3.1 Introduction to PageRank

PageRank (PR) is an algorithm used by Google Search to rank web pages in their search engine results. It is named after the term "web page" and co-founder Larry Page. PageRank is a way of measuring the importance of website pages.

PageRank counts the number and quality of links to a page to determine a rough estimate of the website's importance. The underlying assumption is that more important websites will likely receive more links from other websites.

You may spend around 20 minutes to learn the introduction of PageRank from the following article.

Key reading

Verzhbitskaia, Z. (2021). The past, present & future of Google PageRank. https://www.link-assistant.com/news/google-pagerank-algorithm.html

3.1.1 How PageRank works

PageRank uses the concept from Graph (in-degree and out-degree) and performs the calculation to assign weight to every page.

Watch the following 5-minute 16-second video to learn how Google's PageRank algorithm works.

How Google's PageRank Algorithm Works

https://youtu.be/meonLcN7LD4

Source: (Spanning Tree, 2020)

The PageRank Algorithm: Idea

- The PageRank algorithm gives each page a rating of importance, a recursively defined measure whereby a page becomes important if essential pages link to it. This definition is recursive because a page's importance refers to the importance of other pages that link to it.
- One way to think about PageRank is to imagine a random surfer on the web, following links from page to page. Any page's page rank is roughly the

probability that the random surfer will land on a particular page. Since more links go to the critical pages, the surfer is more likely to end up there.

- The behaviour of the random surfer is an example of a Markov process, which is any random evolutionary process that depends only on the current state of a system and not on its history.

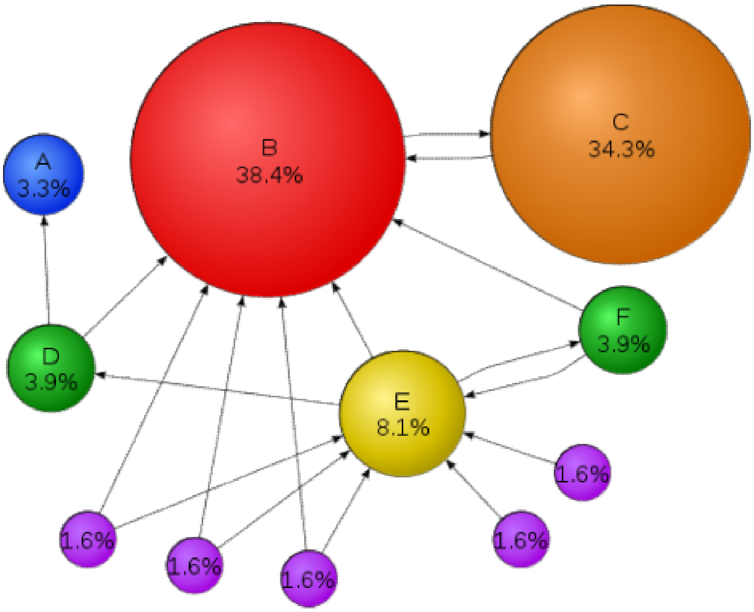
Example for PageRank

Mathematical PageRanks for a simple network is expressed as percentages. (Google uses a logarithmic scale.)

Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from a significant page and is highly valued.

If web surfers who start on a random page have an 82.5% likelihood of choosing a random link from the page they are currently visiting and a 17.5% likelihood of jumping to a page chosen at random from the entire web, they will reach Page E 8.1% of the time. (The 17.5% likelihood of jumping to an arbitrary page corresponds to a damping factor of 82.5%.)

Without **damping**, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages on the web, even though it has no outgoing links of its own.



Source: An illustration of mathematical PageRanks for a simple network (Wikipedia, n.d.)

Further reading

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems 30.1-7 (pp. 107-117). http://infolab.stanford.edu/pub/papers/google.pdf

3.1.2 PageRank algorithm

What is PageRank algorithm?

Watch the following 10-minute 10-second video to learn more.

PageRank Algorithm - Example

https://youtu.be/P8Kt6Abq_rM

Source: (Global Software Support, 2017)

You may continue to watch the following 14-minute 43-second video to learn more about Network Graphs and PageRank algorithm.

Network Graphs and Page Rank Algorithm

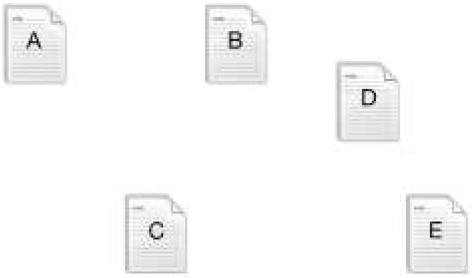
https://youtu.be/ztc6sYgapwA

Source: (Barry Van Veen, 2020)

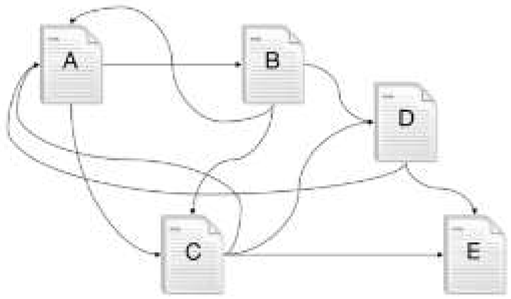
3.1.3 PageRank Steps

Read the steps of a short working example of PageRank:

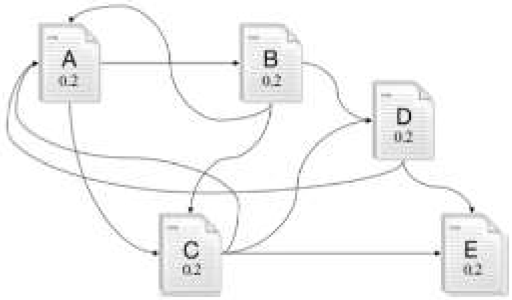
1. Start with a set of pages



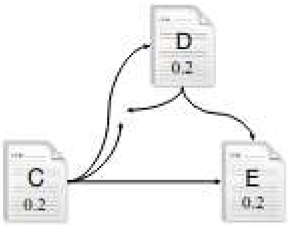
2. Crawl the web to determine the link structure.



3. Assign each page an initial rank of 1/N



4. Successively update each page's rank by adding the weight of every page that links to it divided by the number of links emanating from the referring page.



- In the current example, page E has two incoming links, one from page C and one from page D .
- Page C contributes 1/3 of its current page rank to page E because E is one of three links from page C . Similarly, page D offers 1/2 of its rank to E.
- The new page rank for E is 0.17

$$\mathrm{PR(E)} = \frac{\mathrm{PR(C)}}{3} + \frac{\mathrm{PR(D)}}{2}$$

$$\mathrm{PR(E)} = \frac{0.2}{3} + \frac{0.2}{2} = 0.17$$

Read the following source materials of the content above to gain a more comprehensive understanding of PageRank algorithm:

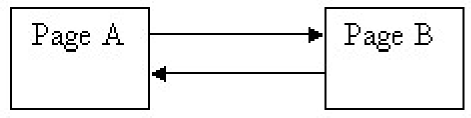
Key readings

1. Roberts, E. (2015). The Google PageRank algorithm [handout]. https://web.stanford.edu/class/cs54n/handouts/24-GooglePageRankAlgorithm.pdf
2. Amine, A. (2021). PageRank algorithm, fully explained. Towards Data Science. https://towardsdatascience.com/pagerank-algorithm-fully-explained-dc794184b4af

3.1.4 Application Activity: Calculating PageRank

Question

In this section, you have learnt what PageRank is and how PageRank for each page is calculated. You can practice calculation by using the following example. Suppose there are only two pages: A & B, and they are just pointing at each other as follows:



As given the outgoing count is 1 for both i.e. C(A)=C(B)=1 and let d = 0.85.
Now, assuming initial PR as 1, 0 and 40 and with a five iteration, calculate the PageRank of A and B for each iteration with the following formula.

$$\mathrm{PR(A)} = (1 - \mathrm{d}) + \mathrm{d}\left(\frac{\mathrm{PR(B)}}{1}\right)$$

$$\mathrm{PR(B)} = (1 - \mathrm{d}) + \mathrm{d}\left(\frac{\mathrm{PR(A)}}{1}\right)$$

You may refer to the article below to help you find the solution.

Reading

Rogers, I. (n.d.). Page Rank explained: The Google Pagerank algorithm and how it works.

https://www.khoury.northeastern.edu/home/vip/teach/IRcourse/4_webgraph/notes/Pagerank%20Explained%20Correctly%20with%20Examples.html

3.2 Applications of PageRank and its impact

In the previous section, you learn about PageRank and how it works. The discussion was focused on web page ranking. The PageRank algorithm was designed for ranking web pages to show the optimized search result. However, over the years, the popularity and usability expanded to other domains which seek to achieve similar tasks but different problem definitions. In this section, we have collected some popular extensions or adaptations of PageRank to solve problems of domains other than web page ranking.

Some of the key variations of PageRank are:

- Sports
- Literature
- Digital health care
- PageRank in Environment management
- Software Development (Debugging, version control etc.)
- Smart City Management (Road traffic management, urban planning etc.)

Reading

Cornell University. (2014). More than just a web search algorithm: Google's PageRank in non-internet contexts. Networks - Course blog for INFO 2040/CS 2850/Econ 2040/SOC 2090. https://blogs.cornell.edu/info2040/2014/11/03/more-than-just-a-web-search-algorithm-googles-pagerank-in-non-internet-contexts/

3.2.1 PageRank in Sports

The application of PageRank in sports is exciting and helps to find many interesting facts about a particular game. For example, using networks of football teams and tennis players, researchers were able to find the best teams and athletes (Jimmy

Connors was returned in the top spot for tennis players). Following are two examples of the different applications of PageRank in sport.

Ranking Football Team

In this, the application of PageRank is to rank football teams via the GEM method. The authors modified and extended the GEM method with more football statistics to look at the possibility of using this method to rank teams more accurately.

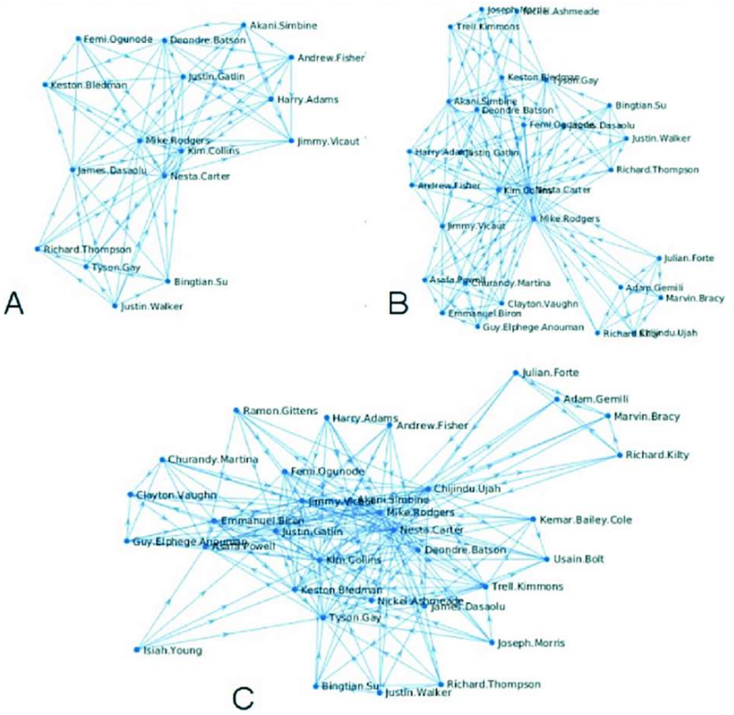
Who beats who!: Ranking Athletes

In this work, the authors utilised a simple who beat who' matrix; the PageRank (PR) and user preference (UP) algorithms can accurately rank track athletes, avoiding the need for subjective assessment.

The authors applied the PR and UP algorithms to the 2015 IAAF Diamond League men’s 100m competition and compared their performance with the Keener, Colley and Massey ranking algorithms.

The top five places computed by the PR and UP algorithms and the Diamond League ‘2016’ points system were all identical, with Kendall’s tau distance between the PR standings and ‘2016’ points system standings being just 15, indicating that only 5.9% of pairs differed in their order between these two lists.

By comparison, the UP and ‘2016’ standings displayed a less strong relationship, with a tau distance of 95, indicating that 37.6% of the pairs differed in their order.



Source: Connectivity networks between athletes after (A) the first three races; (B) the first six races; and (C) all ten races (Beggs, Clive B. , 2017)

Read the sources below to further understand this topic.

Key readings

1. Zack, L., Lamb, R., & Ball, S. (2012). An application of Google's PageRank to NFL rankings. Involve, a Journal of Mathematics, 5(4), 463-471. https://projecteuclid.org/journals/involve-a-journal-of-mathematics/volume-5/issue-4/An-application-of-Googles-PageRank-to-NFL-rankings/10.2140/involve.2012.5.463.full
2. Beggs, C.B., Shepherd, S.J., Emmonds, S., & Jones, B. (2017). A novel application of PageRank and user preference algorithms for assessing the relative performance of track athletes in competition. PLoS ONE, 12(6), 1–26. DOI: 10.1371/journal.pone.0178458. https://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=asn&AN=123384782&scope=site&custid=s7320401

3.2.2 PageRank in Literature

PageRank also found application in literature. In an interesting work, the authors use a network of 19th-century authors to find quantitative evidence that Jane Austin and Walter Scott were found to be the most original authors of the 19th century. In another work, authors applied PageRank to rank the knowledge material to improve collaborative learning.

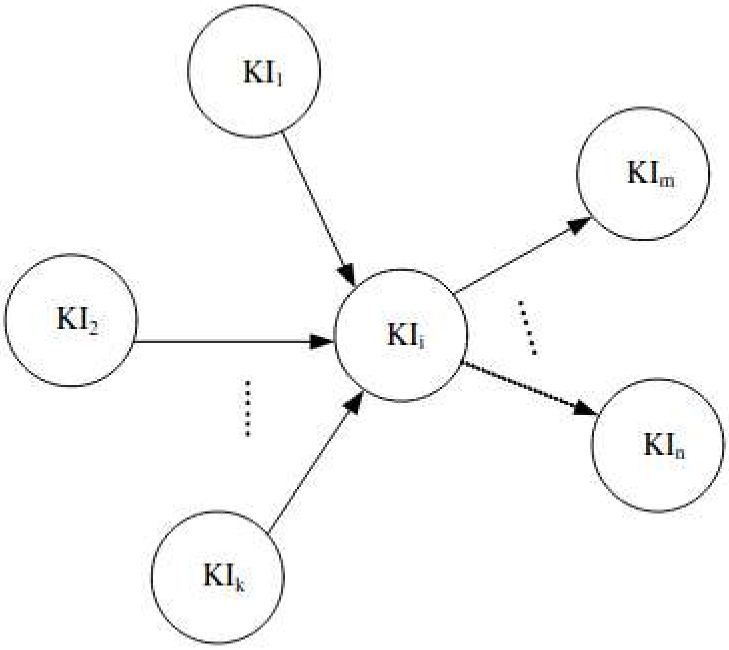
Rank knowledge items: Collaborative Learning

With the rapid development of web 2.0, many realm communities provide free platforms for users to enrich their knowledge through online communication, sharing and socializing without boundaries. As an online system may interact with thousands of users, it is almost impossible for field experts or teachers to give instant help manually, which is inefficient and laborious. An e-learning community should construct an efficient knowledge-acquiring mechanism to cope with it.

Researchers apply a **PageRank-based mechanism** to rank knowledge items synthetically to ensure this mechanism.

The system appraises the knowledge items provided by learners based on their rank, other users' remarks and most importantly, teachers' and realm experts' remarks, thus picking out the KIs to the knowledge base.

The users' grades will be upgraded or degraded by their KIs. Learners are served with the knowledge that best matches their needs and encouraged by each other.



Source: Correlativity model of KE (Yang, Zhao, Zhang & Zhao, 2008)

Key reading

Sangers, A., Heesch, M. V., Attema, T., Veugen, T., Wiggerman, M., Veldsink, J., Bloemen, O., & Worm, D. (2019). Secure multiparty PageRank algorithm for collaborative fraud detection. Financial Cryptography and Data Security, 605–623. DOI: 10.1007/978-3-030-32101-7_35.
https://www.academia.edu/download/80741125/29073.pdf

3.2.3 PageRank in Digital health care

Digital health care is an emerging field, and there are lots of advancements in the field. Recently, researchers used fMRI scans to generate a network where the nodes are voxels on the fMRI scan, and edges between nodes represent that the voxels are strongly time-correlated.

A version of PageRank designed for undirected graphs, neuroscientists were able to identify parts of the brain that change together as subjects aged.

There are many other applications of PageRank for digital health, you can read more from the following reading.

Reading

Gleich, D.F. (2014). PageRank beyond the Web. SIAM Review, 57(3), 321-363.
https://doi.org/10.48550/arXiv.1407.5107

3.2.4 PageRank in Environment management

Environment management has many problems related to ranking or finding the essential resources etc. Researchers have used PageRank for some of the problems related to the environment. For example, the following example is the application of the PageRank algorithm for toxic waste management. R

Waste management

Scientists were able to use PageRank to help determine the position of water molecules in an ionic solution, enabling them to find the best ways to remove nuclear waste and toxic chemicals.

According to Aurora Clark, an associate professor at WSU, once you know the probable positions of different molecules in the solution, “…you can control the chemistry and force certain reactions to occur.”

PageRank maps where toxic chemicals are likely to pool in the solution, enabling a waste cleanup team to quickly and efficiently contain and remove the toxic or radioactive contaminant. [Source: Reading 1]

There are other uses of PageRank for different applications in environmental management that you can read from the mentioned readings.

Readings

1. Garling, C. (2012). Researchers fight toxic waste with Google PageRank. WIRED. https://www.wired.com/2012/02/google-pagerank-water/
2. Gleich, D.F. (2014). PageRank beyond the Web. SIAM Review, 57(3), 321-363. https://doi.org/10.48550/arXiv.1407.5107

3.2.5 PageRank in Software development

The application of PageRank in Software development is continuing with its first use as ranking search engine results. Over the years, the PageRank algorithm has been used to solve various problems in software development. In this section, two interesting applications are discussed, debugging and GIt commits evaluation.

Debugging

MonitorRank is a version of PageRank designed to analyze complex, engineered systems.

The MonitorRank can reduce the time, domain knowledge, and human effort required to find
the root causes of anomalies in such service-oriented architectures.

The algorithm “returns a ranked list of systems based on the likelihood that they contributed to, or participated in, an anomalous situation.”

In other words, MonitorRank is a debugging tool like no other – instead of crawling through error pages and debugging callbacks, it analyzes the structure of the buggy system itself to suggest possible and probable causes of error.

MonitorRank uses each sensor's historical and current time-series metrics as its input, along with the call graph generated between sensors, to build an unsupervised model for ranking. [Source: Reading 3]

Version Control: Commits Evaluation [Source: Reading 1]

In version control systems, a commit is an operation which sends the latest changes to the source code to the repository, making these changes part of the head revision of the repository.

There can be many purposes for making a commit. However, when a commit aims to fix a bug, there can be another preceding commit which makes a reason for the bug fixing. A bug-fixing-based causal relationship links those commits. These commits can be modelled as a directed graph model of causal relationships.

Researchers have used Google's PageRank algorithm in the graph modeling order to evaluate commits' influences on the others.

Through an empirical study with Git repositories of six open source projects, the following factors are shown to be noteworthy:
(1) the number of added files at the commit,
(2) the length of the commit message,
(3) the experience of committing author, and
(4) the number of developers involved in the modified files at the commit.

Readings

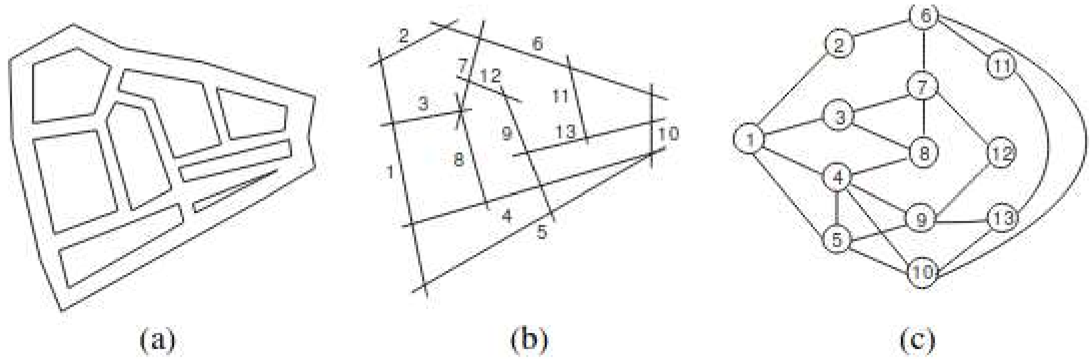
1. Gleich, D.F. (2014). PageRank beyond the Web. SIAM Review, 57(3), 321-363. https://doi.org/10.48550/arXiv.1407.5107
2. Kim, M., Sumbaly, R., & Shah, S. (2013). Root cause detection in a service-oriented architecture. ACM SIGMETRICS Performance Evaluation Review, 41(1), pp 93-104. https://doi.org/10.1145/2494232.2465753
3. Suzuki, S., Aman, H., Amasaki, S., Yokogawa, T., & Kawahara, M. (2017). An application of the PageRank algorithm to commit evaluation on git repository. In 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 380-383. IEEE. https://research.ebsco.com/linkprocessor/plink?id=ed973975-4280-3147-8c79-60708342 64c5

3.2.6 PageRank in SmartCity

The smart city is an evolving concept of a more organized and more comfortable for humans. Many problems need optimized and efficient solutions. Due to multiple entities (for example, bus stops, a network of homes etc.) interaction in smart cities, many of these problems can be represented as graphs so that PageRank can be applied. In this section, we will learn about two applications: 1) Road and foot traffic prediction and 2) Production network ranking.

Predicting road and foot traffic in urban spaces

PageRank has been found to accurately predict traffic flow on individual roads, and connected road maps are represented as graphs, where nodes are streets and intersections are edges (Pop & Dobre, 2012) . Furthermore, PageRank has also been found to accurately reflect observed human mobility through urban spaces (Jiang, 2009).



Source: A fictive urban system (a), its axial map (b) and connectivity graph (c) (Jiang, 2009)

Ranking locations of a production network

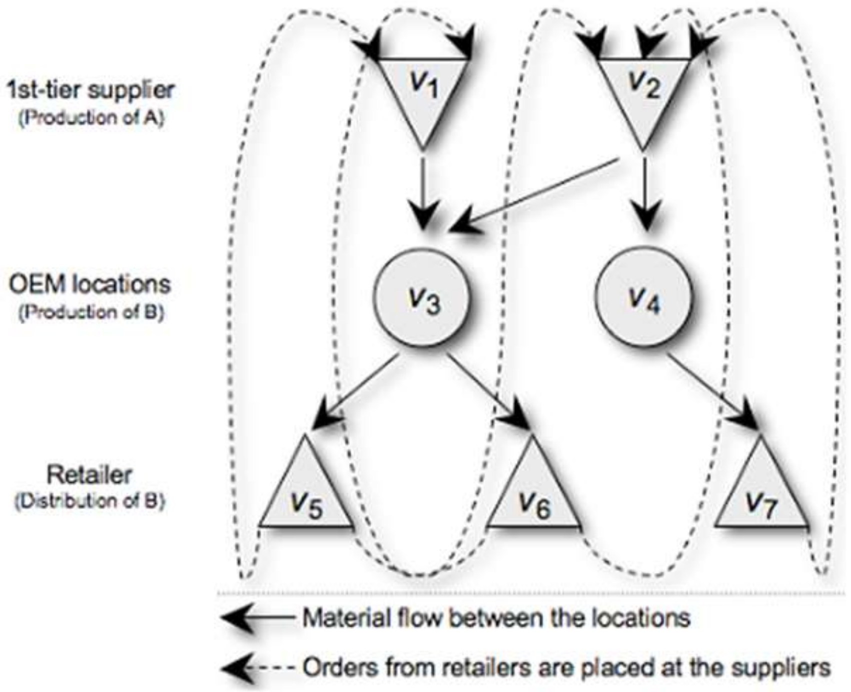
To investigate the dynamics of large-scale production networks, it is essential to derive representative models of lower size.

Against this background, two questions occur:

1. How to identify locations that might be neglected?
2. How do we identify locations that are highly important for the original network?

To answer these two questions, researchers have presented an approach to determine the relative importance of locations within one production network. The approach is based on an adaptation of the PageRank algorithm used by Google.

The algorithm considers both the network structure and the intensity of the material flows. Furthermore, the approach can take changes in material flows over time into account. The following figure shows the mapping of the production network. More details can be learned from the readings.



Source: Production network as graph (Scholz-Reiter et al., 2009)

Key readings

1. Jiang, B. (2009). Ranking spaces for predicting human movement in an urban environment. International Journal of Geographical Information Science, 23(7), 823-837. https://arxiv.org/pdf/physics/0612011.pdf
2. Pop, F. & Dobre, C. (2012). An efficient PageRank approach for urban traffic optimization. Mathematical Problems in Engineering, 2012, 1-9. https://research.ebsco.com/linkprocessor/plink?id=f2d00beb-9bea-3b1d-b862-7d70ad08cdf9
3. Scholz-Reiter, B., Wirth, F., Dashkovskiy, S., Makuschewitz, T., Kosmykov, M., & Schönlein, M. (2009, June). Application of the PageRank algorithm for ranking locations of a production network. In Proceedings of 42nd CIRP conference on manufacturing systems, 3-5. http://www.math.uni-bremen.de/zetem/cms/media.php/256/2009-CIRP-RWDMKS.pdf

3.2.7 Discussion: Applications of the PageRank algorithm

In this section, you were exposed to many applications of the PageRank algorithm other than the initial web page ranking. The application list is incomplete, and there are many more adaptions of the PageRank algorithm.

Time: 60 minutes

Purpose: To explore more applications of the PageRank algorithm.

Task: In this activity, you will need to find two more applications of PageRank and provide a summary of both applications. As submission of this activity, you must provide:

1. Title of the application with reference paper.
2. A summary of the method.
3. Your point of view and any future directions.

3.3 PageRank_implementation

Refer to the article below and try out the exercise.

Reference

GeeksforGeeks. (2022). Page Rank algorithm and implementation. https://www.geeksforgeeks.org/page-rank-algorithm-implementation/