Effective Solving the One-Two Gap Problem in the PageRank Algorithm

Javad Paksima
Department of Engineering, Payame Noor University, Tehran , Iran
Paksima@pnu.ac.ir
Homa khajeh*
Department of Engineering, Science and art University, Yazd, Iran
khajeh121@yahoo.com

Received: 06/Sep/2017 Revised: 17/Dec/2017 Accepted: 08/Apr/2018

Abstract

One of the criteria for search engines to determine the popularity of pages is an analysis of links in the web graph, and various methods have already been presented in this regard. The PageRank algorithm is the oldest web page ranking methods based on web graph and is still used as one of the important factors of web pages on Google. Since the invention of this method, several bugs have been published and solutions have been proposed to correct them. The most important problem that is most noticed is pages without an out link or so-called suspended pages. In web graph analysis, we noticed another problem that occurs on some pages at the out degree of one, and the problem is that under conditions, the linked page score is more than the home page. This problem can generate unrealistic scores for pages, and the link chain can invalidate the web graph. In this paper, this problem has been investigated under the title "One-Two Gap", and a solution has been proposed to it. Experimental results show that fixing of the One-Two gap problem using the proposed solution. Test standard benchmark dataset, TREC2003, is applied to evaluate the proposed method. The experimental results show that our proposed method outperforms PageRank method theoretically and experimentally in the term of precision, accuracy, and sensitivity with such criteria as PD, P@n, NDCG@n, MAP, and Recall.

Keywords: One-Two Gap; PageRank; Search Engine; Web Graph.

1. Introduction

Search has become the predominant way of getting our everyday information in Web. Since online resources are growing rapidly, the use of search tools is required. 91% of search engine users said that when they use the search engine, they usually or more often find the information they need [1]. According to a study conducted in [2], users only examine the results of the top rankings, indicating the importance of ranking. Unit ranking is one of the most important parts of the search engine. Ranking is a process by which the page quality is estimated by the search engine. Currently, there are two major methods to rank web pages. In the first method, the ranking is based on the content of the web page (traditional ranking). Models such as the Probabilistic, Vector Space, and Boolean models are presented for content-based ranking [3]. The second method determines the importance of ranking pages based on web graph and web connections.

Unlike the traditional information retrieval environment, the Web has a large heterogeneous structure, with web pages attached to each other and forming a large graph. Web links include valuable information [4]; therefore, new ranking algorithms are created based on the link. Their main strength is to use the contents of other pages to rank a page [5].

Most search engines use algorithms to score pages based on the web graph. Links represent the quality of a page's content from the perspective of the outer pages (as opposed to the textual content of the page that is fully dependent on its creator). The link text usually contains a descriptive description of a page by other pages; in other words, the ranking is based on a link from the content of other pages to evaluate a page. Most graph-based methods are designed with the assumption that links are created by someone other than the page designer, and the purpose is to advise the page, but this is not always the case.

These algorithms are divided into two major categories: independent of queries, dependent on queries. In query-dependent methods such as PageRank [5] and HostRank [6], ranking is online and using the entire web graph. As a result, the rank of each page is constant for each query; but in query-based methods like HITS [7], ranking is performed only in part of the web graph, which includes query-related pages.

Among the algorithms, the PageRank algorithm is more important because the only algorithm in the search engine is to rank web pages [8]. It is currently used by Google's renowned search engine. Almost every algorithm is presented in a ranking, which has a problem, and PageRank is no exception to this rule. Some of the PageRank problems are addressed in Section 1-2. We encountered a new problem in examining the Web graph and calculating PageRank; that is, if a page has only one backlink, the second page may have a higher score than the first one, which applies to pages with double-top

* Corresponding Author

output levels That is why the name of the problem was "One-Two Gap".

The rest of the article is organized as follows: The PageRank algorithm, and its problems are discussed and used terms in this article are expressed in Section 2. In Section 3, the One-Two Gap problem will be explained. In Section 4, we will analyze this problem analytically and identify the pages that have this problem. Section 5 provides a solution to this problem; and in Section 6, the results of this solution are considered for the TREC Web graph, which is used to better illustrate the problem of One-Two Gap of the number of links between pages. We will conclude and summarize the discussion in the final section.

2. PageRank Algorithm

The PageRank algorithm works independently of the query, and it is used in the Google search engine. This algorithm runs on the entire web graph, and the rank of each page is equal to the total sum of the rank of its input pages; that is, a page with a high rank, with a large number of pages referring to, or pointed pages that have a high ranking [9], [10].

PageRank addresses the links between pages. For example, if the P_1 page has a connection to P_2, then the P_2 issue is probably interesting for the P_1 creator, so the number of links to the web pages indicates the degree of interest in the page for others. Clearly, the degree of interest in the page increases with increasing number of input links. Additionally, when the web page receives links from an important page, naturally it should have a higher ranking. PageRank of page j is displayed with r (j):

$$P_j = \frac{1-d}{n} + d \times \sum_{i \in B(j)} \frac{P_i}{O(i)} \quad (1)$$

Where O(i) represents the number of out-links from page i and B(j) represents the set of pages that refer to page j.

Therefore, PageRank j is equal to the total PageRank of the input pages divided by the degree of output. PageRank of the pages input divided into their out-degree O(i) has two effects. First, the distribution of PageRank to all outputs is fair; and secondly, the sum of the effect of each page and the vector of its page rank is normal. n is the total number of web pages in the web graph.

Parameter d is used to specify the probability of jumping to pages, which is in fact equivalent to random surfer behavior. When a user accesses a page without an out-link, it jumps to another page in random order; therefore, when a user is on a web page, with probability d, he chooses one of the random out-links or jumps to other pages with the probability of 1-d. Because this method is independent of the query, all pages compete with each other and reduce accuracy. This method suffers from a rich-get-richer problem [11]. In addition, the low utility coefficient of this algorithm is due to the lack of a web graph and the limited number of queries. The biggest advantage of PageRank is that it has nothing to do with the input (the query word), so all PageRank values are calculated as offline. It reduces online computing; however, the biggest defect in pagerank algorithm ignores the relevance of the subject with the information. Otherwise, Pages with different PageRanks can exist that have similar contents [12].

2.1 Overview of the Pagerank Problems

In this subsection, some of the similar tasks in the area of resolving PageRank problems are being examined to use existing ideas for further analysis of One-Two Gap problem.

In [13], three problems with using links are as follows:

- Two or more links may be created from a website or from two identical sites.
- Two or more links may be created from two similar sites to a site. In this case, two links should not be considered.
- Some links are created unrealistically for spam pages to raise their rank in search engines.

One of the PageRank problems is suspended pages [14]. Not all web pages have an out-link such as images, PDFs, and some explanatory pages and the like. Suspended pages are those that do not have an out link and they score points to their side like a hole.

A method was suggested for determining the spam linking of suspended pages. This method randomly selects a target page and identifies it by using a special vector and a special amount of spam; and then, by adding and removing the link will fix the problem. Eventually, the PageRank algorithm applies to the modified graph. The major problem with this method is its high execution time and the computational complexity that is practically impossible for large graphs [15].

In [16], a simple algorithm for calculating PageRank is presented. This algorithm considers all suspended webpages as a page and shows that the ranking of non-suspended webpages can be calculated independently of the pending page rank. Their performance has led to a ranking implemented on the smaller matrix. It was shown showed that the PageRank of suspended pages strongly affects non-suspended web pages, but it does not exist on the contrary. The benefits of this method are simple implementation and minimal storage.

In [17], Wang et al. raised the zero-one gap problem in PageRank. One method to calculate the privilege of suspended pages is to disconnect the inputs of these pages. In this method, the score obtained from the input pages is zero, and thus, there is a long difference between pages that do not have an out-link and those that have only one out-link. This problem is called the Zero-One Gap, and Wang et al. presented a new algorithm called DirchletRank to solve this problem. The DirchletRank algorithm is similar to the PageRank algorithm, with the difference that it does not have the Zero-One gap problem.

Bartlett et al. described another problem for graph-based methods [18]. They claimed that a link to a page could be a veto to determine the quality or lack of quality of that page. They described the problems raised in [13] in another way. For example, they identified links to spam pages that should be deleted in the polls or named

repeated links that should be considered once in the voting. To solve these problems, they presented a new model called Super Graph. In the proposed model, the web super graph was constructed to categorize pages in distinct groups, and the main graph links were used to construct hypergraph links. In this way, the graph connections were more uniformly shaped.

Another problem of pagerank is the density in web graph [19]. Experiments show that the web graph is usually a Power Law distribution. In [19]-[25], experiments show that the distribution of PageRank, Out-degree and In-degree usually follow the Power Law distribution for different domains and different number of pages in the web graph. For example, the number of web pages with i inlinks is commensurate with $\frac{1}{i^{2.1}}$. This makes the connections matrix sparse matrix and thus scores assigned to many pages are more negligible and newborn pages receive a very small score.

Pang at el. [26] improved the PageRank algorithm by utilizing the content of the pages and time factor to resolve the problems of topic drift and emphasize older pages (the same problem of rich-get-richer). To reduce the rich-get-richer problem, Setayesh et al. created a new version of the PageRank algorithm that uses the interests of web page users and an ant colony algorithm [27].

The Norm-PageRank algorithm is a new version of PageRank. In each step, this algorithm normalizes web pages PageRank scores to the speed of convergence [28]. The TrustRank algorithm was presented by Google in 2005 [29] to reduce Link Spam. This algorithm considers trusted and well-known pages as the seed pages that do not leak into the spam page.

Xing et al. suggested the Weighted PageRank or in short WPR [30]. In their proposed algorithm, they received more weight, depending on their importance, instead of output pages of one page, which is received the same score from the previous page.

Another problem that is mentioned in [31] is the back button issue when it is redirected from a single page. In the PageRank algorithm, it is assumed that the user chooses one of the output links or jumps to another page, while the third mode is also possible and returns to the previous page, which is unlikely to be zero. Matthew and Bowellit [31] corrected the web graph by establishing a link between each page and the previous page.

When you log in to some pages, the page path will automatically be changed. This issue is also one of the challenges in the Web graph, and research has also been done in this area [32]. On the web, a group of pages may have links to each other and have no links out. This problem is called the spider-trap [33] and the method of removing it is similar to that of suspended pages.

In addition, in [34], a comparison was made between algorithms based on PageRank and methods that have remedied PageRank bugs. The DistanceRank algorithm, which is based on reinforcement learning, reduces one of the PageRank problems that is rich-get-richer [35].

Now, before expressing the One-Two Gap problem of the PageRank algorithm, the used terms are introduced in the next subsection.

2.2 Used Terms

Single node or single page: The nodes of the web graph, with a degree of output, are called single page. In fact, there is only one out-link on the pages equivalent to the single pages. The single page often relates to pages that are redirected automatically. Of course, in other cases, the single pages also appear. For example, some sites first describe their graphic problem and consider a button to opt cancel, which links that button to the main site; or some site designers add links to their sites at the bottom of the pages they are designing. Now, if the design page does not have a specific out link, it will appear as a single page in the graph due to the designer's link.

Single Chain: A redirection may be performed in several steps, and in practice, several single pages are referred to together; this set of single pages is called a single chain.

Length of the single chain: The number of edges that connect the pages in a single chain is called the length of the single chain. For example, in Fig. 2, there is only one single chain and its length is one, or in Fig. 3, the length of the single chain is two. In order to generalize this definition for pages that do not belong to any single chain, we consider the length of zero.

3. One-Two Gap Problem in PageRank

Here are a few examples of the One-Two Gap. Suppose Fig. 1 is a normal web graph with four web pages, with its PageRank values displayed next to the pages.

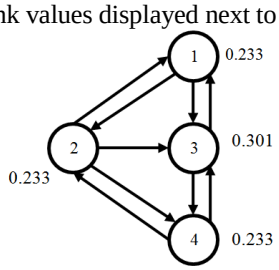

Fig. 1. Web graph without problems, pages suspended without problems one-two

If the graph in Fig. 1 adds the fifth page, that page 4 refers to and takes responsibility for page 4. PageRank scores will be in the form of Fig. 2.

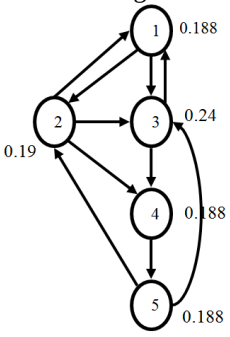

Fig. 2. Web graph with one-two Gap problem in page 4

As can be seen, page 4 has only one out-link and page 5, while having only 4 entries, has more points than page 4, which is not logical. More importantly, a new page that receives 1 entry from a page has a higher score than page 1.

If this trend continues and another page takes on the task of page 5, the scores credit will continue to decline. Fig. 3 shows this case. The new page has a score of page 6 over its equivalent pages.

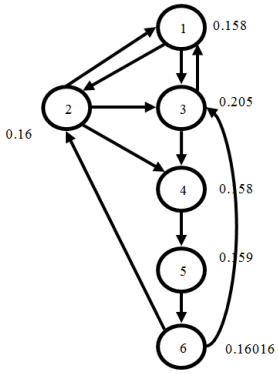

Fig. 3. Web graph with two One-Two Gap problem on Pages 4 and 5.

One way to make pages with out-degree is to use one of the ways to redirect the page to another page [32].

4. Identify Pages with One-Two Gap Problem

In this section, using a single lemma and a theorem, we identify the pages with the problem of One-Two Gap. With the tests performed, it became clear that this problem was not created for important pages, and that the nonsignificant pages had this problem. In Lemma (1), we show that nonsignificant pages with score PageRank are less than the inverse of the number of pages; and in theorem (1), we prove that the low-priority pages with the out degree of 1 have a problem of one-two Gap.

Lemma (1): If n is the total number of graph web pages, the less important pages with PageRank are less than 1/n [36].

Proof: In a normalized version, if the web graph has no pages suspended, the total PageRank scores will be 1 [36]; that is:

$$P_1 + P_2 + ... + P_n = 1 \quad (2)$$

Now, if we assume that all web graph pages have a degree of importance, that is, all pi are equal to 1/n because we have:

$$n \times P_i = 1 \Rightarrow P_i = \frac{1}{n} \quad (3)$$

So, if all web page graphs have a degree of importance, their PageRank score will be 1 / n, so more low importance pages with PageRank are less than 1/n.

Theorem (1): If a PageRank of a page with an out-degree of one is less than 1/n; in other words, the page is not important, the PageRank of the destination page is greater than the source page PageRank.

Proof: Fig. 4 Assuming page A with a PageRank out-degree of one and less than 1/n and page A to point page B.

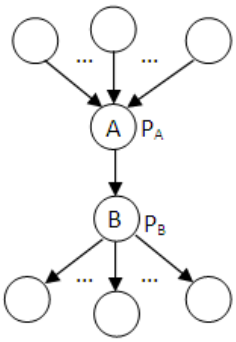

Fig. 4. Page A has an output degree of one

PageRank Score of page B will be:

$$P_B = \frac{1-d}{n} + d \sum_{i \in B_n} \frac{P_i}{O_i} \quad (4)$$

$$P_B = \frac{1-d}{n} + d \times P_A \quad (5)$$

$$P_B - P_A = \frac{1-d}{n} + d \times P_A - P_A \quad (6)$$

$$P_B - P_A = (1-d) \times (1/n - P_A) \quad (7)$$

Relation (4) is the PageRank formula used for page B. Only the sigma of relation (4) is the output of page A, according to this output degree of page A. P_A appears without a denominator in Sigma, and relation (5) is obtained. On the sides of equation (5), we reduce the P_A amount, and relation (6) is obtained, and then by factoring (1-d) relation (7) is obtained.

In relation (7), the expression (1-d) is always positive, and therefore the right-side term sign only on the sign ($1 / n-P_A$); in the event that $P_A < 1/n$, P_B-P_A becomes greater than zero; or in other words $P_B > P_A$, the theorem is proved.

The gap may be diminished by increasing d, but with a large amount of d, it cannot be exploited by other links; therefore, the problem of One-Two gap is an inherent problem.

5. Proposed Model for Solving a One-Two Gap Problem

Perhaps the easiest way to solve the One-Two gap problem is to merge the pages with one output degree to the linked pages, but this solution may be the source of new problems. Firstly, the next pages may have other input links that are ambiguous with this change of status. Secondly, the two pages may not really fit together. For example, some site designers repeat the company's address as a link on each page, and if a page may have the same link, it should be merged into the company page that is not logical.

Another solution, was provided by Mathieu and Bouklit in [31]. They corrected the web graph by setting the link between each page and the previous page and added the role of the Back button to the graph. This method does not resolve one-two gap problem, and the out degree of suspended pages was one, and the same problem occurs for those pages.

The proposed solution is to prevent illogical publishing of scores from the page with the out degree of one to the next pages. PageRank scores are achieved in two ways. One is based on the score of the pages before it in the graph and one to jump to that page. The previous pages score is controlled by d, and since d is always less than one, it does not pass over the previous page's score to the desired page. The factor of increasing the score is the probability of jumping. If the two pages are identical, the probability of jumping is one, so the probability of jump cannot be assumed to be the same and should be halved. For example, the previous address of Yazd University was "www.yazduni.ac.ir" and has now been changed to "www.yazd.ac.ir" and the user can connect with one of two addresses to the Yazd site. Now, assigning two probabilities to a site causes a problem and somehow the whole web graph is affected; therefore, we should prevent from which is more than the source page score. To do this, we first rewrite the PageRank calculation formula as follows:

$$P_j = \sum_{i \in B(j)} \left(\frac{1-d}{n \times |B(j)|} + d \times \frac{P_i}{O(i)}\right) \quad (8)$$

In this relation, the variables are similar to Formula (1). With the difference that we differentiate for page j, in which part of the score is received from each in-link. In the stage, we can minimize the release of a score more than the source page score. For this purpose, the PageRank calculation formula is modified as follows:

$$P_j = \sum_{i \in B(j)} \min\left(P_i, \frac{1-d}{n \times |B(j)|} + d \times \frac{P_i}{O(i)}\right) \quad (9)$$

Formula (9) ensures that maximum released is P_i from page i to j, and this formula is a generalized formula for all degrees. Of course, for pages with zero entry, we still need to use formula (1).

In the following, using Lemma (2) and Theorem (2) we prove that (9) should not always be used, and in the calculation of PageRank, formula (1) can often be used.

Lemma (2): For pages such as page j, received input-link from page such as i only needs to use formula (9), which is $|B(j)| < \frac{1}{n \times P_i}$.

Proof: In the calculation of P_j when P_i is used as a minimum:

$$P_i < \frac{(1-d)}{n \times |B(j)|} + d\frac{P_i}{O(i)} \quad (10)$$

Due to the fact that most of the right side of inequality (10) occurs when the O(i) is equal to one, or, in other words, page i is single page; we have:

$$P_i \times (1-d) < \frac{(1-d)}{n \times |B(j)|} \quad (11)$$

$$|B(j)| < \frac{1}{n \times P_i} \quad (12)$$

And the lemma is proved.

Theorem (2): Equation (9) for a page like j only needs to be used when j is less than the input $\frac{1}{1-d}$.

Proof: According to equation (1), $\frac{1-d}{n}$ is the lowest Pi, so we have:

$$P_i > \frac{(1-d)}{n} \quad (13)$$

$$\frac{1}{1-d} > \frac{1}{n \times P_i} \quad (14)$$

By combining equation (14) with Lemma (2), we conclude that whenever the inputs of page j are to be checked, its input degree is less than the inverse of 1-d, that is:

$$|B(j)| < \frac{1}{1-d} \quad (15)$$

This makes it easier to process pages, and we use special cases of relation (9). Given that d is usually considered to be 0.85, [37], only for pages of less than or equal to six, Equation (9) is used.

In calculating score related to PageRank, scores are calculated recursive and destructive effects of pages with the out-degree of lower than six are transferred to the other pages as recursive.

This problem exists on these kinds of pages but in the moment of calculating the score of pages. Because of the reclusiveness of the calculation and the score of these pages are effected in the scores of pages which have a link between them as a chain-by-links. So the wider range of pages will be affected by the problem. For example, if a page has the One-Two Gap problem and is linked in series to a page with the highest out-degree, the calculation of this score also causes an error.

In this paper, we tried to solve the problem with the least computations and solve the problem of these pages in order to not solve the problem in the whole graph.

The pseudocode of proposed solution is shown in algorithm 1. In Algorithm 1, the PageRank formula based on the previous discussion is used.

Algorithm 1: PageRank algorithm without One-Two Gap problem

```
1:  procedure PageRank_Without_One-Two_Gap(G, iteration)
%G: inlink file, iteration: # of iteration
2:  d ← 0.85                                   %damping factor: 0.85
3:  oh ← G                                     %get outlink count hash from G
4:  ih ← G                                     %get inlink hash from G
5:  N ← G                                      %get # of pages from G
6:  for all p in the graph do
7:      opg[p] ← 1/N    %initialize PageRank_Without_One-Two_Gap
8:  end for
9:  while iteration > 0 do
10:     for all p in the graph do
11:         for all ip in ih[p] do
12:             if |ih[p]|<= 6 then
13:                 npg[p] ← npg[p] + min(opg[ip], (1−d)/(N*|ih[p]|) + (d*opg[ip])/oh[ip])
14:             else
15:                 npg[p] ← npg[p] +
```

$(\frac{1-d}{N*|ih[p]|} + \frac{d*opg[ip]}{oh[ip]})$
%get PageRank_Without_One-Two_Gap from inlinks and get PageRank_Without_One-Two_Gap from random jump
16: **end if**
17: **end for**
18: **end for**
19: $opg \leftarrow npg$ %update PageRank_Without_One-Two_Gap
20: $iteration \leftarrow iteration - 1$
21: **end while**
22: **end procedure**

Fig. 5 shows a modified PageRank algorithm in which relation (1) based on relation (9) is modified.

5.1 Case Study

With an example on a graph with ten pages and 17 edges, we simulate the proposed solution, as shown in Fig. 5. Therefore, we can check the validity of the fixed One-Two gap problem. The simulation is done in the Matlab environment and tested on the Intel core 7 system with a six GB RAM. The proximity matrix A and the transition matrix P graph are created as follows.

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \text{ and}$$

$$P = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

For the PageRank algorithm and the proposed solution, the damping factor d is 0.85 and the threshold of 0.000000001 for the smaller squared error condition is set to be two consecutive repetitions. Fig. 5 shows that pages 1, 5, and 6 contain the problem of One-Two gap. The two left-hand figures have been used in different colors and sizes to show different rankings. This is also enhanced when the fixed problem is also more diverse rank. Two charts on the right side, from top-to-down, show the ranking pages of the pages before and after the One-Two gap problem.

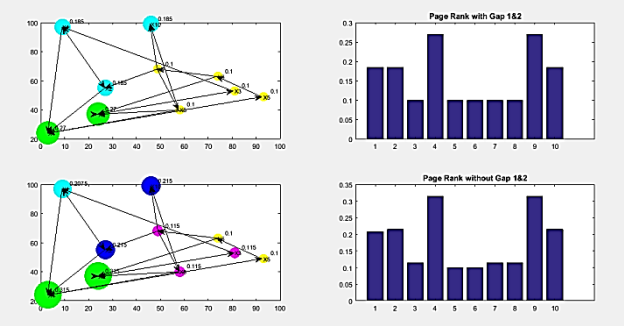

Fig. 5. The output of the page rank and their graph for the algorithms of PageRank and PageRank without One-Two gap problem

6. Experiments

We are using a TREC[1] standard data set in the .GOV domain, which was crawled in 2002. This dataset contains 50 queries, and for each query, the web pages linked to it are identified. The TREC size is about 18 gigabytes and includes 1247753 crawled web pages in the .GOV domain [38]. In this series of experiments, a number of links between pages were removed to contain 39,635 pages, including the terms of theorem (2) in the entire graph of the web, so that there is a possibility to check the validity of the proposed solution.

In this article, some links between pages have been removed in order to increase the damage effect of The One-Two Gap problem on PageRank privileges of many the pages in the web graph. And also because it is evaluated precision in the first ten ranks. Therefore, it is necessary to ensure that there is at least a page contains the One-Two Gap problem on the web graph that is relevant or linked to the relevant pages in the form of link chain. (Note that it is checked that the power law distribution of the Web graph is not lost by deleting the links).

6.1 Experiment 1: Convergence Review

After changing the PageRank algorithm to solve the One-Two gap problem, the first test is to test the method convergence to see if the algorithm's accuracy has not been lost. To prove the empirical convergence of the proposed solution, a similarity test is performed. To illustrate convergence, the results of the repetitions are compared with the last one. For this reason, we obtained the similarities of the repetitions of 10, 20, 30, 40, 50, 60, 70, 80 and 90 with 100th repetition. The similarity of the two lists is calculated according to the following equation [39].

$$\text{Similarity} = \frac{|A \cap B|}{|A \cup B|} \quad (16)$$

Where A and B represent a list of related pages from different iterations. $|A \cup B|$ indicates the total number of pages that have two lists (Union of two lists) and $|A \cap B|$ indicates the number of pages that appear on both lists (Subscribe to two lists). To draw a chart, the list of web pages is based on the N page of the sorted list, which is the

[1] http://trec.nist.gov/

horizontal axis of the graph. When the similarity of the two lists from each of the iterations is close to one, it means that the list of pages is unchanged and convergence is complete.

Fig. 6 shows the convergence of the PageRank algorithm without One-Two gap problem. The algorithm is very close to the one after the 50th iteration.

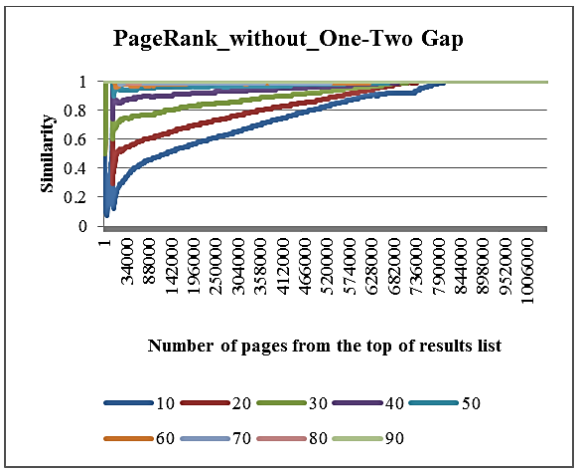

Fig. 6. The convergence of the PageRank algorithm without One-Two gap problem by comparing the similarity of repetitions with 100th repetition

Table 1. The convergence rate of the PageRank algorithm and PageRank without one-two gap problem using a 100-repeat similarity criterion

Start round number of Convergence Over 90%	Start round number of Convergence 100%	Algorithm
50	90	PageRank
50	90	PageRank without One-Two GAP

The results of Table 1 show that the convergence rate is not reduced by applying the change. In both cases, the problem of the One-Two gap and without this problem are obtained in repeats of 50 and 90, respectively, with the convergence of over 90% and complete convergence.

6.2 Experiment 2: Reviewing the Percentage Demoted of the PageRank Algorithm without the One-Two Gap Problem Compared to the PageRank Algorithm

We first consider a relative ranking change (RRC) measurement for web pages. Suppose that page i in position s in PageRank has no gap of One-Two and in position t of PageRank algorithm. The RRC criterion is defined as follows.

$$\mathrm{RRC(i)} = \frac{s - t}{s + t} \tag{17}$$

The RRC (i) is in the range [-1,1]. The positive RRC (i) indicates an increase in rank of page i and an RRC (i) negatively indicating a downgrade of the rank of page i. The RRC is relative to the high-ranking position. We consider the percentage Demoted (PD) based on the RRC for the TREC web graph pages, which is calculated as follows [17].

$$\mathrm{PD(s)} = \frac{\left|\sum_{i\in s,\mathrm{RRC(i)}>0} \mathrm{RRC(i)}\right|}{\left|\sum_{i\in s,\mathrm{RRC(i)}<0} \mathrm{RRC(i)}\right| + \left|\sum_{i\in s,\mathrm{RRC(i)}>0} \mathrm{RRC(i)}\right|} \tag{18}$$

The PD value represents the percentage demoted of pages for RRC. We use without the One-Two gap problem to show the effectiveness of PageRank, which is about 52% more demoted pages than PageRank.

Fig. 7 shows the percentage demoted ranking of pages in the pagerank algorithm without One-Two gap problem. The demoted rank of algorithm is 52%.

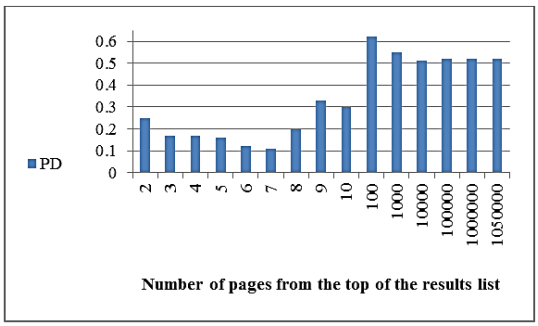

Fig. 7. Percentage demoted of the PageRank algorithm without One-Two gap problem relative to the PageRank algorithm in the different number of pages

6.3 Test 3: Evaluation the Accuracy of the Ranking

The ranking accuracy was performed using the three criteria of P@n, NDCG@n, MAP evaluation on the TREC graph, which removed a series of links between the pages.

6.3.1 Precision Evaluation Criteria

In the retrieval of information, precision and recall are used as criteria for checking the efficiency and quality of the ranking [39]. Precision criteria are used in the position n-th (p@n), mean-average precision (MAP), Normalized Discount Cumulative Gain (NDCG), and recall n-th (R@n) to evaluate the accuracy of information retrieval. The evaluation tools set of LETOR group supports these criteria [38].

- P@n

This criterion indicates the number of relevance pages to user's query in the n position of the ranking list. Relation P@n is as follows.

$$\mathrm{P@n} = \frac{\mathrm{NoR_n}}{n} \tag{19}$$

Where $\mathrm{NoR_n}$ indicates the number of relevance pages in the n position of top ranking list.

- MAP

The Mean Average Precision (MAP) represents the average AP values for all queries provided, and for each query, the AP indicates average of P@n values for all relevance pages.

$$\mathrm{AP} = \frac{\sum_{n=1}^{N}(\mathrm{P@n} \cdot \mathrm{R_e(n)})}{\mathrm{T_R}} \tag{20}$$

Where N, $\mathrm{T_R}$ and $\mathrm{R_e(n)}$ are the number of retrieved pages, the number of relevance pages, and the binary function of the n-th page, respectively, indicating the relevance page with one and the irrelevance page with zero.

- NDCG

The NDCG value of a ranked page in the n-th position is computed as follows:

$$NDCG = Z_n \sum_{i=1}^{N} (2^{r(i)} - 1/\log(i+1)) \quad (21)$$

Where Z_n represents normalization constant and r(i) indicates the relevance level of page i in ranking list. The gain of the i-th page and the discount gain are calculated with the relations $2^{r(m)} - 1$ and $2^{r(i)} - 1/\log(i+1)$. $\sum_{i=1}^{N} (2^{r(i)} - 1/\log(i+1))$ represents the normalized discount cumulative gain in the n-th position.

- R@n

Recall indicates the proportion of retrieved pages that are relevant to the query, and it is called sensitivity [40]. R@n is calculated as follows.

$$R@n = \frac{NoR_n}{NoR\_all} \quad (22)$$

Where NoR_n indicates the number of relevance pages in the top-n result of ranking list. NoR_all shows the whole number of relevance pages to the query.

Regarding Figs. 8-10, the proposed solution, compared to the PageRank algorithm was more appropriate in terms of the P@n, MAP, and NDCG@n evaluation criteria on the TREC2003 benchmark dataset. Due to the fact that the dataset has the conditions for the One-Two gap problem, according to the results, the proposed solution to solve the One-Two gap problem will work perfectly.

According to the two criteria, P@n and NDCG@n, for the first ten pages of the ranking list and MAP criteria, ranking accuracy has been enhanced by solving the problem of One-Two gap.

This precision, according to the P@n criterion in the position of one and two from ranking list, increased by 50% and 100%, respectively. According to the MAP criterion, the overall precision has increased by about 0.04%.

Fig. 11 shows the sensitivity of proposed solution better than the PageRank algorithm. In this figure, it is seen that in the top-10 rank of the ranking list, the PageRank without One-Two gap problem is more recall, and the rhythm of both methods is incremental. It represents that performance of the proposed solving way is much better than PageRank.

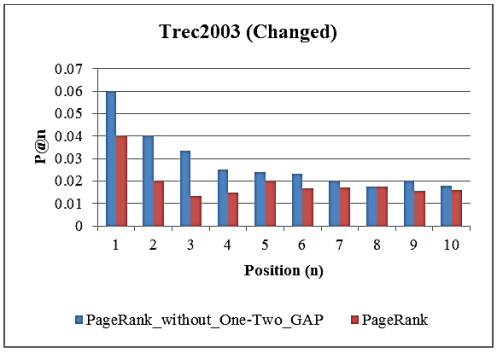

Fig. 8. Comparing the proposed solution with the PageRank algorithm on the TREC2003 data set based on the P@n criterion

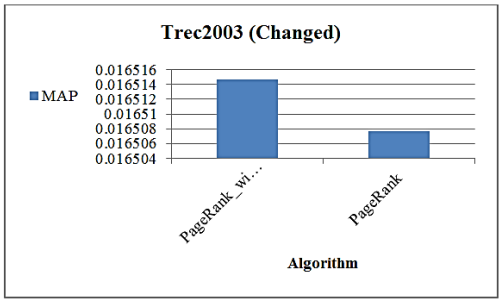

Fig. 9. Comparing the proposed solution with the PageRank algorithm on the TREC2003 data set based on the MAP criterion

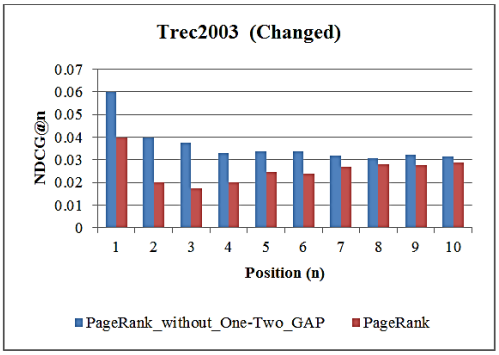

Fig. 10. Comparing the proposed solution with the PageRank algorithm on the TREC2003 data set based on the NDCG@n criterion

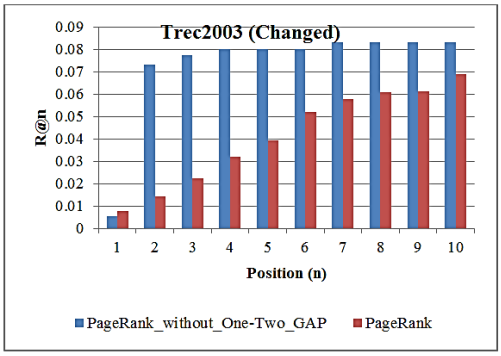

Fig. 11. Comparing the proposed solution with the PageRank algorithm on the TREC2003 data set based on the R@n criterion

6.4 Discussion and Analysis

We have theoretically shown that there is a One-Two gap problem in the PageRank algorithm. The proposed solution is a prerequisites for the One-Two gap problem. Not all the methods that have been developed based on PageRank have paid attention to this problem. In general, the proposed solution provides the right or equal value for PageRank, and can be a good alternative to this algorithm.

Experimental results also emphasize the appropriateness of the proposed solution's performance in terms of accuracy, sensitivity and convergence. The proposed method offers a good combination of sensitivity and accuracy in the top-10 rank of results. Further, the results reveal that this solution contributes to achieving better precision and recall.

The findings also establish the case that the One-Two gap problem decreases precision, accuracy, and recall in PageRank, and this article solves it. In other word, this

method prevents the release of the wrong score to other pages in whole graph.

7. Conclusion

While the PageRank algorithm is used successfully in Google's search engine, there are many researchers who pay attention to it, and many advanced methods have been proposed to improve the precision of this algorithm. In addition, the PageRank algorithm is considered one of the factors that calculates the relevance of web pages. In this paper, we have shown that link-based PageRank algorithm has the One-Two gap problem, which can put the rank of web page more than linked page; and ranks are calculated to be recursive to make errors. We have suggested a solution to this problem, which has been empirically increasing the precision of the ranking. In terms of convergence, the proposed solution is similar to the PageRank algorithm. PageRank without One-Two gap problem acquired the highest recall and precision than PageRank algorithm, on the TREC2003 dataset in domain .gov.

References

[1] K. Purcell, J. Brenner, and L. Rainie, "Search Engine Use 2012," 2012.

[2] Y. Zhang, B. J. Jansen, and A. Spink, "Time series analysis of a Web search engine transaction log," Information Processing & Management, vol. 45, no. 2, pp. 230–245, 2009.

[3] R. Baeza-Yates and B. Ribeiro-Neto, "Modern information retrieval," New York, vol. 9, p. 513, 1999.

[4] Searchmetrics, "Searchmetrics Ranking Factors 2016: Rebooting for Relevance," 2016. [Online]. Available: http://www.searchmetrics.com/knowledge-base/ranking-factors/. [Accessed: 07-May-2017].

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," World Wide Web Internet And Web Information Systems, vol. 54, no. 1999–66, pp. 1–17, 1998.

[6] G.-R. Xue, Q. Yang, H.-J. Zeng, Y. Yu, and Z. Chen, "Exploiting the hierarchical structure for link analysis," Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 186--193, 2005.

[7] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM, vol. 46, no. 5, pp. 604–632, Sep. 1999.

[8] R. Patchmuthu, "Link analysis algorithms to handle hanging and spam pages," 2014.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Rankin: Bringing Order to the Web," World Wide Web Internet And Web Information Systems, vol. 54, no. 1999–66, pp. 1–17, 1998.

[10] M. Bianchini, M. Gori, and F. Scarselli, "Inside pagerank," ACM Transactions on Internet Technology (TOIT), vol. 5, no. 1, pp. 92–128, 2005.

[11] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," Science, vol. 286, no. October, pp. 509–512, 1999.

[12] L. Z. Xiang, "Research and Improvement of PageRank Sort Algorithm Based on Retrieval Results," in Intelligent Computation Technology and Automation (ICICTA), 2014 7th International Conference on, 2014, pp. 468–471.

[13] Z. Dou, R. Song, J.-Y. Nie, and J.-R. Wen, "Using anchor texts with their hyperlink structure for web search," in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 227–234.

[14] A. N. Langville and C. D. Meyer, "A reordering for the PageRank problem," SIAM Journal on Scientific Computing, vol. 27, no. 6, pp. 2112–2120, 2006.

[15] R. K. Patchmuthu, A. K. SINGH, and A. Mohan, "A new algorithm for detection of link spam contributed by zero-out link pages," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 24, no. 4, pp. 2106–2123, 2016.

[16] I. C. F. Ipsen and T. M. Selee, "PageRank computation, with special attention to dangling nodes," SIAM Journal on Matrix Analysis and Applications, vol. 29, no. 4, pp. 1281–1296, 2007.

[17] X. Wang, T. Tao, J.-T. Sun, A. Shakery, and C. Zhai, "Dirichletrank: Solving the zero-one gap problem of pagerank," ACM Transactions on Information Systems (TOIS), vol. 26, no. 2, p. 10, 2008.

[18] K. Berlt, E. S. de Moura, A. Carvalho, M. Cristo, N. Ziviani, and T. Couto, "Modeling the web as a hypergraph to compute page reputation," Information Systems, vol. 35, no. 5, pp. 530–543, 2010.

[19] A. N. Nikolakopoulos and J. D. Garofalakis, "NCDawareRank: a novel ranking method that exploits the decomposable structure of the web," in Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 143–152.

[20] A. Broder et al., "Graph structure in the web," Computer networks, vol. 33, no. 1, pp. 309–320, 2000.

[21] D. Donato, L. Laura, S. Leonardi, and S. Millozzi, "Large scale properties of the webgraph," The European Physical Journal B-Condensed Matter and Complex Systems, vol. 38, no. 2, pp. 239–243, 2004.

[22] A. Flammini, F. Menczer, and A. Vespignani, "The egalitarian effect of search engines," arXiv preprint cs.CY/0511005, 2005.

[23] L. Becchetti and C. Castillo, "The distribution of PageRank follows a power-law only for particular values of the damping factor," in Proceedings of the 15th international conference on World Wide Web, 2006, pp. 941–942.

[24] G. Pandurangan, P. Raghavan, and E. Upfal, "Using pagerank to characterize web structure," Internet Mathematics, vol. 3, no. 1, pp. 1–20, 2006.

[25] N. Litvak, W. R. W. Scheinhardt, and Y. Volkovich, "In-Degree and PageRank of Web pages: Why do they follow similar power laws?," arXiv preprint math/0607507, 2006.

[26] P. Zha, X. Xu, and M. Zuo, "An Efficient Improved Strategy for the PageRank Algorithm," in 2011

International Conference on Management and Service Science, 2011, pp. 1–4.
[27] S. Setayesh, A. Harounabadi, and A. M. Rahmani, "Presentation of an Extended Version of the PageRank Algorithm to Rank Web Pages Inspired by Ant Colony Algorithm," International Journal of Computer Applications, vol. 85, no. 17, 2014.
[28] K. Mohan and J. Kurmi, "A Technique to Improved Page Rank Algorithm in perspective to Optimized Normalization Technique," International Journal, vol. 8, no. 3, 2017.
[29] N. L. Amy and D. M. Carl, "Google's PageRank: The Math Behind the Search Engine," Priceton university Press, Nol, vol. 3, pp. 335–380, 2004.
[30] W. Xing and A. Ghorbani, "Weighted PageRank algorithm," in Proceedings of the Second Annual Conference on Communication Networks and Services Research, 2004, pp. 305–314.
[31] F. Mathieu and M. Bouklit, "The effect of the back button in a random walk: application for pagerank," in Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, 2004, pp. 370–371.
[32] M. Zhukovskii, G. Gusev, and P. Serdyukov, "URL redirection accounting for improving link-based ranking methods," in Advances in Information Retrieval, Springer, 2013, pp. 656–667.
[33] Z. Bahrami Bidoni, R. George, and K. Shujaee, "A Generalization of the PageRank Algorithm," ICDS 2014, The Eighth International Conference on Digital Society, pp. 108–113, 2014.
[34] A. K. Singh, "A comparative study of page ranking algorithms for information retrieval," International journal of electrical and computer engineering, vol. 4, pp. 469--480, 2009.
[35] A. M. Zareh Bidoki and N. Yazdani, "DistanceRank: An intelligent ranking algorithm for web pages," Information Processing and Management, vol. 44, no. 2, pp. 877–892, 2008.
[36] B. Poblete, C. Castillo, and A. Gionis, "Dr. searcher and mr. browser: a unified hyperlink-click graph," in Proceedings of the 17th ACM conference on Information and knowledge management, 2008, pp. 1123–1132.
[37] R. Baeza-Yates, P. Boldi, and C. Castillo, "Generalizing pagerank: Damping functions for link-based ranking algorithms," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 308–315.
[38] T. Qin, T. Y. Liu, J. Xu, and H. Li, "LETOR: A benchmark collection for research on learning to rank for information retrieval," Information Retrieval, vol. 13, pp. 346–374, 2010.
[39] T. H. Haveliwala, "Efficient computation of PageRank," 1999.
[40] M. Arora, U. Kanjilal, and D. Varshney, "Evaluation of information retrieval: Precision and recall," International Journal of Indian Culture and Business Management, vol. 12, no. 2, pp. 224–236, 2016.

Javad Paksima received the B.Sc. degree in Software engineering from Sharif University, Tehran, Iran, in 1996. He received M.Sc. degree in Software engineering from Sharif University, Tehran, Iran, in 1998. He received Ph.D. degree in Software engineering from Yazd University, Yazd, Iran, in 2018. He is a faculty member of Payam Noor University (PNU). His research interests include Search engines, Algorithms design and Parallel Programming.

Homa Khajeh received the B.Sc. degree in Software Engineering from Islamic Azad University, Najafabad Branch (IAUN), in Isfahan, Iran, in 2009 and her M.Sc. degree of Software Engineering from Science and Art University in Yazd, Iran, in 2014. Her research interests are mainly in the field of Information Retrieval, Search Engine, Machine Learning, and Big Data.