Comparing Personalized Relevance Algorithms for Directed Graphs

Luca Cavalcanti
DISI, University of Trento
Trento, Italy
0009-0001-2529-1639

Cristian Consonni[§]
Joint Research Centre, European Commission
Ispra, Italy
0000-0002-2490-8967

Martin Brugnara
DISI, University of Trento
Trento, Italy
0000-0002-7348-5147

David Laniado
Eurecat - Centre Tecnològic de Catalunya
Barcelona, Spain
0000-0003-4574-0881

Alberto Montresor
DISI, University of Trento
Trento, Italy
0000-0001-5820-8216

Abstract—We present an interactive Web platform that, given a directed graph, allows identifying the most relevant nodes related to a given query node. Besides well-established algorithms such as PageRank and Personalized PageRank, the demo includes Cyclerank, a novel algorithm that addresses some of their limitations by leveraging cyclic paths to compute personalized relevance scores. Our demo design enables two use cases: (a) algorithm comparison, comparing the results obtained with different algorithms, and (b) dataset comparison, for exploring and gaining insights into a dataset and comparing it with others. We provide 50 pre-loaded datasets from Wikipedia, Twitter, and Amazon and seven algorithms. Users can upload new datasets, and new algorithms can be easily added. By showcasing efficient algorithms to compute relevance scores in directed graphs, our tool helps to uncover hidden relationships within the data, which makes of it a valuable addition to the repertoire of graph analysis algorithms.

Index Terms—algorithms, directed graphs, personalized pagerank, cyclerank, wikipedia link graph, interactive dashboard

I. Introduction

The relevance of nodes within a graph can be computed globally, or in relation to some specific node. The latter is useful when identifying the nodes that are most relevant within a specific context. For instance, in the graph of hyperlinks between Wikipedia articles (Wikilink network), that can be viewed as a semantic graph connecting the corresponding concepts and entities [1], one may want to identify the concepts that are most relevant to a given one. Or in the network of interactions between users in a social media platform, such as Twitter, one can look for the users that are more relevant with respect to a given user.

One established algorithm from the state of the art for this problem is Personalized PageRank (PPR), a variation of PageRank (PR) in which one can specify one or more nodes as query, and obtain for all the other nodes in the graph a score that captures the degree of relatedness against the nodes of interest. This algorithm presents some shortcomings, since it tends to assign a high score to nodes with high global centrality

---

§Work done while at *Eurecat - Centre Tecnològic de Catalunya.*

in the graph, regardless of the query node. For example, in the case of the English Wikipedia, articles such as United States or New York Times (nodes that are globally central) tend to appear among the most relevant topics for any query, even when they are not specifically relevant to the query node.

To overcome these limitations, we have developed an algorithm, called CYCLERANK, that leverages the cycles that exist in a directed graph to compute a relevance score for the other nodes in relation to a query node selected by the user [2].

In our web demo, users can compute personalized relevance scores on a set of example graphs including Wikilink networks from different language editions, networks of interaction on Twitter about specific topics, and networks of co-purchased products from Amazon. Users can compare the results obtained with different algorithms: Cyclerank, PageRank, Personalized PageRank and several variations that have been introduced in the literature such as CheiRank and 2DRank.

The paper is organized as follows. In Section II, we describe the algorithms included in our demo, with a focus on CycleRank, our proposal. Section III shows the architecture of our system and its implementation. Section IV-C compares the results obtained by Cyclerank with those coming from PageRank and various other algorithms, with examples covering a set of diverse topics. Finally, in Section V we present some directions for future work and we draw our conclusions.

II. Overview of the Algorithms

In the following section, we outline briefly the algorithms that are showcased in this demo: *Cyclerank, PageRank, Personalized PageRank, CheiRank* and *2DRank*.

PageRank. PageRank [3] is a metric based on incoming connections, where connections from relevant nodes are given a higher weight. Intuitively, the PageRank score of a node represents the probability that, following a random path in the network, one will reach that node. PageRank can be computed in an iterative process, as the score of a node depends on the score of the nodes that link to it, however

more efficient algorithms are available. The idea at the basis of PageRank is that of simulating a stochastic process in which a user follows random paths in a hyperlink graph. From each node, the algorithm assumes equal probabilities of following any hyperlink included in the page and a certain probability of "teleporting" to another random page in the graph. The damping factor α, generally assumed to be 0.85, defines the probability of continuing surfing the graph versus teleporting.

Personalized PageRank. Personalized PageRank [3] is a variant of the original PageRank algorithm, in which teleporting is not directed to all nodes randomly, but to a specific node or set of nodes. In this way, the algorithm models the relevance of nodes around the selected set of reference nodes, as the probability of reaching each of them, when following random walks starting from this chosen set.

Cheirank. Chepelianskii [4] introduced the idea of computing the PageRank score of nodes on the transposed graph and called the algorithm CheiRank. It can be seen as a kind of PageRank based on outgoing instead of incoming connections.

2DRank. Zhirov [5] combined CheiRank and PageRank to produce a single two-dimensional ranking of Wikipedia articles, 2DRank. This method does not assign a score to each node, but just produces a ranking.

Cyclerank. The goal of Cyclerank is to assign to all the nodes in a graph a score that measures their relevance to a given reference node provided as input. Intuitively, a node that is linked from the reference node, but does not link back to it, is likely not to be related to that subject, even if it is relevant in general. Conversely, a node that links to the reference node, but is not linked from it, is likely to be related, but not relevant. Nodes that are linked both from and to a reference node are the ones that we expect to be both relevant and related to it.

We want then to be able to quantify this kind of "mutual relevance" accounting also for the indirect links, i.e., for the number of paths that can be found linking a node from and to the reference node. We do this by counting the number of cycles of various lengths that contain the reference node and any other node. As short distances represent a stronger relationship, short cycles receive a higher weight.

Given a directed graph $G(V,E)$, a reference node $r \in V$ and an integer $K > 1$, the Cyclerank score of any node $i \in V$ with respect to r is given by:

$$CR_{r,K}(i) = \sum_{n=2}^{K} \sigma(n) \cdot c_{r,n}(i) = \sum_{n=2}^{K} \frac{c_{r,n}(i)}{e^n} \qquad (1)$$

where $c_{r,n}(i)$ is the number of cycles of length n that contain nodes i and r, K is a parameter representing the maximum length considered for cycles, and $\sigma(n)$ is the general form of a scoring function that weights the score assigned for each cycle. We set it to be $\sigma = e^{-n}$.

In this way, given a reference node, the Cyclerank score of a node i represents the number of cycles including both the reference node r and node i, multiplied by a factor depending
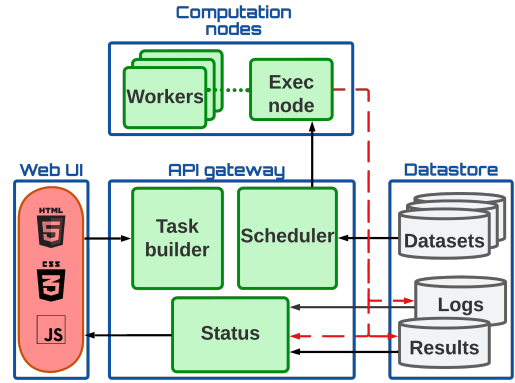

Fig. 1: Architecture of the demonstration

on the chosen scoring function and the length of each cycle. For example, for Wikipedia we have experimentally found that the best choice for the scoring function is an exponential damping $\sigma = e^{-n}$. By definition, the reference node gets the maximum Cyclerank score as it is included in all the cycles considered.

III. SYSTEM ARCHITECTURE

Our system consists of four components, which are presented in the diagram of Figure 1:

- The *Datastore* is responsible for storing and managing datasets. It also provides storage for results and logs produced by the system.
- The *API gateway* acts as a mediator between the computational nodes and the web user interface (Web UI). It acts as entry point for all incoming requests from the Web UI and routes them to the relevant computational nodes. Additionally, it could provide security, authentication, and authorization features for all requests to the system.
- The *Computational nodes* are responsible for processing data requests and can be scaled up or down depending on the system's workload. They interact with the data stores to retrieve or store data and then return the results to the API gateway.
- The *Web UI* is responsible for collecting user input and presenting the computation results to the user. It interacts with the API gateway to send requests to the computational nodes and display results to the user. The user interface is designed to be user-friendly and responsive to different devices and screen sizes.

Each component is containerized to provide isolation and to simplify development and deployment. Both the API gateway and the computational node (workers) are developed as microservices.

The user interacts with the system through the Web UI. When a request is made the following process is started:

1) a task, which is a triple consisting of a dataset, an algorithm and a set of parameters is built by the Task Builder component and is sent to the Scheduler;
2) when the Scheduler receives the task, it fetches the dataset and invokes an Executor node;

3) the computation needed to perform the task is off-loaded to the worker nodes; while the computation is running, the Status component polls the Executor node to monitor its progress;

4) when the computation is completed, results and logs are written to the datastore; the Status component can access them in response to user requests;

5) finally, the API returns the results of the completed task to the Web UI which displays them to the user.

Our demo design enables the possibility of adding new algorithms to the demo. We do not offer this possibility directly because it would expose us to several security issues that are outside the scope of this demo.

IV. RESULTS AND DISCUSSION

In this section we present the demo resources and interface, and how to build queries. Finally, we showcase some use cases and discuss them through concrete examples.

A. Resources

The demo can be accessed at https://cricca.disi.unitn.it/cyclerank-demo/, the code is available on GitHub at https://github.com/CycleRank/cyclerank-demo.

B. Datasets

In principle, all algorithms showcased can be run on any directed graph. We provide these datasets because they present interesting use cases for the problem of personalized relevance in graphs:

- *WikiLinkGraphs* [6]: networks of links between Wikipedia articles (wikilinks). Yearly snapshots for 9 languages (`de`, `en`, `es`, `fr`, `it`, `nl`, `pl`, `ru`, and `sv`) for the years 2018, 2013, 2008, and 2003.
- *Amazon* [7]: product ids and metadata about $548,552$ different products (Books, music CDs, DVDs) that were co-purchased ("Customers who bought X also bought Y") on Amazon.
- Twitter: networks of users that made some tweet containing a set of keywords of interest. Each user is a node and two users are connected if they have interacted in some way (retweet, reply, quote or mention). In particular we provide two datasets from Twitter: `cop27` (COP27 climate conference); and `8m` (8th of March, International Women's Day).

Users can upload their own dataset to run the provided algorithms on their dataset. We support commonly used graph formats such as: `edgelist` (CSV)[1], `pajek`[2], and our own `ASD` format. We detail the supported formats in the Instructions page of the demo.

[1]A comma-separated value with source and target indexes, see: https://gephi.org/users/supported-graph-formats/csv-format/.

[2]A format where each line is an element, and the list of edges follows the list of nodes, see https://gephi.org/users/supported-graph-formats/pajek-net-format/.

C. Usage

The Web UI enables users to select the dataset and the algorithm to execute, along with its parameters. For Cyclerank, these parameters include a reference node r, the maximum cycle length K, and the scoring function σ, as defined in Equation 1. For PageRank, CheiRank and 2DRank, it is sufficient to specify the value of the transition probability α, while their personalized variants also take a reference node r.

Fig. 2: Task builder interface

The user can create arbitrary query sets by choosing different reference nodes, datasets and parameters, which are then visualized in the task builder interface, as presented in Figure 2. A query set defines a task, and a unique identifier is assigned to it, serving as a permalink to retrieve its results. A query set can be emptied by clicking on the trash bin symbol (🗑) on the top right, or queries can be eliminated individually (⊠).

D. Use cases

We present here two use cases for our demonstration: algorithm comparison and dataset comparison.

Algorithm comparison. Our demo can showcase the results of all the presented algorithms allowing users to compare the strenghts and limitations of each algorithm. Tables I and II present a comparison between the top-5 results with the highest scores obtained by PageRank, Cyclerank and Personalized PageRank, computed with different parameters on the wikilink graph of the English Wikipedia taken as of March 1st, 2018 and the Amazon co-purchase graph, respectively.

These results highlight the limitation of Personalized Page-Rank: in Table I, Personalized PageRank promotes results that have a very high in-degree and have among the highest values of the PageRank score in the overall network. This problem has been studied in depth in [2]. Similarly, on the Amazon co-purchase graph in Table II, Personalized PageRank suggest popular items such as the *"Harry Potter"* book series, while Cyclerank does not.

Dataset comparison. Another use case for our demo consists in applying the same algorithm on different datasets to gain new insights. Table III presents a cross-cultural comparison between the relevant nodes in the Wikipedia link graph identified by Cyclerank starting from the article *"Fake news"* for different editions, showing how the same concept is framed differently across different language communities. A similar analysis can also be performed by comparing snapshots of a graph at different points in time, another functionality available in the demo.

TABLE I: Top-5 articles with the highest PR ($\alpha = 0.85$), CR ($K = 3, \sigma = e^{-n}$) and PPR ($\alpha = 0.3$) scores computed on the 2018-03-01 English Wikipedia snapshot. The reference articles for CR and PPR are *"Freddie Mercury,"* and *"Pasta"*.

page		Freddy Mercury			Pasta	
#	PageRank	Cyclerank	Pers.PageRank		Cyclerank	Pers.PageRank
1	United States	Freddie Mercury	Freddie Mercury		Pasta	Pasta
2	Animal	Queen (band)	Queen (band)		Italian cuisine	Bolognese sauce
3	Arthropod	Brian May	The FM Tribute Concert		Italy	Carbonara
4	Association football	Roger Taylor	HIV/AIDS		Spaghetti	Durum
5	Insect	John Deacon	Queen II		Flour	Italy

TABLE II: Top-5 articles with the highest PR ($\alpha = 0.85$), CR ($K = 5, \sigma = e^{-n}$), and PPR ($\alpha = 0.85$) scores computed on the Amazon co-purchase dataset. The reference items for CR and PPR are *"1984,"* and *"The Fellowship of the Ring."*

item		1984		The Fellowship of the Ring	
#	PageRank	Cyclerank	Personalized PageRank	Cyclerank	Personalized PageRank
1	Good to Great	Animal Farm	The Catcher in the Rye	The Hobbit	The Silmarillion
2	The Catcher in the Rye	Fahrenheit 451	Lord of the Flies	The Return of the King	The Hobbit
3	DSM-IV	The Catcher in the Rye	Animal Farm	The Silmarillion	Harry Potter (Book 1)
4	The Great Gatsby	Brave New World	Fahrenheit 451	The Two Towers	Harry Potter (Book 2)
5	Lord of the Flies	Lord of the Flies	To Kill a Mockingbird	Unfinished Tales	The Return of the King

TABLE III: Top-5 articles with the highest Cyclerank ($K = 3, \sigma = e^{-n}$) scores computed on different Wikipedia language editions (`de`, `es`, `fr`, `it`, `nl`, `pl`) using the reference article *"Fake news"*.

#	Fake News (de)	Fake news (en)	Fake news (fr)	Fake news (it)	Nepnieuws (nl)	Fake news (pl)
1	Barack Obama	CNN	Ère post-vérité	Disinformazione	Facebook	Dezinformacja
2	Tagesschau.de	Facebook	Donald Trump	Post-verità	Journalistiek	Propaganda
3	Desinformation	US pres. election, 2016	Facebook	Bufala	Hoax	Media społecznościowe
4	Fake	Propaganda	Hoax	Debunker	Donald Trump	-
5	Donald Trump	Social media	Alex Jones (complotiste)	Clickbait	-	-

V. Conclusions and Future Work

In this demonstration, we showcased Cyclerank, a novel algorithm that leverages cyclic paths to identify the most relevant nodes related to a given query node. We also described two use cases:

- *algorithm comparison*: we compared Cyclerank with 6 established algorithms: PageRank, CheiRank, 2DRank and their personalized variants. We highlighted how, on different datasets from Wikipedia and Amazon, Cyclerank could produce more relevant answers and mitigate the issue of over-representation of popular nodes. Our demo design enables the possibility of adding new algorithms to the demo.
- *dataset comparison*: we showed how Cyclerank can be used to compare different datasets, providing insights into the relationships within the data. We presented a cross-cultural comparison of the results obtained on the Wikipedia link graph for the same topic in six different languages. Users can also upload their own datasets. We support three dataset formats and we plan to add new ones in the future.

In summary, our flexible demo design enables the comparison of relevance ranking algorithms on directed graphs.

References

[1] N. Markusson, T. Venturini, D. Laniado, and A. Kaltenbrunner, "Contrasting medium and genre on Wikipedia to open up the dominating definition and classification of geoengineering," *Big Data & Society*, vol. 3, no. 2, p. 2053951716666102, 2016.

[2] C. Consonni, D. Laniado, and A. Montresor, "Cyclerank, or there and back again: personalized relevance scores from cyclic paths on directed graphs," *Proceedings of the Royal Society A*, vol. 476, no. 2241, p. 20190740, 2020.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[4] A. D. Chepelianskii, "Towards physical laws for software architecture," *CoRR*, vol. abs/1003.5455, 2010.

[5] A. O. Zhirov, O. V. Zhirov, and D. L. Shepelyansky, "Two-dimensional ranking of wikipedia articles," *The European Physical Journal B*, vol. 77, pp. 523–531, 2010.

[6] C. Consonni, D. Laniado, and A. Montresor, "Wikilinkgraphs: A complete, longitudinal and multi-language dataset of the wikipedia link networks," *Proceedings of ICWSM*, vol. 13, no. 01, pp. 598–607, 2019.

[7] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, pp. 5–es, 2007.