcasing its adaptability in personalized content delivery [6].

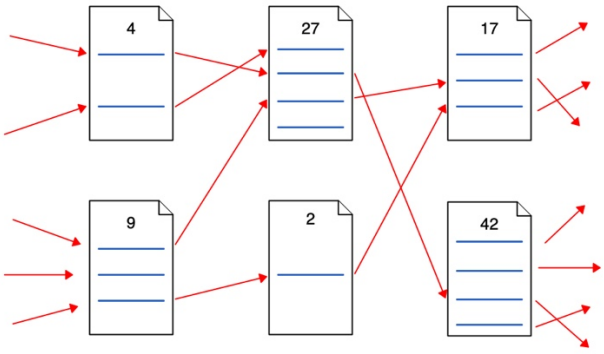

Fig. 2 PageRank in Web Social Network

Despite its broad usefulness, PageRank has limitations. In dynamic networks, where link structures evolve rapidly, their static nature may struggle to capture evolving influence patterns. Susceptibility to "link spam" and manipulative tactics challenges its accuracy in reflecting true authority. Moreover, in cases of disconnected components or "spider traps," PageRank's effectiveness diminishes.

To address these issues, contemporary research explores variations and enhancements to the original algorithm. The following sections analyze the potential alignment between simulated multi-particle environments and PageRank. This examination assesses the environment's ability to shed light on PageRank's performance and adaptability across diverse scenarios.

The multi-particle environment emulates graph networks effectively [7, 8]. It represents particles as nodes and their interactions as edges, closely resembling graph characteristics. Particle interactions mirror graph network dynamics, simulating phenomena like information propagation and

connectivity patterns [9]. This versatility makes it valuable for studying complex network phenomena, algorithm evaluation, and understanding node behavior and connections. Its emulation of graph networks positions it as an apt choice for investigating graph-based algorithms like PageRank and understanding the intricate dynamics of interconnected systems.

3. Methodology

3.1. PageRank

PageRank is an algorithm used to assess the importance of web pages by leveraging their interlinking relationships. The algorithm's fundamental concept introduces the notion of page authority, which derives solely from the web's topology and is independent of page content. PageRank views a page's authority akin to citations in scientific literature. The authority of page p depends on hyperlinks directed towards it (akin to citations) and the authority of pages p that cite it. The algorithm computes scores iteratively, where higher scores lead to prominent search result rankings. The traditional algorithm is defined by the formula 1:

$$PR(p) = (1 - \mathrm{d}) + d \sum_i \frac{PR(p_i)}{C(p_i)} \tag{1}$$

In this equation, PR(p) represents the PageRank score assigned to page p, while d stands for a damping factor within the range of 0 to 1(Typically set to 0.85), and $\sum_i \frac{PR(p_i)}{C(p_i)}$ represents the cumulative PageRank scores of all pages i that link to page p divided by the out-degree $C(p_i)$ of page i. The formula signifies that the calculated value is the summation of PageRank scores from all pages that are linked to it, which is then normalized by the count of outbound links originating from each linking page i. This result is then scaled by the damping factor d and added to a constant factor $(1 - \mathrm{d})$. The damping factor d represents the probability of random hopping while browsing, while the constant $(1 - \mathrm{d})$ accounts for the probability of not hopping.

This iterative calculation process can be efficiently implemented using matrix-vector multiplication, enhancing computational efficiency. Overall, your description comprehensively captures the essence of PageRank's principles and formula, effectively conveying its significance in web page ranking.

3.2. Motif-Based PageRank

Motif-Based PageRank enhances traditional algorithms by accurately calculating node authority scores. It introduces matrices based on motifs (subgraphs) to capture complex relationships beyond direct node connections. Motif-Based PageRank uses motif-based adjacency matrices to assess pairwise node relationships, and edge-based adjacency matrices to evaluate direct links. The algorithm is defined by formula 2:

$$PR(p) = (1 - \mathrm{d}) + d \sum_i \frac{PR(p_i)}{C(p_i)} * H_{M_k} i \tag{2}$$

Where $H_M i$ is the fusion of matrices based on motifs and edges. According to Zhao H et al [10]. there are two ways to combine which are linear and non-linear, formula 3 demonstrates the linear combination, and formula 4 illustrates the non-linear combination:

$$H_{M_k} = a * W + (1 - a) * W_{M_k} \tag{3}$$

$$H_{M_k} = W^a * W_{M_k}^{(1-a)} \tag{4}$$

Where H_{M_k} represents the amalgamated adjacency matrix, W signifies the adjacency matrix based on direct connections, W_{M_k} denotes the adjacency matrix capturing higher-order relationships, and α stands for a weight parameter aimed at equilibrium between the significance of them both.

Finally, to calculate the authority score for each node, the computation of the motif-based adjacency matrix is an essential undertaking, serving to encapsulate the nuanced pairwise relationships that transpire between two nodes within each distinct subgraph. For a given motif M, its motif-based adjacency matrix is denoted as W_M, where $(W_M)_{ij}$ denotes the number of times node i and node j appears in the motif M. Specifically, $(W_M)_{ij}$ is calculated by formula 5:

$$(W_M)_{ij} = \sum_{v,X_A(v)\in M} 1(\{i,j\} \sqsubset X_A(v)) \tag{5}$$

Where $v, X_A(v)$ is an instance in motif M, $X_A(v)$ is the set of neighbouring nodes of node v in the set of anchors A, and $1(\{i,j\} \sqsubset X_A(v))$ is an indicator function that is 1 when both node i and node j are neighbours of node v and 0 otherwise.

Compared to the traditional PageRank algorithm, Motif-based PageRank takes into consideration higher-order relationships among nodes beyond direct linkages. This heightened sensitivity to complex node relationships leads to more precise authority score calculations, especially within intricate networks like social networks.

3.3. Topic-Specific Crawling PageRank

The algorithm is a topic-specific crawling algorithm based on a relevant context graph. A context graph is constructed using search engine data and word distributions are calculated to determine the relevance of a page to a given topic. Iteratively calculates PageRank value for each page's importance, uses depth-first search strategy based on relevance graph to crawl pages related to topic. Formula 6 constructs a relevant context graph:

$$C(t) = \{d \mid d \in D \wedge t \in T(d)\} \tag{6}$$

Where $C(t)$ denotes the set of documents related to query t, D denotes the set of documents, and $T(d)$ denotes the set of words in document d. Formula 7 for calculates the word distribution:

$$P(w|t) = \alpha * P(w|G) + (1-\alpha) * P(w|T(t)) \tag{7}$$

Where $P(w|t)$ denotes the distribution of word w under query t, $P(w|G)$ denotes the distribution of word w under a generic language model, $P(w|T(t))$ denotes the distribution of word w under a topic-specific language model, and α is a smoothing parameter.

4. Experiment

The experiments are based on the theoretical foundations of two algorithms: Topic-Specific Crawling PageRank, and Motif-Based PageRank, both of which are further supported by provided code for reproducibility. Leveraging the multi-agent environment facilitated by pettingZoo, fundamental scenarios were crafted as a basis for simulating these algorithms.

Topic-Specific Crawling PageRank is tailored for topic-specific web crawling, enabling the acquisition of web pages pertinent to a given topic from the internet. This algorithm aids search engines and related tools in efficiently gathering information aligned with specific themes, thereby enhancing information retrieval efficiency and accuracy. Inputs to this algorithm encompass two key facets: the designated topic and a set of initial URLs. The topic represents the specific subject of interest for web crawling, such as "Computer Science" or "Mountaineering". The specified initial URLs serve as starting points for the algorithm's automated extraction of topic-related pages. Priority computation, based on the relevance-context graph retaining word probabilities, determines the sequence for accessing unvisited pages. This algorithm selectively retrieves an internet subset relevant to the specified topic and excels in establishing page-topic relevancy, surpassing other methods. In the experimental setup, agents are designated as distinct web pages, each associated with an independent URL and topic. These data are stored in vector format, facilitating inter-agent communication to represent directional connections.

Motif-Based PageRank, on the other hand, is geared towards addressing higher-order relationships within citation networks. In this context, higher-order relationships pertain to indirect connections between researchers that are not established through direct citations. For instance, two researchers might not directly cite each other, yet they might have co-authored papers with researchers who do cite each other. Through analysis of these indirect relationships, Motif-Based PageRank better captures researchers' influence and authority in citation networks. Unlike Topic-Specific Crawling PageRank, the input data for Motif-Based PageRank primarily consists of citation relationships. In the proposed multi-agent environment, each agent embodies a distinct author, marked with varying degrees of influence, representing different researchers' impacts and authorities. They communicate by sending messages to denote citation relationships.

Throughout the experiments, the established multi-agent environment is instrumental in forming an agent graph network through message exchanges, where each agent signifies a unique node [11, 12]. The experiment records all transmitted messages among nodes. Upon the program's conclusion, the recorded data is processed and input into the respective PageRank algorithms, leading to conclusions that are subsequently compared with the attributes set for the experiment's agents.

4.1. Setup

Table 1 and Table 2 display the hardware and software setup for this experiment.

Table 1. Hardware Configuration

Component	Specification
Operation System	Windows 10 & macOS 12.4
Memory	16GB
CPU	Intel (R) Core (TM) i7-10750 & Apple M1 Pro
GPU	RTX2060 & Apple M1 Pro

Table 2. Software Configuration

PettingZoo	Motif-Based PageRank	Topic-Specific Crawling PageRank
Python 3.9.12	Python 2.7.12	Python 3.9.12
Numpy 1.22	Numpy 1.11	Numpy 1.22
NetworkX 2.7.1	NetworkX 1.11	NetworkX 2.7.1

4.2. Dataset

The experimental data section encompasses two distinct categories of data sources. The first category involves data collected from a simulated environment, while the second category draws upon publicly available datasets from online sources, including well-known datasets such as DBLP. In the simulated environment, a multi-particle framework facilitates the creation of a controlled setting mimicking real-world interactions among web pages or researchers. This artificial yet meticulously designed environment allows for generating data reflecting various scenarios and network dynamics. Interactions, agent attributes, citation relationships, and collaborative patterns are captured and recorded to generate the synthetic dataset. Complementing this simulated dataset, the second category features publicly available datasets from reputable sources like DBLP. These datasets encapsulate real-world scholarly interactions, citations, and collaborations among researchers, introducing authenticity and complexity to the experiments.

Utilizing both synthetic and publicly available datasets provides a comprehensive perspective on algorithm evaluation. While simulated data enables controlled testing, real-world datasets offer nuanced insight into algorithm performance under complex conditions. The combination of these two types of data enhances the experimentation's robustness, allowing for a well-rounded assessment of algorithm capabilities across varying data sources. This hybrid approach validates algorithm adaptability and emphasizes their potential in real-world scenarios, reinforcing the credibility of experimental outcomes.

4.3. Performance Analyse

In this experiment, a graph network is constructed through a series of operations using raw data, depicting relevant web pages related to a specific topic and their interlinking relationships, as shown in Fig.3 and Fig.4. This ensures the alignment of the experimental data with the scenarios of Topic-Specific Crawling PageRank and Motif-Based PageRank.

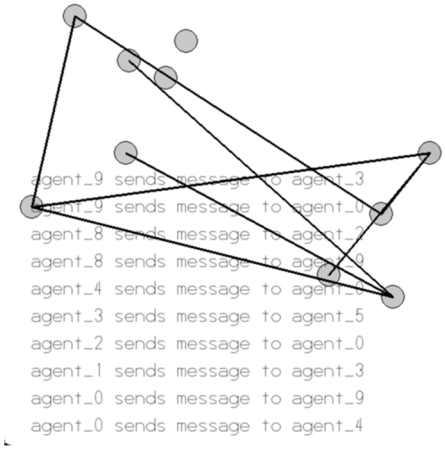

Fig. 3 Multi-Agent Environment in Experiment

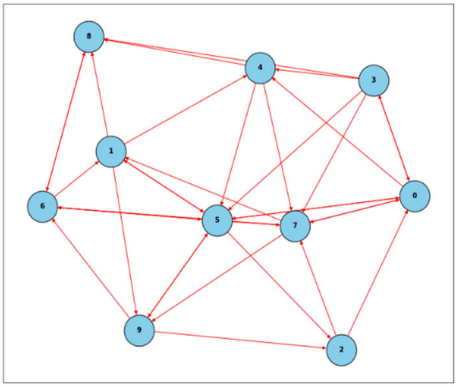

Fig. 4 Agent Graph Network based on MAE.

The Topic-Specific Crawling PageRank algorithm computes this graph network. It starts with a priority calculation to access unvisited pages, relying on a relevancy context graph to preserve term probabilities. This approach estimates the correct sequence for accessing pages. The algorithm iteratively computes the PageRank value for each webpage to ascertain their significance and relevance. This setup aligns with the algorithm's objective of retrieving web pages relevant to a designated topic from the internet.

For Motif-Based PageRank, the experiment uses data from the multi-particle environment and applies formula 2 to calculate PageRank values. Calculated values are sorted, and accuracy is determined based on agent levels, as shown in Fig.5 and Fig.6. Agent1 and Agent4 exhibit the highest citation and co-citation counts, illustrated in Fig.6. These agents also possess the highest levels and maximum PageRank values. Similarly, Agent7 and Agent5 have the lowest levels, and the algorithm assigns them the lowest PageRank values, positioning them at the bottom of the ranking.

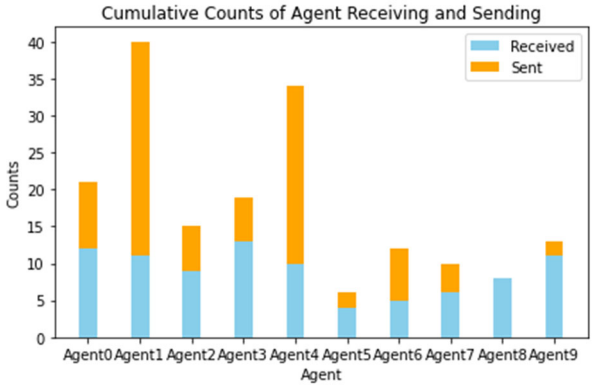

Fig. 5 Cumulative Counts of Agents Receiving and Sending

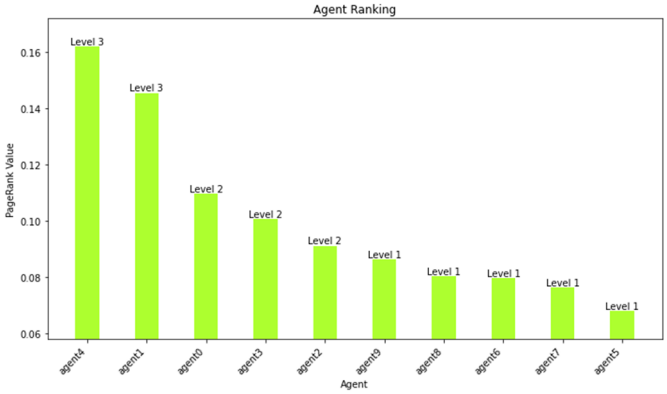

Fig. 6 Agent Ranking Value

Based on the analysis of the experimental results, it can be concluded that the Multi-Particle Environment constructs targeted search networks for Topic-Specific Crawling PageRank and networks with higher-order relationships for Motif-Based PageRank. The environment simulates scenarios and data needed for PageRank algorithms, demonstrating adaptability and realism. With support from the pettingZoo library, virtual scenarios offer control for adjusting agent attributes. Note that the Multi-Particle Environment's data is virtual and may lack realism of real-world data in certain aspects. Yet, it reduces human resource costs, providing an efficient experimental approach. In conclusion, the Multi-Particle Environment shows potential in experiments, but consider differences between virtual and real data when using it.

5. Conclusion

This study investigates two novel algorithms, modeling corresponding application scenarios: topic-specific crawling PageRank and topic-based PageRank. Each algorithm addresses distinct aspects of web analytics and information retrieval. The practical application of topic-specific crawling PageRank demonstrates its effectiveness in targeted web crawling, resulting in a refined and efficient collection of web pages related to specific topics. Conversely, topic-based PageRank captures higher-order relationships in citation networks, providing insights into researchers' influence and authority. The experimental data, generated through the Multi-Particle Environment, successfully showcases its compatibility with PageRank algorithms.

Furthermore, this study illuminates the potential of the introduced algorithms and their alignment with the Multi-Particle Environment. Several promising avenues for future research and development emerge. Integrating real-world data and scenarios into the Multi-Particle Environment would bridge the gap between simulation and reality. This could involve incorporating actual web data, user behavior patterns, and dynamic content updates for more accurate algorithm evaluations. Exploring algorithm behavior in dynamic environments, where network structures, content, and user preferences evolve, could reveal adaptability and stability insights. Incorporating user behavior modeling into

algorithms could enhance their personalization capabilities. By considering factors like click-through rates, browsing history, and preferences, algorithms could provide more tailored and relevant search results, enhancing user satisfaction and engagement. In conclusion, these proposed directions aim to enhance the applicability and performance of the introduced algorithms. By integrating real-world scenarios, exploring network complexities, adapting to dynamic environments, employing hybrid approaches, and incorporating user behavior insights, future research can contribute to refining and advancing web analytics and information retrieval techniques.

References

[1] Edmunds A, Morris A. The problem of information overload in business organisations: a review of the literature [J]. International journal of information management, 2000, 20(1): 17-28.

[2] Sunil Datt M B A. The information explosion: Trends in technology 2011 review [J]. The Journal of Government Financial Management, 2011, 60(4): 46.

[3] Liu J Q, Li X R, Dong J C. A survey on network node ranking algorithms: Representative methods, extensions, and applications [J]. Science China Technological Sciences, 2021, 64(3): 451-461.

[4] Terry J, Black B, Grammel N, et al. Pettingzoo: Gym for multi-agent reinforcement learning [J]. Advances in Neural Information Processing Systems, 2021, 34: 15032-15043.

[5] Ding Y, Yan E, Frazho A, et al. PageRank for ranking authors in co-citation networks [J]. Journal of the American Society for Information Science and Technology, 2009, 60(11): 2229-2243.

[6] Al_Janabi S, Kadiam N. Recommendation system of big data based on pagerank clustering algorithm[C]//Big Data and Networks Technologies 3. Springer International Publishing, 2020: 149-171.

[7] Kravets A G, Kravets A D, Korotkov A A. Intelligent multi-agent systems generation [J]. World applied sciences journal, 2013, 24(24): 98-104.

[8] Liu Y, Wang W, Hu Y, et al. multi-agent game abstraction via graph attention neural network [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 7211-7218.

[9] Veličković P. Everything is connected: Graph neural networks [J]. Current Opinion in Structural Biology, 2023, 79: 102538.

[10] Zhao H, Xu X, Song Y, et al. Ranking users in social networks with motif-based pagerank [J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(5): 2179-2192.

[11] Kraus S, Azaria A, Fiosina J, et al. AI for explaining decisions in multi-agent environments [C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(09): 13534-13538.

[12] Veličković P. Everything is connected: Graph neural networks. Current Opinion in Structural Biology. 2023 Apr 1;79:102538.