# Reviewing PageRank for Citation Recommendation

*ITCS 6114: Algorithms and Data Structures*

*Dr. Berardinelli*

David Gary

October 13, 2022

## 1 Introduction

The number of research publications in the field of computer science have grown tremendously in the last decade. While this signals many positives for the field, it has also created an obfuscation obstacle that researchers must overcome when attempting to find related works. As a result, many researchers are developing recommendation systems that automate the process of finding relevant citations, with graph representations as the core for how papers are connected. The current state of this research, unfortunately, has not yet managed to produce a system that can account for the unique nature of developing fields, and common errors in research like overcitation and unrelated citations tend to cause unfair edge weighting and poor recommendations [1]. Still, it is worth learning the current graph based approaches to creating these citation recommendation systems. In this paper, we will discuss a base-level graph representation and algorithm (PageRank [2]) that provides citation recommendations.

## 2 The Problem

Following the works of Kucuktunc et al. [1], it is assumed that the researcher already has a base set of citations created, and the goal of the recommendation system is to find more papers that are related to the base set. As such, let $G = (V, E)$ be a directed graph where each vertex $v \in V$ represents a paper, and each edge $e \in E$ represents a citation relationship between two papers in the form $e = (u, v)$ where $u$ is a citation of $v$. The goal of the recommendation system is to find a set of papers $R \subseteq V$ of a given size $k$ that are related to the base set $B \subseteq V$.

**Problem Statement:** Given a directed graph $G = (V, E)$, where each vertex $v \in V$ represents a paper, and each edge $e \in E$ represents a citation relationship between two papers in the form $e = (u, v)$ where $u$ is a citation of $v$, the goal of the recommendation system is to find a set of papers $R \subseteq V$ of a given size $k$ that are related to the base set $B \subseteq V$.

### 2.1 The PageRank Approach

Let $G' = (V, E')$ be the undirected graph where $E' = E \cup E^{-1}$, where $E^{-1}$ is the set of edges where the direction is reversed. It then follows that a random walk on $G'$, from start position $v \in V$, will have a probability of $1/|V|$ of landing on any vertex $u \in V$. PageRank utilizes a random chance of teleporting to any vertex in the graph, based on a given damping factor $d \in [0, 1]$, where the resulting teleportation probability is $1 - d$. A transition matrix $P$ is created:

$$P_{(u,v)} = \begin{cases} (1-d)\frac{1}{|V|} + d\frac{1}{|N(v)|} & \text{if } (u,v) \in E' \\ (1-d)\frac{1}{|V|} & \text{otherwise} \end{cases} \tag{1}$$

This lets us construct a probability distribution equal to the discrete time evolution equation $p_{t+1} = Pp_t$. A convergence threshold $\epsilon$ is given and the algorithm terminates when $p_{t+1} - p_t < \epsilon$. The resulting probability distribution $p$ is then used to rank the vertices in $V$. To construct our output set $R$, we first find the top $k$ vertices in $p_t$ that are not in $B$. Then, we find the top $k$ vertices in $p_t$ that are not in $B$ and $R$; the ranking of these vertices is determined by the sum of the probabilities of the vertices in $R$ that are connected to them. This is repeated until $|R| = k$.

### 2.2 Inputs and Outputs

The table below shows the expected inputs and outputs of the algorithm. It should be noted, however, that the method of weighting an edge is not specified for this problem. Edge weights can factor in any number of desired features. PageRank is more concerned with the structure of the graph and the probability of a random walk, and does not take into account the content of the papers. These features can be added in later stages of the algorithm, but they will be highly dependent on the components of the input graph and base citation sets. As an example, most algorithms from the Kucuktunc et al. [1] paper used metadata information to build out the vertices, but this metadata exclusively contained citation information. If one were to weight their algorithm by word count similarities in the abstracts, then the input graph would need to be constructed with the abstracts as a component of the vertices.

| **Input** | **Description** |
|---|---|
| $G = (V, E)$ | the directed graph |
| $B \subseteq V$ | the base set of citations |
| $k$ | the number recommendations |
| $d \in [0, 1]$ | the damping factor |
| $\epsilon$ | the convergence threshold |
| **Output** | **Description** |
| $R \subseteq V$ | the set of recommended citations |

### 2.3 Algorithmic Form

**Algorithm 1** PageRank($G, B, k, d, \epsilon$)

$G' = (V, E')$
$E' = E \cup E^{-1}$
$P = \begin{bmatrix} (1-d)\frac{1}{|V|} + d\frac{1}{|N(v)|} & \text{if } (u,v) \in E' \\ (1-d)\frac{1}{|V|} & \text{otherwise} \end{bmatrix}$
$p_1 = \begin{bmatrix} \frac{1}{|V|} \end{bmatrix}$
$t = 1$
**while** $p_{t+1} - p_t \geq \epsilon$ **do**
  $p_{t+1} = Pp_t$
  $t = t + 1$
**end while**
$R = \emptyset$
**while** $|R| < k$ **do**
  $R = R \cup \{v \in V \mid v \notin B \wedge v \notin R\}$
  $R = R \cup \{v \in V \mid v \notin B \wedge v \notin R \wedge \sum_{u \in R} p_t(u) \geq \sum_{u \in R} p_t(v)\}$
**end while**
**return** $R$

# 3 Conclusion

As shown above, PageRank defines the citation recommendation problem in a highly modular and intuitive graph representation. Using the method above, including more specified approaches to edge weighting and teleportation is as easy as changing a few variables in the algorithm. For instance, say that each $v$ contains information about the date a paper was published, and I wanted to filter out papers that were published after a certain date $D$. I could simply change the transition matrix $P$ to be:

$$P_{(u,v)} = \begin{cases} (1-d)\frac{1}{|V|} + d\frac{1}{|N(v)|} & \text{if } (u,v) \in E' \wedge date(v) < D \\ (1-d)\frac{1}{|V|} & \text{otherwise} \end{cases} \tag{2}$$

The adaptability of this algorithm is what has made it the base behind countless graph-based citation recommendation systems.

# References

[1] Küçüktunç, O., Saule, E., Kaya, K., and Çatalyürek, U. V. Diversifying citation recommendations. *ACM Trans. Intell. Syst. Technol. 5*, 4 (dec 2014).

[2] Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.