Study of a Random Navigation on the Web Using Software Simulation

Mirella Amelia Mioc
Department of Computer Science, Faculty of Automatics and Computers "Politehnica" University of Timisoara, Bd. V. Parvan nr. 2, Timisoara, 300223, Romania
mmioc@cs.upt.ro

Stefan-Gheorghe Pentiuc
Faculty of Electrical Engineering and Computer Science, Research Center in Computer Science, "Stefan cel Mare" University of Suceava, Romania
pentiuc@eed.usv.ro

Abstract

The general information about the World Wide Web are especially nowadays useful in all types of communications. The most used model for simulating the functioning of the web is through the hypergraph. From the known algorithms used for web navigation in this simulation the surfer model was chose. The main objective of this paper is to analyze the Page Rank and its dependency of Markov Chain Length. In this paper some software implementation are presented and used. The experimental results demonstrate the differences between the Algorithm Page Rank and Experimental Page Rank.

Keywords: Hyperlink, Web graph, Markov Channel, Page Rank, Pseudo-Random Sequence, Surfer Navigation.

1. Introduction

The World Wide Web contains millions of web sites. From the beginning it has a continuing growth in complexity and size. In the same time, the World Wide Web is a massive global hypertext system. Generally speaking, a graph is a mathematical structure consisting of several vertices or nodes, which are connected through lines called edges. A website has a link structure, which could be represented by a direct graph. Thus, the web pages available at the web site are vertices and if and only if page A contains a link to page B, there is a directed edge between them (A and B). The web graph describes the directed links between pages of the www. Properties of such kind of graphs have been extensively analyzed (Broder et al., 2000) and used in multiple applications. These pages are written in many languages and are made available by many different authors. Today everyone is using more and more private computers, as well as networks formed by them, resulting in the biggest network ever: a global network known as the Internet. Day after day, an increasing amount of information is being transferred through different media (wired and wireless) over the internet. The type of information includes plain text, HTML, audio, video, multimedia, images and various other applications or file types.

The rest of the paper is organized as follows:

- the next section renders a brief description;
- Section 3 presents the software implementation for the analysis;
- Section 4 contains some experimental results;
- Section 5 is the conclusion.

2. Description of the analysis

This analyze has as kernel a hyperlink matrix associated with a web graph having m nodes, where m is between 15 and 25. The elements of this matrix are read from an input file "hyperlink.in", which can be randomly generated using the Mersenne generator. For obtaining the Google matrix, G, the used parameter $\alpha = 0.85$.

In this experimental analyze the intuitive significance of the page rank of one site has been demonstrated. The Experimental Page Rank of one web page is the asymptotic frequency with which one surfer visits that page or the number of visits that one surfer pays to the site from a number of steps

taken during his entire navigation. For a good simulation it is very important to find methods for navigating through the web (Levene and Wheeldon, 2004). John Kemeny and Laurie Snell have proposed the use of Markov models for web simulations (Kemeny and Snell, 1960). Cadez et al. (2000) used Markov models for classifying the sessions into different categories for browsers. Some other proposed techniques choose to combine different order Markov models for obtaining low state complexity and improving accuracy, as Deshpande and Karypis (2004). Dongshan and Junyi (2002) used for predicting the access providing good scalability and high coverage a hybrid-order tree-like Markov model. As an alternative to the Markov model Pitkow proposed a longest subsequence model (Pitkow and Pirolli, 1999), also for predicting the next page accessed by the user Sarukkai chose Markov models (Sarukkai, 2000).

Transitions are simulated using the Markov Chain nodes, Google matrix and an arbitrary initial probability distribution. Examples can be seen in Figure 1 and Figure 2.

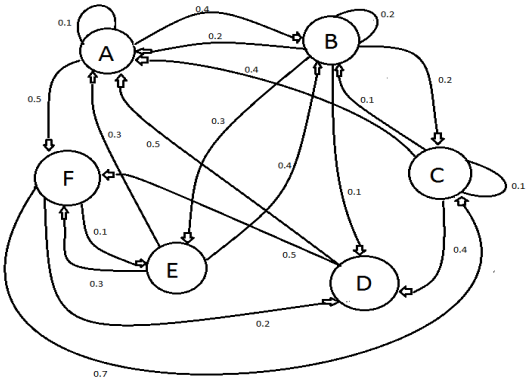

Figure 1. Markov Chain Model

	A	B	C	D	E	F
A	0.1	0.4	0	0	0	0.5
B	0.2	0.2	0.2	0.1	0.3	0
C	0.4	0.1	0.1	0.4	0	0
D	0.5	0	0	0	0	0.5
E	0.3	0.4	0	0	0	0.3
F	0	0	0.7	0.2	0.1	0

Figure 2. Transition matrix

Algorithmic Page Rank is calculated using the power method, having $e=10^{-5}$. In the software implementation the simulated path has a Markov length of 200.

3. Software implementation

For analyzing the functioning of different possible web graphs it was used a software implementation. In the following it will be presented the implementation for the most important parts of this simulation. For this it has been used the hyperlink matrix H. In the outlink function the sum is calculated in order to be used in obtaining the page ranks.

```
void fct_outlinks()
{
    int i,j,sum;
    for (i=0;i<m;i++)
```

```
        {
            sum=0;
            for (j=0;j<m;j++)
            {
                sum=sum+H[i][j];
            }
            pag[i].outlinks=sum;
        }
    }
```

Figure 3. Outlinks function

Function LantMarkov() is used in three different ways. This function calls simDiscr() which simulated a discreet distribution of probabilities.

```
int *LantMarkov (int m,double
*p0,double **L)
{

    for (i=0;i<m;i++)
      x[i]=i;

    s[0]=simDiscr(m,x,p0);
    i=s[0];
    for(k=0;k<Z;k++)
    {
      p=L[i];
      pag[i].nrviz++;
      s[k]=simDiscr(m,x,p);
      i=s[k];
    }

    for (i=0;i<m;i++)

      pag[i].PgRkE=(double)pag[i].nrviz
/Z;

    return s;
}
```

Figure 4. Markov Chain Implementation

PageRank() function is based on using the power method, whith e= 10^{-5} and an uniform initial setting for probability distribution.

Algorithmic Page Rank calculus also involves the use of Google matrix, G (see Figure 5).

Algorithmic Page Rank has been calculated for a number of m = 16 sites. The Page Rank coefficient is a numerical value beginning with 0 and having it's maximum value of 10. It is well known that Google only take into account sites with a Page Rank greater than 4.

```
double *PageRank(double **G,int m,int *pas)
{

    for (i=0;i<m;i++)
    {
      pi[i]=(1.0)/m;
    }
    *pas=0;

    double eps=1.e-5;
    do
    {
      for (i=0;i<m;i++)
          piprim[i]=pi[i];
      for (i=0;i<m;i++)
          pi[i]=0;
      for (i=0;i<m;i++)
          for (j=0;j<m;j++)
                  pi[i]+=G[j][i]*piprim[j];
      *pas=*pas+1;
    }
    while (norma(m,diferenta(m,pi,piprim))>=eps);

    return pi;
}
```

Figure 5. Page Rank Implementation

4. Experimental results

Some of the most important aspects of the analysis is obtaining the parameters which can give the main information about a web. In the simulation implementation information as: page, number of inlinks, number of outlinks, value for Algorithmic Page Rank and Experimental Page Rank will be processed for obtaining the results of the analysis. For the first part it was necessary to use experimental values as: inlinks, outlinks and in and out frequencies.

Table 1. Experimental values for simulation 1

inlinks	outlinks	PgRankAlgo	inFreq*1000	outFreq*1000
4	9	0.033767	307	692
5	6	0.046948	454	545
5	9	0.041963	357	642
7	12	0.051136	368	631
7	9	0.056731	437	562
7	11	0.051806	388	611
7	5	0.056256	583	416
7	8	0.056521	466	533
8	8	0.058508	500	500
8	9	0.06016	470	529
8	7	0.057091	533	466
8	6	0.05742	571	428
9	10	0.067977	473	526
9	7	0.068807	562	437
9	8	0.067305	529	470
10	6	0.087018	625	375
10	10	0.078762	500	500
11	9	0.07874	550	450
11	8	0.08311	578	421
12	7	0.093128	631	368

Due to the large size of hyper graph the continuously changing documents and links determine an impossibility to catalogue all vertices and edges (Albert, Jeong and Barabási, 1999).

The challenge in obtaining a general map of the web is illustrated by distributions like in Figure 6 for outlinks and in Figure 7 for inlinks.

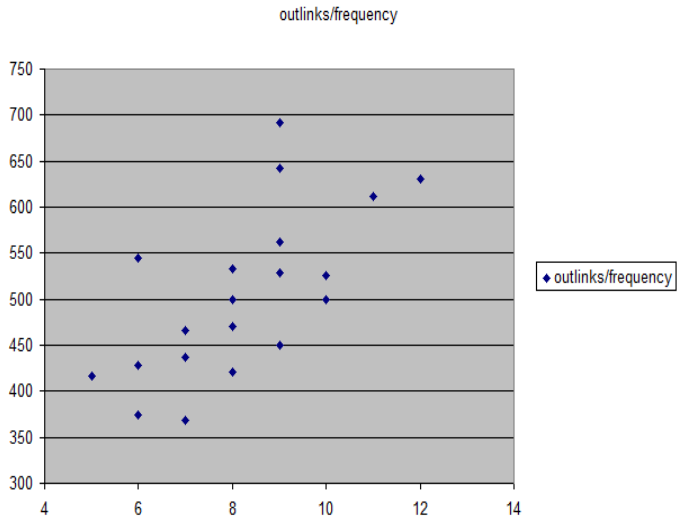

Figure 6. Outlinks degree distribution for all web sites

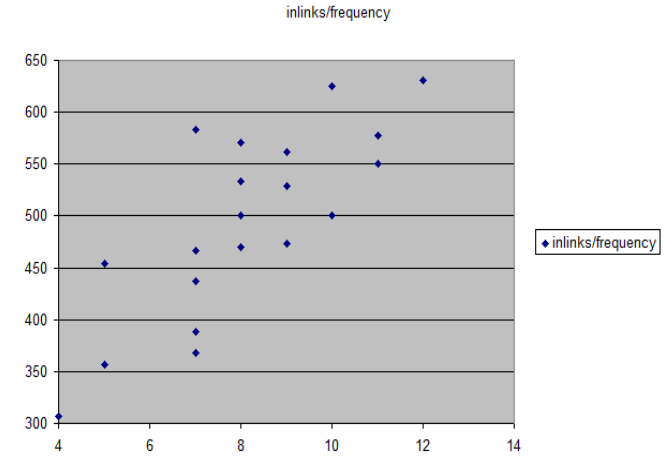

Figure 7. Inlinks degree distribution for all web sites

In the following tables there are the values obtained for each of the 16 sites for the first simulation. Table 2 contains the Algorithmic Page Ranks, while Table 3 contains the Experimental Page Ranks for several Markov chain lengths (n).

Table 2. Experimental values for simulation 2

SiteId	inlinks	outlinks	PageRankAlgo
0	9	10	0.067977
1	8	8	0.058508
2	10	6	0.07826
3	7	12	0.051136
4	8	9	0.06016
5	8	8	0.053896
6	8	7	0.057091
7	11	9	0.07874
8	7	9	0.056731
9	8	9	0.062105
10	5	6	0.046948
11	9	7	0.068807
12	9	8	0.067305
13	11	8	0.08311
14	7	11	0.051806
15	8	6	0.05742

Table 3. Experimental Page Rank Values for simulation 2

n=50	n=100	n=200	n=300	n=400	n=500
0.0672	0.0634	0.0662	0.0622	0.0636	0.0644
0.0664	0.066	0.0666	0.0572	0.0668	0.0612
0.0624	0.0682	0.0646	0.0622	0.066	0.0594
0.0654	0.0584	0.0662	0.0636	0.0628	0.067
0.0632	0.066	0.066	0.0608	0.068	0.0616
0.0662	0.0624	0.063	0.0628	0.0616	0.063
0.058	0.0548	0.0626	0.061	0.0624	0.0638
0.0556	0.0604	0.0636	0.0616	0.0612	0.0606
0.066	0.0726	0.061	0.0682	0.0574	0.0606
0.0662	0.0622	0.0594	0.0594	0.0638	0.0612
0.0564	0.0616	0.0614	0.0652	0.059	0.0622
0.0634	0.0648	0.0624	0.059	0.0686	0.0664
0.0572	0.0602	0.057	0.065	0.06	0.06
0.0618	0.059	0.055	0.0678	0.0576	0.0634
0.0654	0.0558	0.062	0.0628	0.0596	0.0626
0.0592	0.0642	0.063	0.0612	0.0616	0.0626

The next diagram proves that the values for Experimental Page Rank depend on the length of the Markov Chain, while Algorithmic Page Rank remains constant.

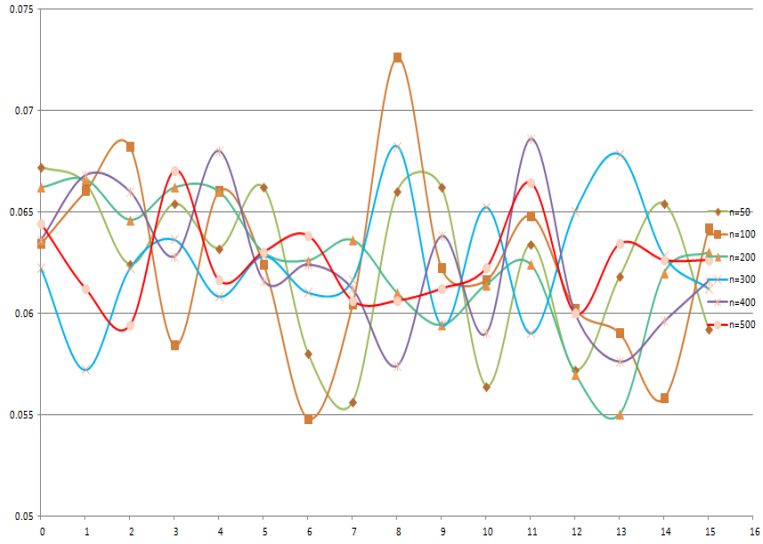

Figure 7. Experimental Page Rank dependency on Markov Chain length

Another diagram shows that for each site there is only one constant value independent of the length of Markov Chain. Algorithmic Page Rank only depends on the number of inlinks and outlinks.

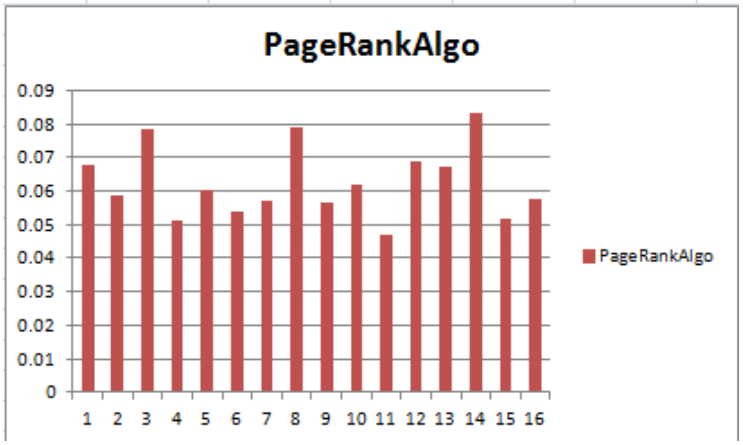

Fig. 8. Algorithmic Page Rank

Table 4 contains the values for Experimental Page Rank with balanced distribution. In the following figure it is shown the diagram for the values obtained for Experimental Page Rank.

The Experimental Page Rank values oscillate between the same limits independently of the change of Markov Chain length (N).

Tabel 4. Experimental Page Rank for simulation with balanced distribution

N=50	N=100	N=200	N=300	N=400	N=500
0.0606	0.061	0.0656	0.0578	0.059	0.0666
0.0664	0.0582	0.0658	0.058	0.0592	0.0594
0.0668	0.0578	0.0576	0.0586	0.0638	0.065
0.0546	0.0618	0.0574	0.0686	0.059	0.0604
0.0614	0.0682	0.061	0.0614	0.0644	0.0654
0.0552	0.0622	0.0556	0.061	0.0624	0.0672
0.0642	0.0604	0.0618	0.0634	0.0646	0.059
0.0616	0.0578	0.0626	0.068	0.0622	0.063
0.0628	0.06	0.0642	0.0628	0.0686	0.0644
0.0644	0.0678	0.061	0.0622	0.0602	0.0584
0.0636	0.0642	0.0624	0.0672	0.0648	0.0554
0.0616	0.0618	0.065	0.0648	0.0642	0.065
0.0642	0.0622	0.0634	0.0594	0.0596	0.0614
0.0602	0.0654	0.0654	0.0624	0.0594	0.0622
0.0672	0.0658	0.0682	0.0548	0.0608	0.0612
0.0652	0.0654	0.063	0.0696	0.0678	0.066

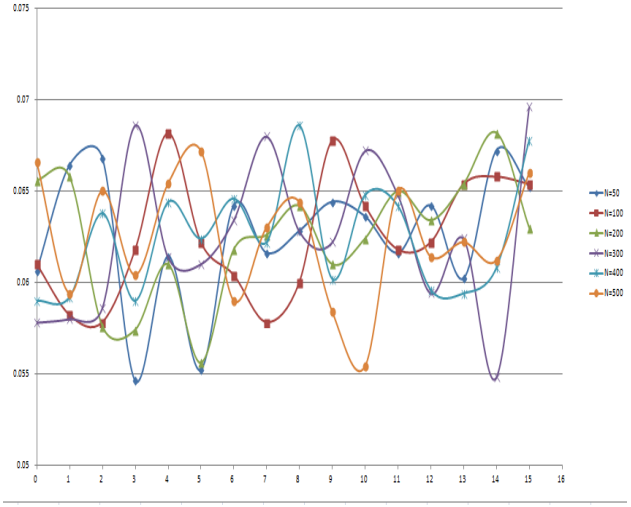

Figure 9. Experimental Page Rank dependency on Markov Chain length with balanced distribution

5. Conclusions

This paper explored different implementation choices for analyzing the most important parameters about a web. Many researchers explored the use of new search engines for studying the evolution of the web (Ntoulas, Cho and Olston, 2004). Another important research is realized about the Link Structure Graph (LSG). The LSG captures a complete hyperlink structure from the web and models link associations reflected in the page layout (Rodrigues, Milic-Frayling and Fortuna, 2007). For further works ideas like extrapolation methods for accelerating page rank calculation can be developed (Kamvar et al., 2003).

References

Albert, R., Jeong, H., & Barabási, A. L. (1999). Internet: Diameter of the world-wide web. *Nature*, 401(6749), 130-131.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2000). Graph Structure in the Web. *Computer networks*, 33(1), 309-320.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., & White, S. (2000). Visualization of navigation patterns on a web site using model based clustering. In *Proceedings of the Sixth International KDD conference*, pages 280–284, 2000.

Deshpande, M., & Karypis, G. (2004). Selective Markov models for predicting Web page accesses. *ACM Transactions on Internet Technology (TOIT), 4(2)*, 163-184.

Dongshan, X., & Junyi, S. (2002). A new markov model for web access prediction. *Computing in Science & Engineering*, 4(6):34–39.

Kamvar, S. D., Haveliwala, T. H., Manning, C. D., & Golub, G. H. (2003). Extrapolation methods for accelerating pagerank computations. In *Proceedings of the World Wide Web Conference*, Budapest, 2003.

Kemeny, J. G., & Snell, J. L. (1960). *Finite Markov Chains*. D. Van Nostrand, Princeton, New Jersey.

Levene, M., & Wheeldon, R. (2004). Navigating the world wide web. In *Web Dynamics* (pp. 117-151). Springer Berlin Heidelberg.

Ntoulas, A., Cho, J., & Olston, C. (2004). What's New on the Web? The Evolution of the Web from a Search Engine Perspective. In *Proceedings of the 13th international conference on World Wide Web* (pp. 1-12). ACM.

Pitkow, J., & Pirolli, P. (1999). Mining longest repeating subsequences to predict world wide web surfing. In *Proceedings of the Second Usenix Symposium on Internet Technologies and Systems*, Colorado, USA, October 1999.

Rodrigues, E. M., Milic-Frayling, N., & Fortuna, B. (2007). Detection of Web Subsites: Concepts, Algorithms, and Evaluation Issues. In *IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 66-73). IEEE.

Sarukkai, R. R. (2000). Link prediction and path analysis using markov chains. In *Computer Networks*, 33(1), 377-386.

.