2010/08/30

Experiments of RedSVD

Daisuke Okanohara

RedSVD

- RedSVD is a C++ library for matrix decompositions
  - New BSD license
  - http://code.google.com/p/redsvd/
- The implementation of the algorithm in [1]
  - [1] “Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions”, N. Halko, P.G. Martinsson, J. Tropp, arXiv 0909.4061

RedSVD (contd.)

- RedSVD differs from the original work [1].
  - To reduce the memory requirement further, redsvd sample both rows and columns and solve the smaller SVD problem.
- RedSVD is optimized for truncated SVD and for sparse matrices

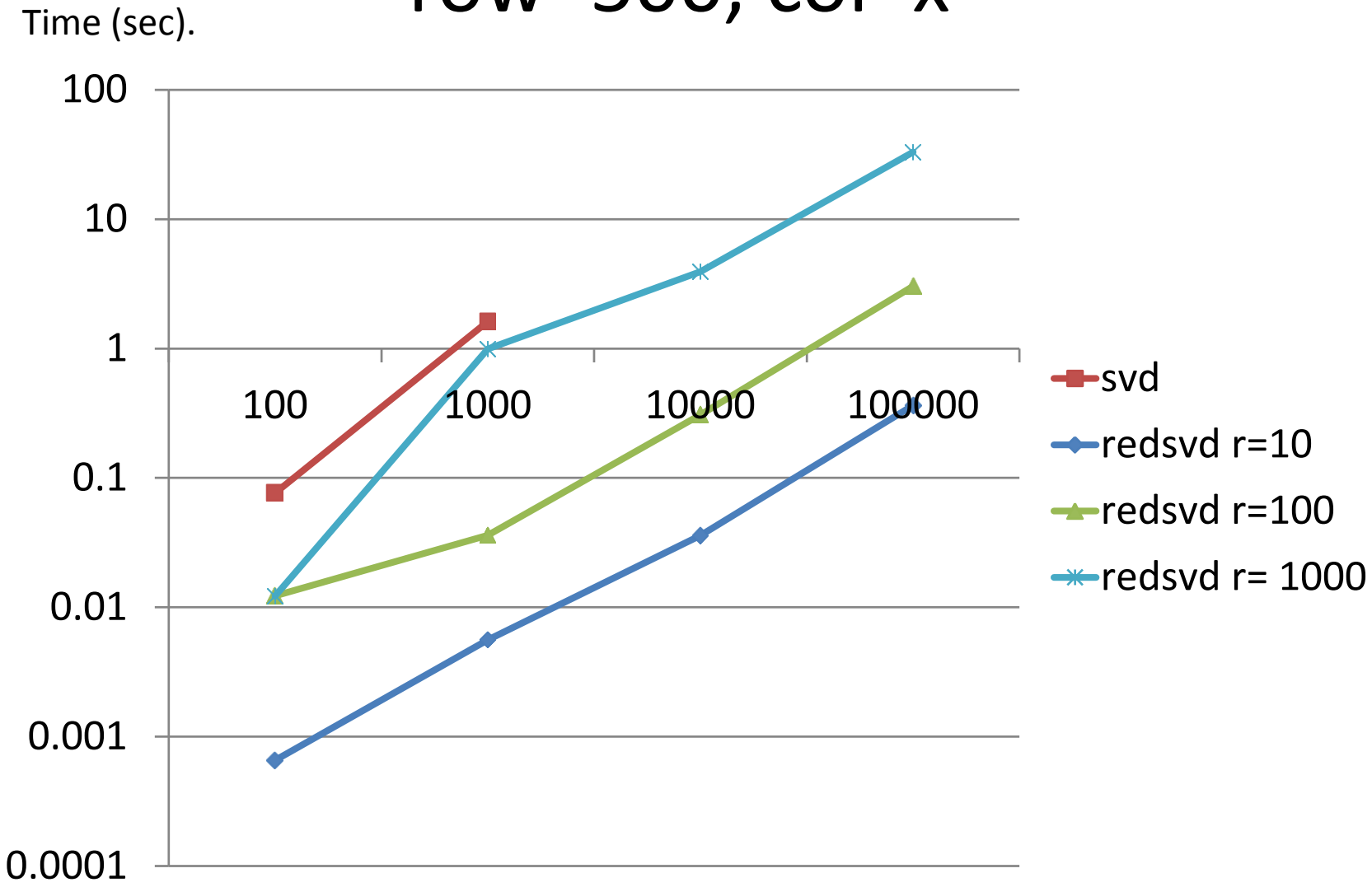
Experiments

- Conducted the following two experiments
  - Performance
  - Accuracy

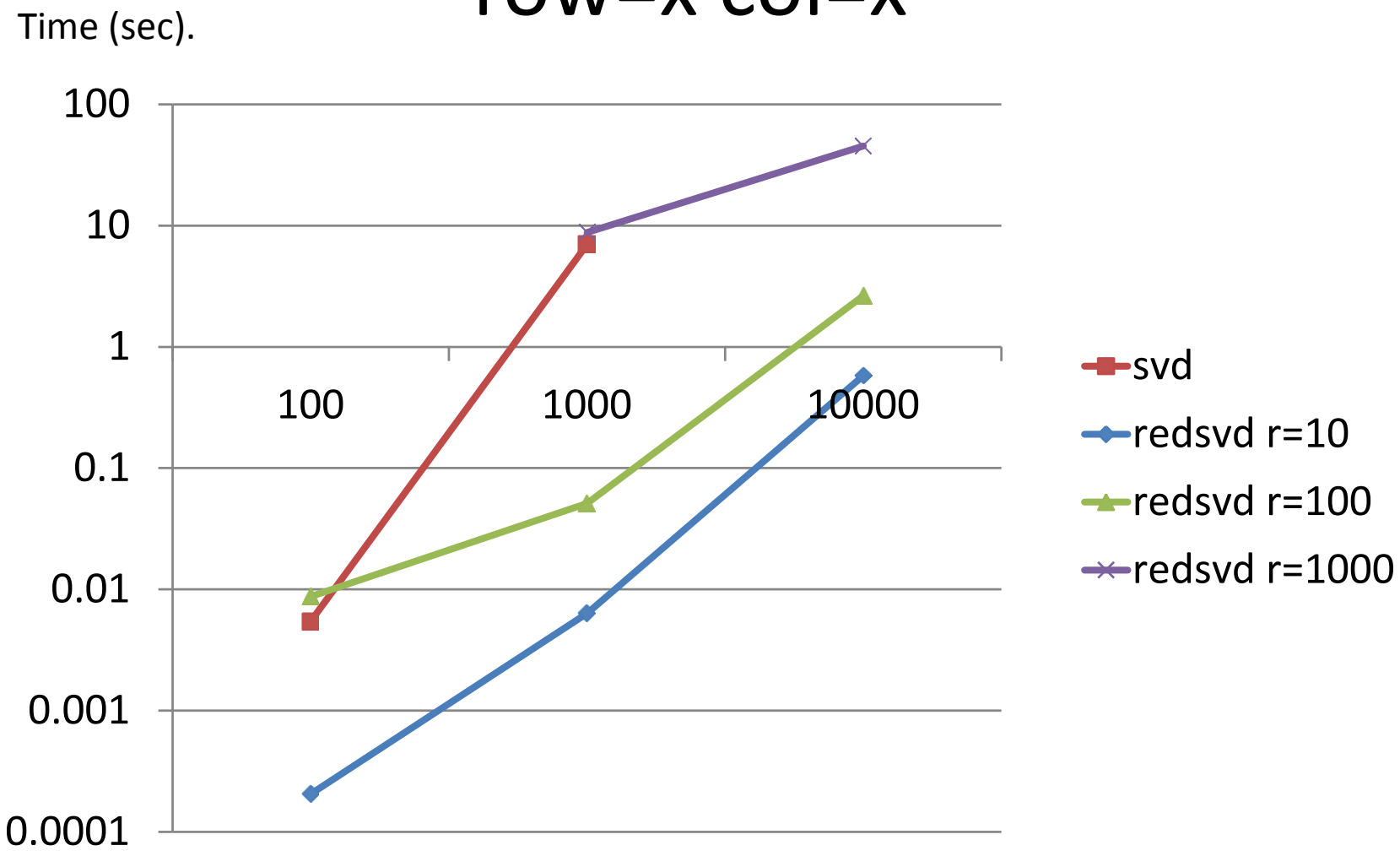
Performance Test Setup

- Compare the result for the following cases
  - Dense Matrix
    - # row is fixed, and # col is increased
    - Square matrix
  - Sparse Matrix
    - Square matrix and the nonzero ratio is changed
  - svd : Eigen::SVD Version 3.0
  - redsvd : REDSVD::RedSVD Version 0.0.3
  - Some results are not examined due to the lack of memory requirement

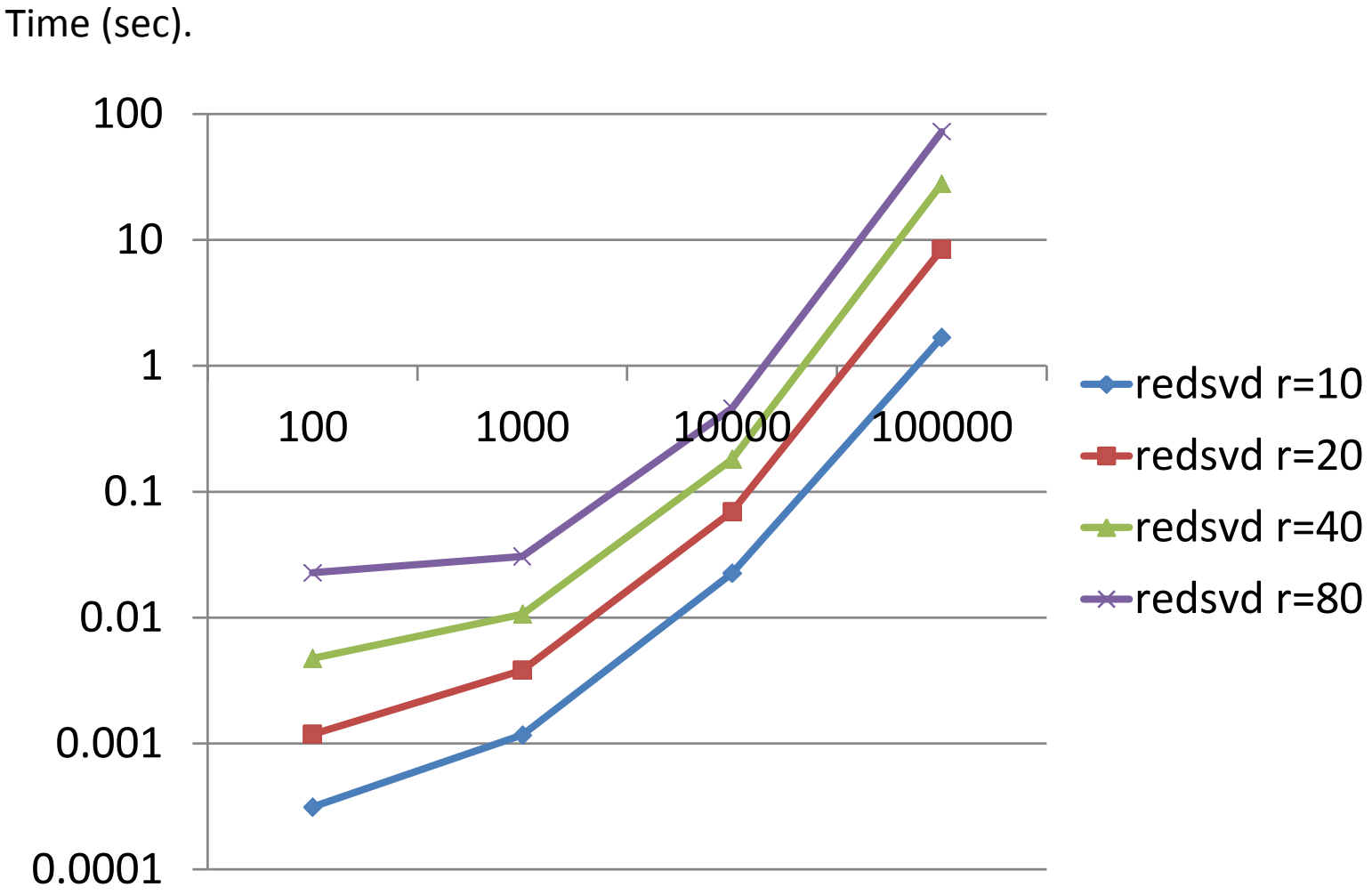
The result of SVD for dense matrices row=500, col=x



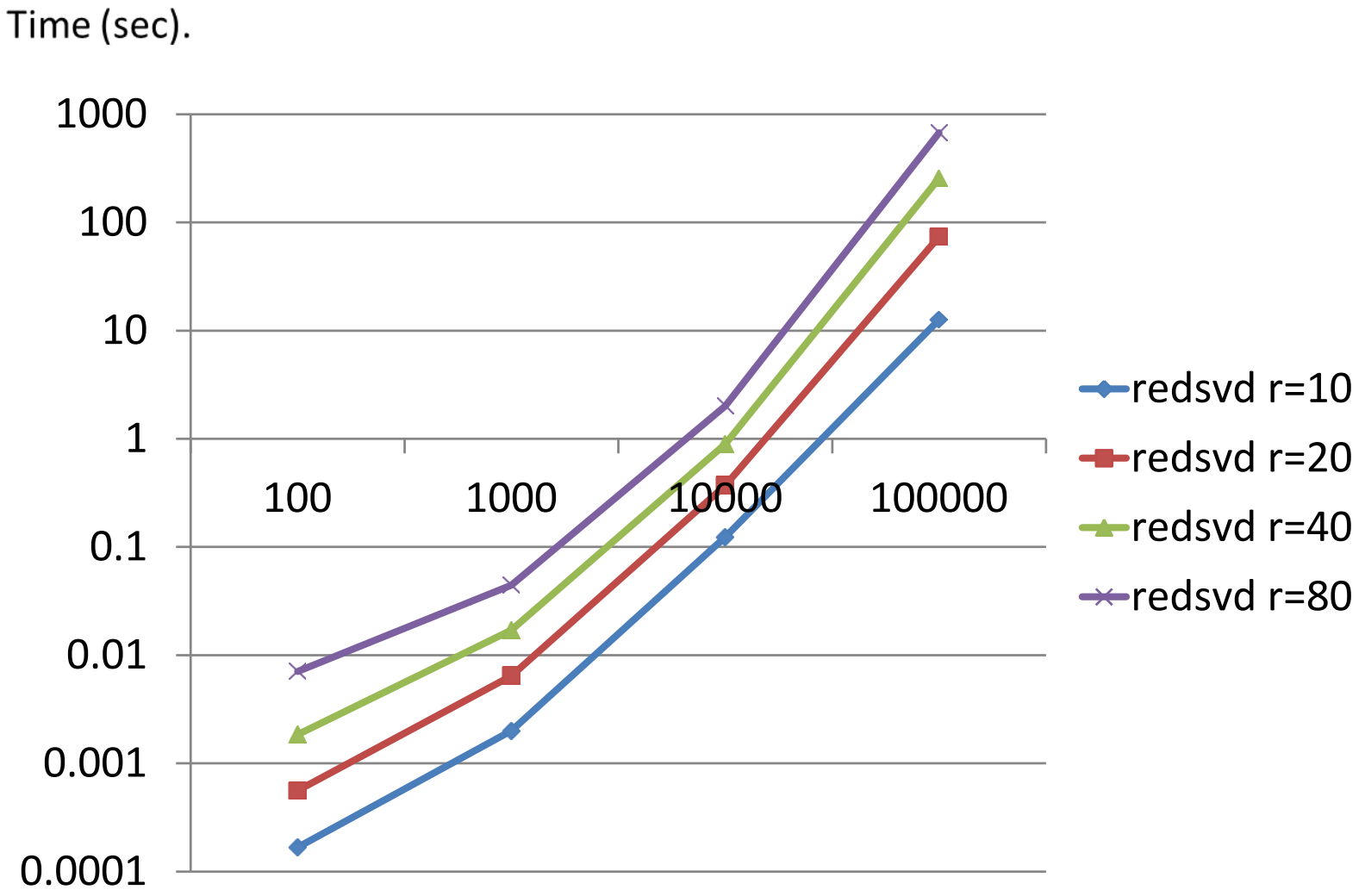
The result of SVD for dense matrices row=x col=x



The result of SVD for sparse matrices row = x, col = x, nonZero ratio = 0.1%



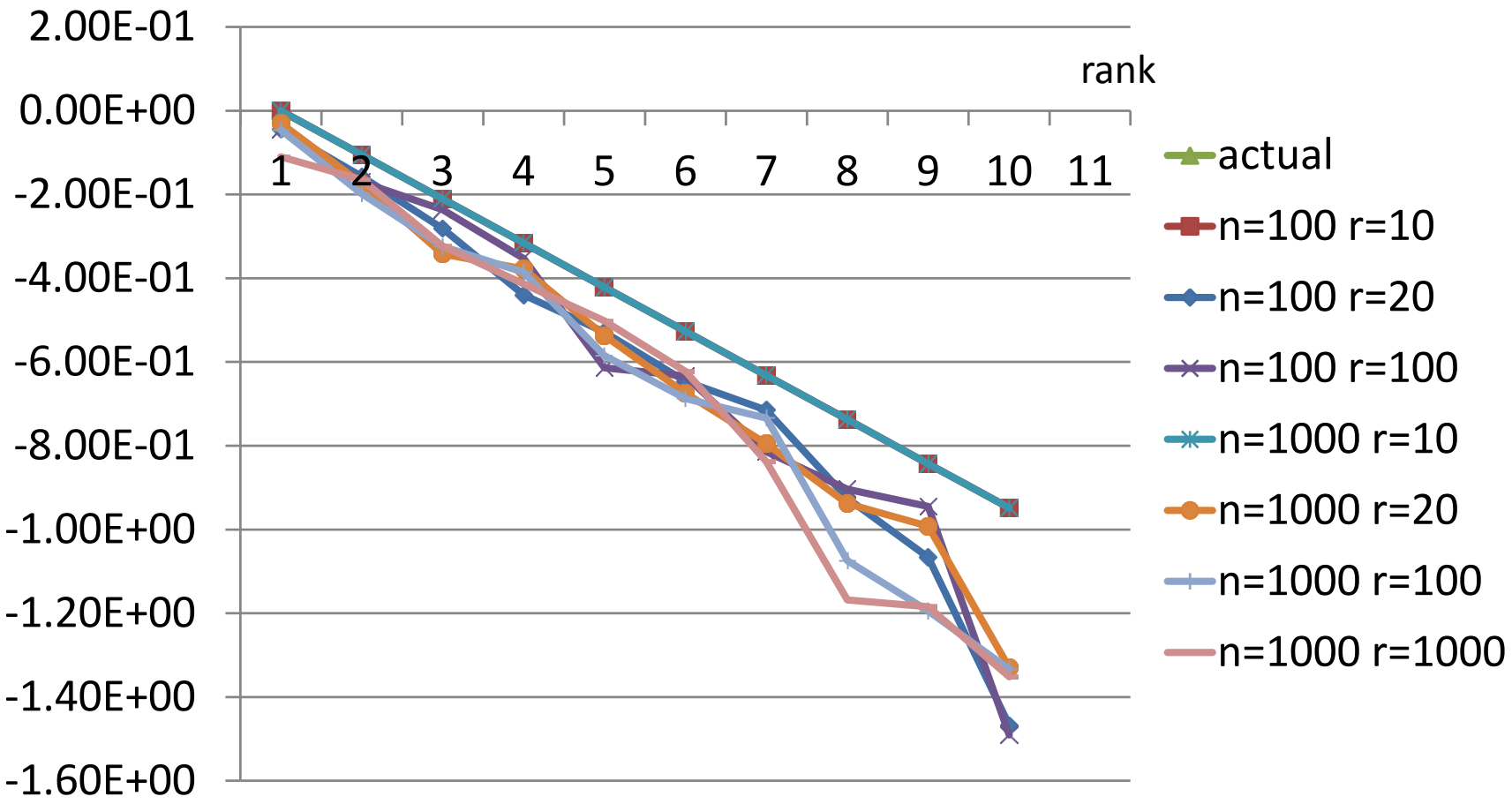
The result of SVD for sparse matrices
row = x, col = x, nonZeroRatio = 1%



Accuracy Test Setup

- Generate random square matrices U and V
  - These are ortho-normalized by Gram-Schmidt
- Set a singular vector as $S_i = 0.9^i$
- Set a sample matrix $A := USV^T$
- Compute a SVD of $A := U_t S_t V_t^T$ with top-10 singular values
  - n : the row/col of A
  - r : the actual rank of A

Singular Values



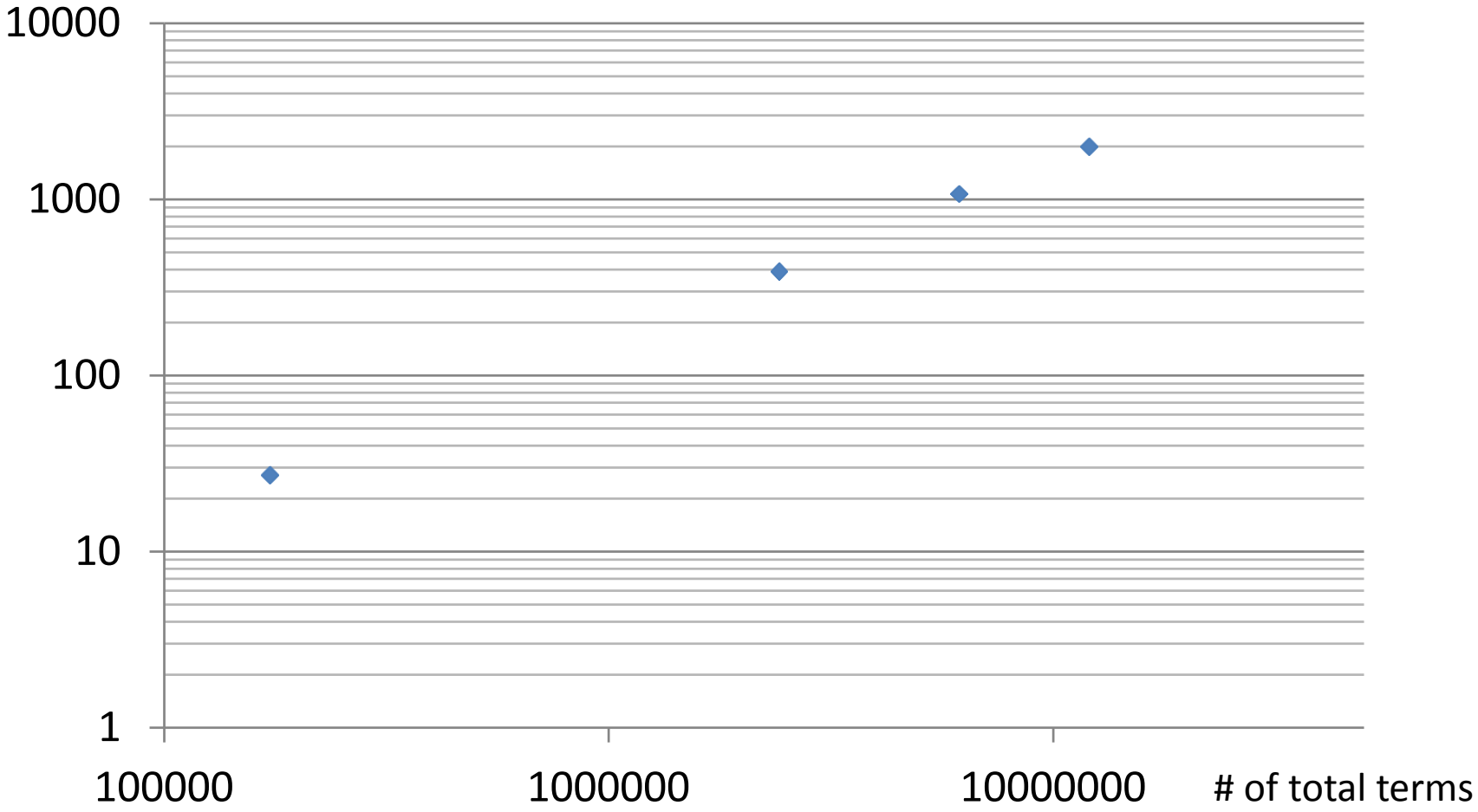
Order of Magnitude = $\log(S_i)$

LSA
(SVD for Doc-Term Matrix)

- Data: English Wikipedia
- A_{ij} = I(term j is appeared in a doc i)
  - I(x) returns 1 if x is true and 0 otherwise
  - A is very sparse; the nonzero ratio is 0.2% - 0.5%
- SVD of A is known as Latent Semantic Analysis

The result of LSA

Time Performance of RedSVD



- The numbers of docs are 3560, 46857, 118110, 233717, and those of terms are 27106, 147144, 261495, 402239