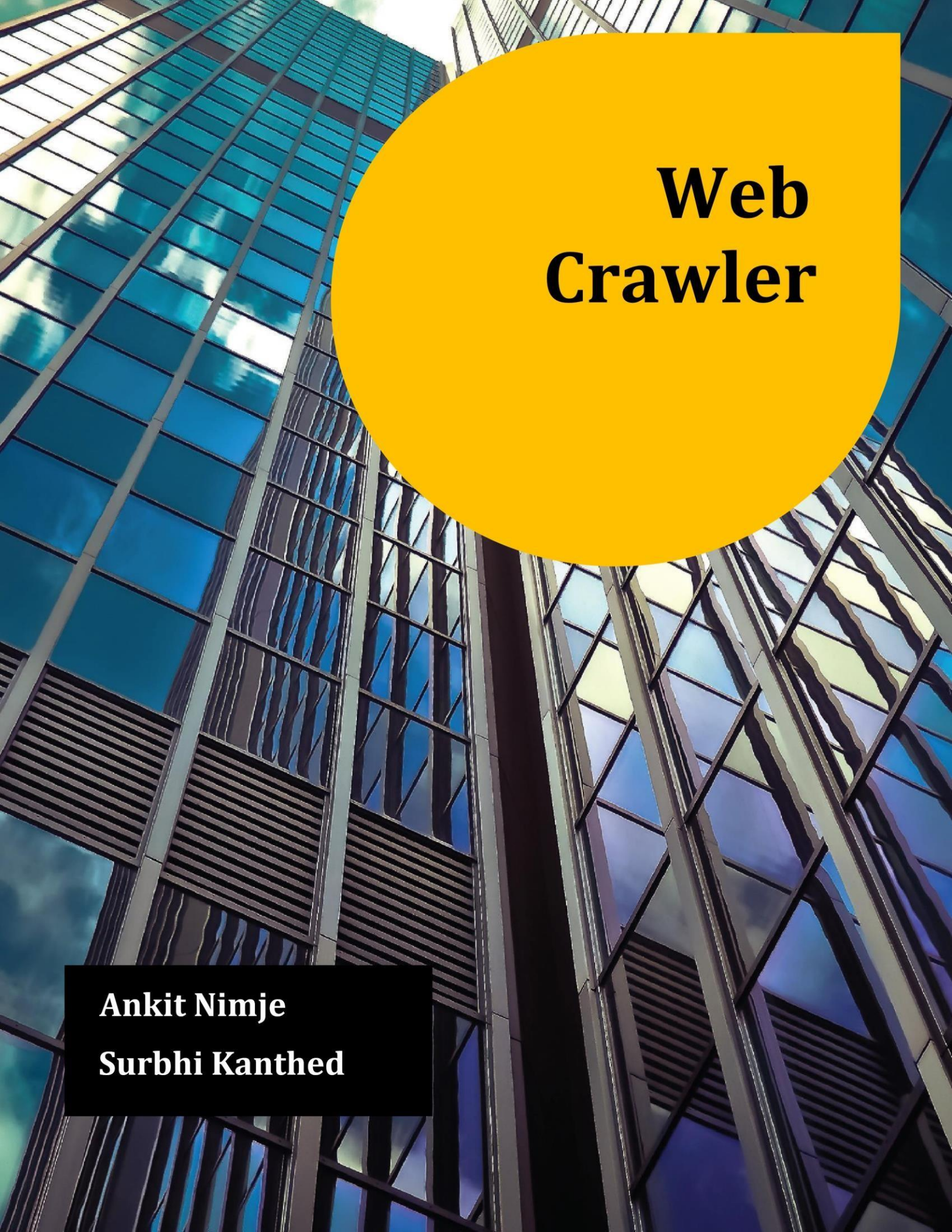

Web Crawler

Ankit Nimje

Surbhi Kanthed

In this week we were working on Page Rank Algorithm. Page Rank is used by Google Search Engine and also named after Larry Page, one of the founders of Google. It is a way of measuring importance of web pages. According to Google, "*PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites."*

Page Rank Algorithm is based on following formula:

$$PR_{t+1}(P_i) = \sum (PR_t(P_j) / C(P_j))$$

Where,

t → number of iterations

P_j → Page rank from previous iteration

P_i → Page rank of current iteration

C → Number of nodes pointing to current node

Page Rank Algorithm works as a series of iterations with a method defined as pageIterations() with parameters such as:

- G → a directed graph (all undirected graphs will be converted to directed graphs before passing)
- Alpha → damping parameter for page rank
- Maximum Iterations → Number of iterations after which pageIterations() will be terminated
- nStart → Starting value of Page Rank for each node.
- Weight → Weight of edge
- Converge → To check convergence of iterations

Thus, using PR algorithm after all iterations, we get a score for each web page. Higher the score higher the page rank. For instance, to make your webpage display on first few pages of Google Search you should consider provide links to your webpage on many other relevant web pages.

References:

1] S. Chakrabarti, M. Berg, and B. Dom," Focused crawling: a new approach to topic specific Web resource discovery", Computer Networks, vol. 31, pp. 1623-1640, 1999.

2] M. Najork and J. Wiener," Breadth-first search crawling yields high-quality pages", in Proc. 10thInt. World Wide Web Conf., pp. 114-118, 2001.

3] Page Rank Algorithm and Implementation – Geeks for Geeks