Assessment of ranking algorithms in complex networks

P. WOŁOSZYN

pawel.woloszyn@wat.edu.pl

Military University of Technology, Faculty of Cybernetics
Institute of Computer and Information Systems
Kaliskiego Str. 2, 00-908 Warsaw, Poland

A particularly helpful search of a network such as the Internet or a citation network not only finds nodes that satisfy some criteria but also ranks those nodes for importance to create what amounts to a "reading list". In the recent past, there has been a large interest across a number of research communities in the analysis of complex networks. The selected set of pages from the World Wide Web can be modeled as a directed graph, where nodes are designated as individual pages, and the links as a connection between them. As the number of webpages to be ranked is in the billions, the computation is time-consuming and can take several days or more. Algorithms like PageRank, HITS, SALSA and their modifications has a challenge to deal with the size of the processed data. The need for accelerated algorithms is clear. This article presents the characteristics of three best known ranking algorithms and the assumptions for new algorithm development with first test runs.

Keywords: link analysis, web search, HITS algorithm, Kleinberg's HITS algorithm, PageRank algorithm, SALSA algorithm, hubs, authorities.

1. Introduction

Obtaining information from the World Wide Web is a much greater challenge than getting information stored in traditional databases. The Web is rapidly changing structure, each day the parties are added, deleted and modified. Size of the Web can be estimated at over one billion pages, and many of these sites may contain information repeated or false. Search tools in the Web must be able to distinguish pages with quality content from sites with low quality content [4].

The Web contains, in addition to useful knowledge, a huge amount of information noise. Obtaining useful and valuable knowledge of the gigantic repository is a difficult task. Methods of connections exploration analyze the link structure and links between Web documents to develop a ranking of documents. These methods are mainly used in the search engines, but not only, because thanks to them, you can also define the space around the organization.

Simple text search in a large set of Web pages, often returns useless results. The network structure is exposed to a so-called artificial intervention in order to improve the performance rankings in search engines. Good search algorithm and ranking of Web pages should be immune to such actions.

With the amount of information that is constantly increasing due to the widespread use of computers and the Internet, the network information based on the filtering of data by using tools such as ranking algorithms to attract the attention of researchers from various fields [1].

It is expected that a good ranking algorithm should produce an impartial ranking, where both recent and old nodes have the same chance to appear at the top. The networks also arises the problem of evolving new nodes that may have important information. To explain: Recently added nodes will receive only a few links, because their weight is much less than the mass of older nodes that have already accumulated a lot of links. The problem may also be that the old nodes already point to many other nodes and new ones haven't any outgoing calls yet.

Ranking individual elements within the network, including nodes and links allows you to specify the most important subsets, components and priorities of resources, such as finding the most authoritative websites related to the search topic on the Internet, discovering the most influential people in a social network, identify the most cited articles of scientific networks or identification of the most vulnerable components in the infrastructure systems (eg. transport networks, power grids, water systems). Quantitative assessment of

the criticality of network components helps to inform decision-makers about their management strategy. For example, designers of infrastructure can define the safety objectives or levels of reliability elements (for example, roads and bridges) in the transport network. Therefore, the aim is to effective and rapid methods of calculation, because the classical methods are too computationally demanding.

Ranking of nodes and links in the network connects to the question "Is it important nodes or links in the network are reasonably identified while maintaining a low amount of input data and computing resources?". An example would be one of the easiest methods of indicating the importance of a network node – the level of the apex denoting the number of calls connected to it. Adjacency matrix is used to determine the patterns of connections of nodes and links in the network. Adjacency matrix has also been used as input for advanced eigenvalue and eigenvector analysis, which are still used in practice.

One of the most successful solutions in the calculation of the ranking in a complex network algorithm is PageRank (Brin, Page), which is the basis for the ranking of web pages at Google, in which the parties are nodes and links are the connections between them [7].

Another popular algorithm ranking is Hypertext Induced Topic Selection (HITS), developed by J.Kleinberg. HITS algorithm defines two types of nodes in the network, hubs and authorities, the result ranking is calculated with them in a mutually supportive. However, both the PageRank and HITS are limited to ranking network nodes and are not of substantial importance links [4].

Rankings of research for the Internet and network citations are mainly focused on the ranking of nodes, because connections in such networks do not have a practical application. This is in sharp contrast to some types of networks, such as transport networks, whose links are at least as important as the nodes and the supply chain network, which links are important to find the flow of the product or service to the end customer.

That is why we need new tools, computationally efficient, in particular, on the basis of network analysis for the common ranking of all nodes.

2. PageRank algorithm

The selected set of pages from the Web can be modeled as a directed graph, where nodes are designated as individual pages, and the links as a connections between them.

The theory of algorithms, link analysis assumes that the hierarchy of the validity of pages and documents is based on the structure of their connections. Important pages and documents have more incoming links. Also relevant is the fact that incoming links from relevant sites are important in calculating rank.

PageRank is probably the most popular ranking alghoritm, that have been implemented in real systems, the ranking information, infrastructure, etc. Despite its unique popularity and wide use in various fields of science, the relationship between efficiency and the properties of the network on which algorithm works is not yet fully understood [10].

Examining performance PageRank network model based on actual data, we can show that the real effects make the PageRank could fail at the most valuable nodes and that it is dependent on the scope of the model parameters [8].

The most popular of the ranking algorithms, is designed to rank websites in the results displayed by the search engine. The algorithm is based on the idea: "A node is importand if it is indicated by other important nodes" [7]. The essential role played by the PageRank algorithm in Google search, led to intensive research of its properties. PageRank is very often used beyond its original range: in the ranking of scientific papers, authors, journals, ranking images in the search engine, ranking of urban roads according to the flow and traffic, measuring the importance of biochemical reactions and metabolic, etc.

Unusual characteristics and stability of the algorithm makes it suitable candidate for calculating the rank of nodes in the network, for example such as the World Wide Web, where the information is often not completely reliable.
PageRank is a global ranking of all pages, independent of their content, based on their position in the structure of the network graph. PageRank value for node P_i, marked as $r(P_i)$, is a sum of all ranking values of nodes points to P_i

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r\left(P_j\right)}{\left|P_j\right|} \qquad (1)$$

where, B_{P_i} is a block of all nodes that points to P_i, while $|P_j|$ is a number of outgoing links from node P_j. It should be noted, that the PageRank on of the incoming links $r(P_j)$ in the above

equation P_j, designated $|P_j|$. The problem that is reflected with this record, is that the value of $r(P_j)$, which is a number of nodes P_i, are unknown. To deal with this problem we use an iterative algorithm. At the begining, we assume that rank of all nodes has the same value. (for egzample 1/n, where n is the total number of pages in our repository). In that case, we have possibility to calculate $r(P_i)$ for each node P_i from our index. Thus, the above equation can be applied by substituting the value of the previous iteration as $r(P_j)$.

Let $r_{k+1}(P_i)$ be the PageRank value for node P_i for iteration k+1, then:

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|} \quad (2)$$

This proces is initialized with the value of $r_0(P_i)=1/n$ for every nodes P_i and repeated until the results of Pagerank will strive to set the value.

Simple definition of PageRank is suitable for the interpretation of the model based on random checkpoints – named by Lawrence Page – Random Surfer Model.

Random surfing the web is equivalent to a random walk on the graph base. This type of random walk is a well-studied combination-problem. It can be shown that for the PageRank algorithm, the vector r is proportional to the stationary distribution of the probability of random walking. In this way, the rank PageRank is proportional to the frequency with which random surfer will visit it [7, 8].

It appears here an issue with ranking function. We assume that the two sides refer to each other, but nowhere else. We suppose also that there is somewhere a page that has a link to one of these two. During the iteration, PageRank value will be accumulated in a loop between the nodes, but will not be distributed further, because there is no way out of the loop. This is how the situation is defined as Rank Sink.

To circumvent this trap, we can use a Random Surfer Model [7, 8]. It refers to the random movement of the graph. Random Surfer simply successively clicks on random links. If true network user would get into a small loop between the parties, it is unlikely to continue to move in it without end. Instead, you jump to another page. This behavior is defined as periodic boredom and jump to another page.

Another threat to the correct calculation PageRank model are links leading to sites with no outgoing link (eg. Privacy policy). During the study it can be found that the database can contains a large number of such links. Since it is not possible to distribute them in rank to other sites, it was decided to remove from system all such links before the PageRank is calculated. PageRank is well defined only in the case where the compound of links (link graph) is strongly associated. Each strongly (internally) connected cluster of Web pages, which do not overlook external links (to the world) leads to volatilization rank – a phenomen of Rank Sink. Individual party, which has no links to the outside causes the leakage of rank – a phenomen of Rank Leak.

Technically speaking, Rank Leak is a special case of Rank Sink, Rank Leak also causes other problems. In the case of Rank Sink nodes receive the rank of 0, which means that we can't recognize the validity of such nodes.

In order to solve this problem, PageRank was modified, for node r(i):

$$r(\mathrm{i}) = \mathrm{d} \sum_{j \in B(\mathrm{i})} r(\mathrm{j}) / N(\mathrm{j}) + (1-\mathrm{d})/\mathrm{m} \quad (3)$$

where:

d – dumping factor (average value 0.85, it is salso used as α)

$B(i)$ – set of nodes pointing to node i

$N(\mathrm{j})$ – set of outgoing links from node

m – number of all nodes in graph.

Please note that Simple PageRank is a special case occurs when d = 1. In the case of Random Surfer, modification of the model shows that the surfer may occasionally be "boring" and jump to a random Web page (instead of random pages of the current page).
To prevent a crash surfer between the nodes, other nodes must have a nonzero rank.

3. HITS algorithm

Regardless to Brin and Page, 1998 J. Kleinberg proposed a different definition of the meaning and validity of the ranking calculation nodes on the Web. Kleinberg said that it is not necessary that only highly traded nodes point to other nodes by adding value to them in the rankings.

Instead, there are special nodes that act as hubs, which contain a collection of links to valuable nodes which he called authorities. It was proposed two-level identification of hubs and authorities. In this framework, each party may be treated as having two identities [4].

HITS (Hypertext Induced Topic Search) algorithm is based on a search through the links and references. In contrast to the PageRank

technique that assigns a global rank of every page, HITS algorithm is dependent on the ranking of links. Instead of producing a single result ranking algorithm HITS produces two rankings – the authority and hub. Document authority (Authority Page) pages are, they point to other pages or documents. Document hub (hub page) is a document that is not necessarily an authority, but refers to documents authorities.

There are two reasons for a closer look at authorities. First, the hubs used in the HITS algorithm for calculating the sides of authorities. Second, the hubs are in themselves very useful compendium of answers to a user's query. Hub indicates a document authority. Document authority is a party, which indicates multiple hubs pages.

The basic idea of HITS algorithm is to identify a small subgraph of the Web and application analysis of the links in the vertex sets to locate sites of authorities and hubs for a given query. Subgraph is selected depending on the question asked by the user. Subgraph choosing a small, typically a few thousand pages, not only focuses on the analysis of references substantial part of the web, but also reduces the amount of calculation in the next step. Since the selection subgraph and its analysis are performed during query execution, it is important that this is done quickly.

Focused subgraph is generated by creating a set of roots R – random collection of pages that contain the string of inquiries – and extend it to hand over thematically similar to R.

As input, algorithm takes: string as a query, and two parameters *t* and *d*. The *t* parameter limits the size of the set of roots, while parameter *d* limits the numer of pages added to concentrated subgraph. This data control limits can be used to limit the impact on the search engine by very popular sites, for example, in order to prevent the domination of the result list.

In the phase of analyzing the links, HITS algorithm uses an internal indirect connections to identification of authorities and hubs from extended S set.

Let B(i) be a set of pages points to page i.

Let F(i) be a set of pages pointed by page i.

Link analysis algorithm creates an authoritative assessment of A_i, and assessment of H_i hub for each page of S set.

The algorithm is iterative and performs two types of operations in each step, they are referred to as "I" and "O" operation.

During the operation, and the result of the assessment of the authority of each page is updated to the sum of the results of the evaluation hubs for all pages that points them.

$$a_i = \sum_{j \in B(i)} h_j \tag{4}$$

During the operation O hubs evaluation of each page is updated to the sum of the results of the assessment authority on all sides, they point to.

$$h_i = \sum_{j \in F(i)} a_j \tag{5}$$

Stages *I* and *O* shows us that a good authority document is indicated by many good hubs. By the way, note that the page can be, and often is, both the authority and the hub. HITS algorithm calculates the results for both of them.

Iterative algorithm repeats two stages, with the normalization, until the evaluation of hubs and authorities will give consistent results.

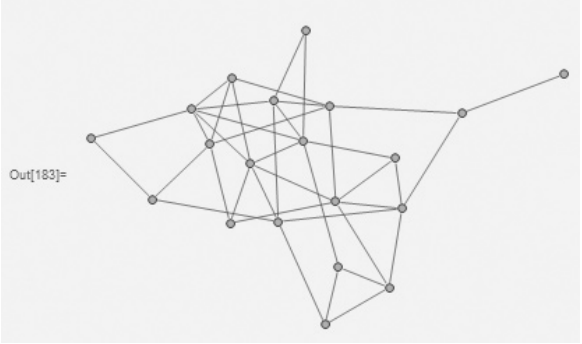

Fig. 1. Visualization of the the random graph G, as a start for HITS algorithm calculations
Source: Self developed

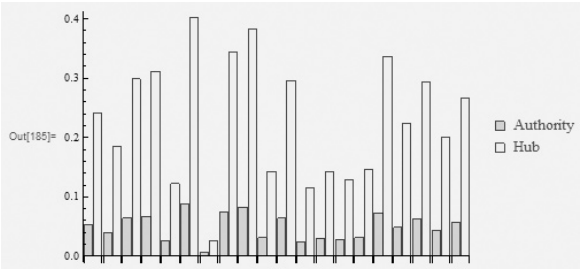

Fig. 2. Visualization of a list values of authorities and hubs in HITS algorithm for the the graph G
Source: Self developed

4. SALSA algorithm

Stochastic Approach for Link-Structure Analysis (SALSA) is the ranking algorithm designed by R. Lempel and S. Moran, assigning high scores for hubs and authority complex network, based on the number of links between them [5]. It is an alternative algorithm that combines the features of algorithms HITS and PageRank.

As with the HITS, the nodes are divided into authorities and showing no hubs. SALSA

algorithm performs a random passage (random walk) by a set of hubs and authorities set.

Like the HITS algorithm assigns two ranking results for each node: the result of the hub and the result of the authority.

In the phase of selection we have focused subgraph, which consists of nodes most appropriate for a given topic (eg page top-N returned by the search engine algorithm text), and then calculate the authorities and hubs. In this way, the two sets of nodes are dependent on the search topic.

SALSA algorithm can be seen as an extension of the algorithm HITS. The values of hubs and authorities in this algorithm are calculated at query execution time, and thus can significantly affect the response time of the search.

5. The structure of connections in a complex network

In order to research and analysis of the structure calls were collected real data from the Internet, using a properly configured crawlers Apache Nutch. This robot allows to crawl the network segment defined by the dimensions of the width of the combined domains and their depth. The result is a copy of a selected segment of a network that is stored in an internal database of Apache Nutch.

After removing information noise we receive data describing the nodes and their connections. Such data set can be imported by a code to the Wolfram Mathematica software.
Imported data we use to build a matrix switch with information about the connections between the nodes, can perform complex network visualization on a directed graph as on Figure 3.

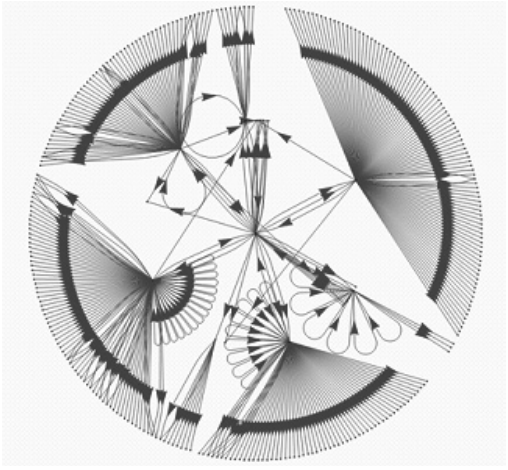

Fig. 3. Visualization of the first domain level of wat.edu.pl
Source: Self developed

At this point, we are able to calculate the values of PageRank to the first level of a network. Based on the above example, the five highest ranking values arranged as follows:

Tab. 1. PageRank values for the five highest ranks
Source: Self developed

Node	PageRank value
http://studentwat.edu.pl/	0.0544951
http://www.bg.wat.edu.pl/	0.0388971
http://intranet.wat.edu.pl/	0.0388971
http://www.cgs.wat.edu.pl/	0.0369862
http://usos.wat.edu.pl/	0.0369862

Tab. 2. HITS values for the five highest ranks
Source: Self developed

Hub Node	HITS value
http://studentwat.edu.pl/	0.0495771
http://www.bg.wat.edu.pl/	0.0366853
http://intranet.wat.edu.pl/	0.0364632
http://www.cgs.wat.edu.pl/	0.0318841
http://usos.wat.edu.pl/	0.0267361

It is easy to observe a strong leakage phenomenon occurring rank associated with a lot of dangling nodes. In this example, this is due to reduction of the data nodes of the first network level. This phenomenon can also be observed with the addition of another level dangling nodes, it always result in a loss of rank.

6. New algorithm assumptions

In many cases, prior to the calculation, using the selected algorithms, hanging nodes are removed as not importand values (for example, links to image files, etc.) and resulting loss of rank, also in order to speed up the calculation of the large amounts of data.

However, dangling nodes are very important for several reasons: they are necessary for proper representation of the network structure. They tend to act “on the border” of network, which is where the connection is fast changing. Most of the websites are dangling nodes. It is estimated that dangling nodes can exceed the number of other nodes almost four times. Some types of URLs are naturally dangling. They are used for download files such as PDF, diagrams, multimedia files, which should have a high ranking.

It is expected that a good ranking algorithm should produce an impartial ranking, where both recent and old nodes have the same chance to appear at the top. The networks also arises the problem of evolving new nodes that may have important information. To clarify: last nodes get some links, because their weight is much less than the mass of older nodes that have already accumulated a lot of links. The problem may also be that the old nodes already point to many other nodes and new ones are not yet any outgoing calls (as a dangling nodes can be removed from the final calculations).

The structure of the matrix including all the nodes of a network and the connection between them allows the implementation and use of new algorithm for ranking without having to remove the nodes, and comparing the results with data of the algorithms described.

An interesting approach seems to propose an algorithm based on the idea of PageRank with implemented support hanging nodes and using the methods of aggregation. Such a solution avoids the need to remove nodes to speed up the calculations, and also affect the accuracy of the ranking, and can significantly shorten the time required for calculation. In addition, evolving networks can be expected that the next crawling our segment of the network obtains a connection to another segment or domain through just such a dangling nodes. Such a combination will increase the accuracy of calculation of he ranking, and will make the structure of crawled network at any given time [12].

All nodes can be group in parametrized blocks, such as dangling nodes as one block, nodes without connection to dangling nodes as second block and all other nodes as third block.
Expected results can be:

- significant improvement of computational efficiency using aggregation methods – acceleration with increasing data sets.
- implementation of link-update and page--update operations in the matrix in terms of value changes and size changes.
- new and affordable way to position nodes in a network that can be practically used in applications.

The next step is implementation of this proposition of algorithm (called FastRank), based on the method of divide and conquer, which can be used to group nodes in exactly parameterized blocks, allow for more efficient updating of existing nodes and adding or removing nodes from the matrix calculations, while maintaining the maximum high degree of accuracy of the calculations rankings.

We split our set of nodes for a three blocks according to specifity, and used each block results to compute other block results.

Implementation and results of above theory shows that there is an improvement of computational efficiency during first tests.

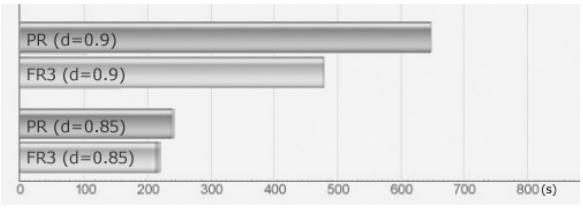

Fig. 4. Compare of time efficiency of classic PageRank method (PR) with FastRank method (FR) and different values of *d*.
Source: Self developed

We can see that there is not much difference with 0.85 parameter *d* value, between PageRank and Fastrank method. But if we look at 0.9 parameter *d* value, it is about a quarter better performance. In this case we observe shorter computation time with increased *d* value, and it means that there is improve in calculation accuracy. Test set was about 120.000 nodes of real data crawled from World Wide Web.

The computational complexity still remains the same as in classic PageRank and HITS methods.

This result shows us that there is a sill a lot to do here. I hope (and have some reasons to believe) that block divide should lead to a better performance. The next goal is to investigate this.

7. Bibliography

[1] Markov Z., Larose D., *Eksploracja zasobów internetowych*, Wydawnictwo Naukowe PWN, Warszawa 2009.

[2] Fronczak A., Fronczak P., *Świat sieci złożonych. Od fizyki do internetu*, Wydawnictwo Naukowe PWN, Warszawa 2009.

[3] Lay D.C., *Linear Algebra and Its Applications*, Addison-Wesley, 2000.

[4] Kleinberg Jon M., "Authoritative sources in a hyperlinked environment", *Journal of the ACM* (*JACM*), 46.5: 604–632 (1999).

[5] Farahat A., LoFaro T., Miller J.C., Rae G., and Ward L., "Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization", *SIAM J. Sci. Comput*., 27(4), 1181–1201 (2006).

[6] Borodin A., Roberts G.O., Rosenthal J.S., Tsaparas P., *Link Analysis Ranking*

Algorithms Theory And Experiments, The Pennsylvania State University, 2005.
[7] Page L., Brin S., Motwani R., Winograd T., *The PageRank Citation Ranking: Bringing Order to the Web. Technical Report*, Computer Science Department, Stanford University, 1998.
[8] Brin S., Page L., *The anatomy of a large--scale hypertextual Web search engine*, Computer Networks and ISDN Systems, 1998.
[9] Langville A.N., Meyer C.D., "Deeper Inside PageRank", *Internet Mathematics*, February 2004.
[10] Kamvar S., Haveliwala T., Golub G., *Adaptive Methods for the Computation of PageRank. Technical Report*, Computer Science Department, Stanford University, 2003.
[11] Kamver S., Haveliwala T., Manning C., Golub G., *Exploiting the block structure of the web for computing PageRank.* Tech. Rep. SCCM03-02, Stanford University, http://www-sccm.stanford.edu/nf-publicationstech.html, 2003.
[12] Berry W.M. and Browne M., *Understanding Search Engines*: *Mathematical Modeling and Text Retrieval*, SIAM, Philadelphia, 2nd edition, 2005.

Algorytmy rankingu w sieciach złożonych

P. WOŁOSZYN

W ostatnich latach zaobserwować można duże zainteresowanie środowisk naukowych obszarem sieci złożonych. Zbiór stron z sieci World Wide Web można zamodelować jako graf skierowany, gdzie węzły są wyznaczone jako poszczególne strony, a linki jako połączenie pomiędzy nimi. Liczba stron internetowych, które biorą udział w rankingu, podana jest w miliardach, zatem obliczenia są czasochłonne, uzależnione od użytych algorytmów oraz oczekiwanego stopnia dokładności. Algorytmy takie jak PageRank, HITS, SALSA i ich modyfikacje mają do czynienia z problemem ilości przetwarzanych danych. Dlatego potrzebne są nowe narzędzia, wydajne obliczeniowo w szczególności w oparciu o analizy sieci dla wspólnego rankingu wszystkich węzłów. W prezentowanym artykule przedstawiam charakterystykę trzech najbardziej znanych algorytmów rankingu oraz propozycję założeń do opracowania nowego algorytmu wraz z pierwszymi testami na zestawie realnych danych.

Słowa kluczowe: analiza linków, przeszukiwanie sieci Web, HITS, PageRank, SALSA, autorytety i koncentratory.