Computationally efficient PageRank algorithm exploiting graph sparsity

Dmitry Kamzolov

Scientific advisor: Yury Maximov, Alexander Gasnikov
Moscow Institute of Physics and Technology
Department of Control and Applied Mathematics

20 May 2015

Purpose of research

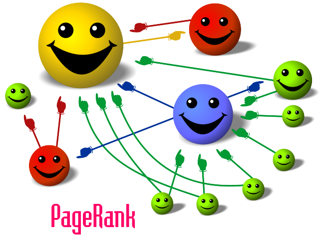
Checking theoretical estimation algorithm based on the Nesterov ideas ranking web-pages on the sparse graphs.

Application

It lets solve a large class ranking problems in logarithmic largest space time.

Notation

- Given a directed graph with n vertices.
- The vertices — sites.
- Oriented edge — links.
- Matrix $\mathbf{P}$ is adjacency matrix for the graph.
- Vector PageRank $\mathbf{x}$ is vector quantity characterizing the importance of each site.
- Sparsity coefficient s is the maximum number of non-zero elements in each column and each row of the matrix $\mathbf{P}$.



Problem Statement

The problem of finding the vector PageRank = The problem of searching left eigenvector $\mathbf{x}$, that

$$\mathbf{x}^T = \mathbf{x}^T \mathbf{P}, \text{ where } \sum_{k=1}^{n} x_k = 1.$$

The problem of searching left eigenvector = The optimization problem

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}\|_2^2 + \frac{1}{2}\sum_{k=1}^{n}(-x_k)_+^2 \longrightarrow \min_{\langle \mathbf{x}, \mathbf{e} \rangle = 1},$$

where matrix $\mathbf{A} = \mathbf{P}^T - \mathbf{I}$, $\mathbf{I}$ - the identity matrix, $\mathbf{e} = (1, \dots, 1)$,

Key ideas

- Matrix **A** is sparse with Lipschitz constant L.
- Componentwise descent.

$$x_{k+1} = x_k + \operatorname*{argmin}_{h:\langle h,e\rangle=0} \left\{ f(x_k) + \langle \nabla f(x_k), h\rangle + \frac{L}{2}\|h\|_1^2 \right\},$$

- The number of iterations $O(1/\epsilon^2)$
- The complexity of a single iteration $O(s^2 \ln n)$
- The complexity of the algorithm

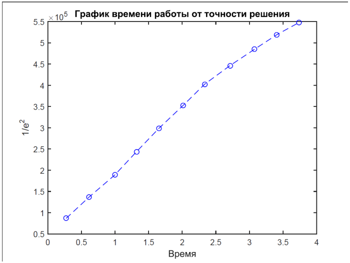
$$O\left(\frac{s^2 \ln n}{\varepsilon^2}\right).$$

- The goal of computational experiment is to check theoretical estimates of the complexity.
- **Input**:the dimension n, the sparsity coefficient s, the generated sparse matrix $\mathbf{A}$ Erdos-Renyi type with parameters n, s, the initial value of the function f_0 and the initial value of the gradient of g_0, precounted $\mathbf{A}^T\mathbf{A}$.
- **Output**: Page Rank $\mathbf{x}$, time.

Results

- Checking dependence $O(1/\epsilon^2)$. Timelines execution of the accuracy of the solution (1-2). Charts have a linear dependence. $\Longrightarrow$ The complexity $O(1/\epsilon^2)$. Theoretical estimates are confirmed.
- The checking of dependence. $O(s^2 \ln n)$. Unable to use the sparsity of the matrix due to the nature of MatLab and memory arrays. Theoretical estimates have not been confirmed.
- Getting Dependence $O(n)$. Timeline of execution of of the dimension (3). Charts have a linear dependence. $\Longrightarrow$ complexity $O(n)$.
- The final complexity: $O(n/\epsilon^2)$

Charts



pic.1 The dependence of the runtime by the number of iterations of the algorithm for $n = 12500$ sites

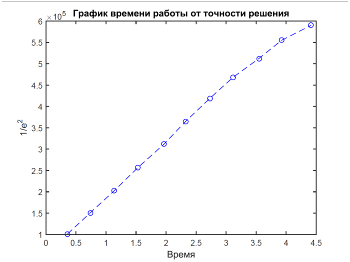


Рис.2 The dependence of the runtime by the number of iterations of the algorithm for $n = 15000$ sites

Графики

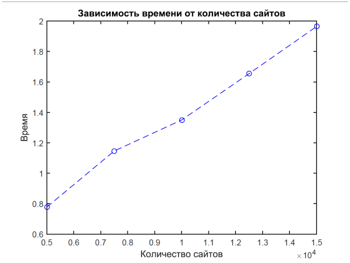


Рис.3 The dependence of the the runtime from the dimension n where 21000 iteration

Comparing with other algorithms

It received the complexity:

- $O\left(\frac{n}{\epsilon^2}\right)$, in practice without sparsity,
- $O\left(\frac{s^2 \ln n}{\varepsilon^2}\right)$ in theory with sparsity.
- The table shows the rate of convergence of the most advanced fast algorithms.
- Our algorithm has the lowest the complexity.

Table

Method	Condition	Complexity
Nazin-Polyak 2008	no	$O\left(\frac{n\ln(\frac{n}{\sigma})}{\epsilon^2}\right)$
Nesterov 2012	S	$O\left(\frac{sn\ln n}{\epsilon^2}\right)$
Juditsky et al 2009	no	$O\left(\frac{n\ln(\frac{n}{\sigma})}{\epsilon^2}\right)$
Grigoriadis-Hachiyan 2009	S	$O\left(\frac{s\ln n\ln(\frac{n}{\sigma})}{\epsilon^2}\right)$
Polyak-Tremba 2012	S	$\frac{2sn}{\epsilon}$
Gasnikov-Dvurechensky 2015	S	$O\left(\frac{s^2\ln n}{\epsilon^2}\right)$

Estimates of the rate of convergence of algorithms PageRank, where S sparsity-condition.

Final

In our work:

- The research method of ranking web pages with sparse graphs.
- It is shown that the theoretical estimate of the number of steps corresponds to the experimental data.
- It is shown that the theoretical estimate of the complexity of the algorithm step does not correspond to the experimental data, due to the nature of programmatic implementation.
- Through the use of 1-norm achieved $O(n)$ arithmetic operations on a step of the algorithm.
- The result is a new, and even with the deterioration estimates of complexity algorithm is one of the fastest algorithms.