Experiments with Randomized Algorithms in the Text to Matrix Generator Toolbox

Eugenia Maria Kontopoulou, Dimitrios Zeimpekis
and Efstratios Gallopoulos

Department of
Compute Engineering and Informatics
University of Patras

6th International Conference of the ERCIM WG on
Computational and Methodological Statistics (ERCIM 2013)

London, 15/12/2013

Outline

The Big Data Problem

Large & Complex data sets $\left\{ \begin{array}{l} \text{Storage Problems} \\ \text{Interpretation Problems} \end{array} \right.$

The Big Data Problem

Large & Complex data sets $\begin{cases} \text{Storage Problems} \\ \text{Interpretation Problems} \end{cases}$

$\Downarrow$ $\Downarrow$

Secondary Memory $\longrightarrow$ Slow Communications

High Dimensional & Noisy Data

The Big Data Problem

Large & Complex data sets
- Storage Problems
- Interpretation Problems

$\Downarrow$ $\Downarrow$

Secondary Memory $\longrightarrow$ Slow Communications

High Dimensional & Noisy Data

$\Downarrow$

Processing / Managing Methods

The Big Data Problem

Large & Complex data sets $\left\{ \begin{array}{l} \text{Storage Problems} \\ \text{Interpretation Problems} \end{array} \right.$

$\Downarrow$ $\Downarrow$

Secondary Memory $\longrightarrow$ Slow Communications

High Dimensional & Noisy Data

$\Downarrow$

Processing / Managing Methods

$\Downarrow$ $\Downarrow$

Reduce Passes & Blocks in RAM

Dimensionality Reduction

The Big Data Problem

Large & Complex data sets $\begin{cases} \text{Storage Problems} \\ \text{Interpretation Problems} \end{cases}$

$\Downarrow$ $\Downarrow$

Secondary Memory $\longrightarrow$ Slow Communications

High Dimensional & Noisy Data

$\Downarrow$

Processing / Managing Methods

$\Downarrow$ $\Downarrow$

Reduce Passes & Blocks in RAM

Dimensionality Reduction

$\Downarrow$

Randomized Techniques

Randomization inroads into Matrix Computations

BLENDENPIK: SUPERCHARGING LAPACK'S LEAST-SQUARES SOLVER*

HAIM AVRON†, PETAR MAYMOUNKOV‡, AND SIVAN TOLEDO†

Abstract. Several innovative random-sampling and random-mixing techniques for solving problems in linear algebra have been proposed in the last decade, but they have not yet made a significant impact on numerical linear algebra. We show that by using a high-quality implementation of one of these techniques, we obtain a solver that performs extremely well in the traditional yardsticks of numerical linear algebra: it is significantly faster than high-performance implementations of existing

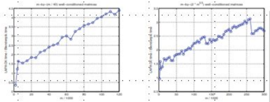


FIG. 1.1. *Comparison between LAPACK and the new solver for increasingly larger matrices. Graphs show the ratio of LAPACK's running time to Blendenpik's running time on random matrices with two kinds of aspect ratios.*

An efficient distributed randomized solver wit[h] application to large dense linear systems

Marc Baboulin*, Dulceneia Becker†, George Bosilca‡, Anthon[y] Danalis§, Jack Dongarra¶

Project-Team Grand-Large

Research Report n° 8043 — Aout 2012 — 20 pages

Abstract: Randomized algorithms are gaining ground in high-performance com[puting applica]tions as they have the potential to outperform deterministic methods, while still pro[viding accurate] results. We propose a randomized algorithm for distributed multicore architecture[s to efficiently] solve large dense symmetric indefinite linear systems that are encountered, for i[nstance, in pa]rameter estimation problems or electromagnetism simulations. This solver combi[nes]

RANDOMIZED ALGORITHMS FOR ESTIMATING THE TRACE OF AN IMPLICIT SYMMETRIC POSITIVE SEMI-DEFINITE MATRIX

HAIM AVRON AND SIVAN TOLEDO

Accelerating linear system solutions using randomization techniques

Marc Baboulin[1], Jack Dongarra[2,3,4], Julien Herrmann[1], and Stanimire Tomov[2]

Randomized Extended Kaczmarz for Solving Least Squares

Anastasios Zouzias* Nikolaos M. Freris†

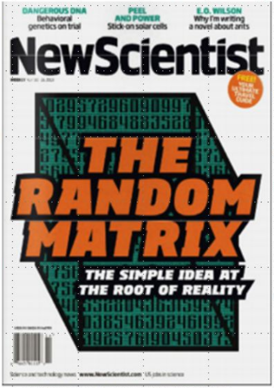
Site Map

PNAS

Proceedings of the National Academy of Sciences of the United States of America

CUR matrix decompositions for improved data analysis

original data are). We present an algorithm that preferentially chooses columns and rows that exhibit high "statistical leverage" and, thus, in a very precise statistical sense, exert a disproportionately large "influence" on the best low-rank fit of the data matrix. By selecting columns and rows in this manner, we obtain ...ved relative-error and constant-factor approximation guarantees in worst-



random matrix theory. Originally developed more than 50 years ago to describe the energy levels of atomic nuclei, the theory is turning up in everything from inflation rates to the behaviour of solids. So much so that many researchers believe that it points to some kind of deep pattern in nature that we don't yet understand. "It really does feel like the ideas of random matrix theory are somehow buried deep in the heart of nature," says electrical engineer Raj Nadakuditi of the University of Michigan, Ann Arbor.

All of this, oddly enough, emerged from an effort to turn physicists' ignora[nce] into an advantage. In 1956, when ...

FAST MONTE CARLO ALGORITHMS FOR MATRICES I: APPROXIMATING MATRIX MULTIPLICATION*

PETROS DRINEAS†, RAVI KANNAN‡, AND MICHAEL W. MAHONEY§

[FA]ST MONTE CARLO ALGORITHMS FOR MATRICES III: COMPUTING A COMPRESSED APPROXIMATE MATRIX DECOMPOSITION*

PETROS DRINEAS†, RAVI KANNAN‡, AND MICHAEL W. MAHONEY§

SAMPLING FROM LARGE MATRICES: AN APPROACH THROUGH GEOMETRIC FUNCTIONAL ANALYSIS

MARK RUDELSON AND ROMAN VERSHYNIN

Effective Resistances, Statistical Leverage, and Applications to Linear Equation Solving

Petros Drineas * Michael W. Mahoney †

FINDING STRUCTURE WITH RANDOMNESS: PROBABILISTIC ALGORITHMS FOR CONSTRUCTING APPROXIMATE MATRIX DECOMPOSITIONS

N. HALKO†, P. G. MARTINSSON†, AND J. A. TROPP‡

RELATIVE-ERROR *CUR* MATRIX DECOMPOSITIONS*

PETROS DRINEAS†, MICHAEL W. MAHONEY‡, AND S. MUTHUKRISHNAN§

Randomization inroads into Matrix Computations

"*One great and underused technique at this scale is sampling.* ***For most things that you want to do with data, 100.000 randomly selected rows is as good as 10.000.000 rows*** *and working at R or SPSS scale allows for a much faster analysis cycle*"

Lukas Biewald

CEO of CrowdFlower

Documents

Labels	Titles
B1	Identifying users of social networks from their data footprint: An application of large-scale matrix factorizations
B2	Data fusion based on coupled matrix and tensor factorizations
B3	On incremental deterministic methods for dominant space estimation for large data sets
B4	Fast projection methods for robust separable nonnegative matrix factorization
B5	Experiments with randomized algorithms in the text to matrix generator toolbox

Term-Document Matrix (TDM)

33×5

terms	B1	B2	B3	B4	B5
	Documents				
algorithm	0	0	0	0	2.3219
applic	2.3219	0	0	0	0
base	0	2.3219	0	0	0
coupl	0	2.3219	0	0	0
data	0.7370	0.7370	0.7370	0	0
determinist	0	0	2.3219	0	0
domin	0	0	2.3219	0	0
estim	0	0	2.3219	0	0
experi	0	0	0	0	2.3219
factor	0.7370	0.7370	0	0.7370	0
fast	0	0	0	2.3219	0
footprint	2.3219	0	0	0	0
fusion	0	2.3219	0	0	0
gener	0	0	0	0	2.3219
identifi	2.3219	0	0	0	0
increment	0	0	2.3219	0	0
larg	1.3219	0	1.3219	0	0
matrix	0.3219	0.3219	0	0.3219	0.3219
method	0	0	1.3219	1.3219	0
network	2.3219	0	0	0	0
nonneg	0	0	0	2.3219	0
project	0	0	0	2.3219	0
random	0	0	0	0	2.3219
robust	0	0	0	2.3219	0
scale	2.3219	0	0	0	0
separ	0	0	0	2.3219	0
set	0	0	2.3219	0	0
social	2.3219	0	0	0	0
space	0	0	2.3219	0	0
tensor	0	2.3219	0	0	0
text	0	0	0	0	2.3219
toolbox	0	0	0	0	2.3219
user	2.3219	0	0	0	0

✓ tf-idf

✓ Stemming

Text Data
. . . for text mining tasks

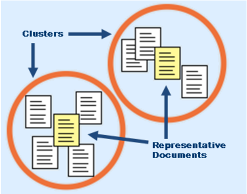
Retrieval

Text Data

. . . for text mining tasks

Retrieval

Clustering

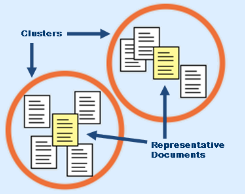


Text Data

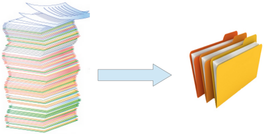
. . . for text mining tasks

Retrieval

Clustering



Classification

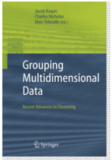


Text to Matrix Generator

TMG: A MATLAB Toolbox for Generating Term-Document Matrices from Text Collections (ZG06)

What is TMG:

- Toolbox developed in University of Patras for text mining tasks over document collections
- Educational and Research tool



Grouping Multidimensional Data

Recent Advances in Clustering
Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds.)
2006, XII, 268 p.

Grouping Multidimensional Data
2006, pp 187-210

TMG: A MATLAB Toolbox for Generating Term-Document Matrices from Text Collections

D. Zeimpekis, E. Gallopoulos

Text to Matrix Generator

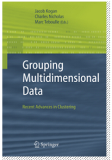
What is TMG:

- Toolbox developed in University of Patras for text mining tasks over document collections
- Educational and Research tool

Implementation:

- over 17.000 lines of `matlab` and `perl`
- takes advantage from sparse technology provided by MATLAB
- first version by Zeimpekis (´06)

TMG: A MATLAB Toolbox for Generating Term-Document Matrices from Text Collections (ZG06)



Grouping Multidimensional Data

Recent Advances in Clustering
Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds.)
2006, XII, 268 p.

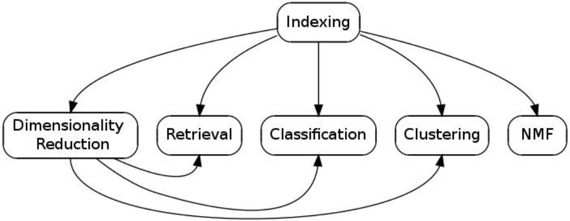
Grouping Multidimensional Data
2006, pp 187-210

TMG: A MATLAB Toolbox for Generating Term-Document Matrices from Text Collections

D. Zeimpekis, E. Gallopoulos

Six basic modules:

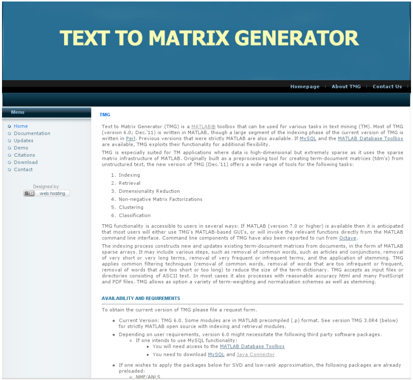
1. Indexing
2. Dimensionality Reduction
3. Non-Negative Matrix Factorizations
4. Retrieval
5. Clustering
6. Classification



How can I find TMG?

Free under request from:

http://scgroup20.ceid.upatras.gr:8000/tmg/


TEXT TO MATRIX GENERATOR

Homepage | About TMG | Contact Us

Menu

- Home
- Documentation
- Updates
- Demo
- Citations
- Download
- Contact

Designed by: web hosting

TMG

Text to Matrix Generator (TMG) is a MATLAB® toolbox that can be used for various tasks in text mining (TM). Most of TMG (version 6.0; Dec.'11) is written in MATLAB, though a large segment of the indexing phase of the current version of TMG is written in Perl. Previous versions that were strictly MATLAB are also available. If MySQL and the MATLAB Database Toolbox are available, TMG exploits their functionality for additional flexibility.

TMG is especially suited for TM applications where data is high-dimensional but extremely sparse as it uses the sparse matrix infrastructure of MATLAB. Originally built as a preprocessing tool for creating term-document matrices (tdm's) from unstructured text, the new version of TMG (Dec.'11) offers a wide range of tools for the following tasks:

1. Indexing
2. Retrieval
3. Dimensionality Reduction
4. Non-negative Matrix Factorizations
5. Clustering
6. Classification

TMG functionality is accessible to users in several ways: If MATLAB (version 7.0 or higher) is available then it is anticipated that most users will either use TMG's MATLAB-based GUI's, or will invoke the relevant functions directly from the MATLAB command line interface. Command line components of TMG have also been reported to run from Octave.

The indexing process constructs new and updates existing term-document matrices from documents, in the form of MATLAB sparse arrays. It may include various steps, such as removal of common words, such as articles and conjunctions, removal of very short or very long terms, removal of very frequent or infrequent terms, and the application of stemming. TMG applies common filtering techniques (removal of common words, removal of words that are too infrequent or frequent, removal of words that are too short or too long) to reduce the size of the term dictionary. TMG accepts as input files or directories consisting of ASCII text. In most cases it also processes with reasonable accuracy html and many PostScript and PDF files. TMG allows as option a variety of term-weighting and normalization schemes as well as stemming.

AVAILABILITY AND REQUIREMENTS

To obtain the current version of TMG please file a request form.

- Current Version: TMG 6.0. Some modules are in MATLAB precompiled (.p) format. See version TMG 3.0R4 (below) for strictly MATLAB open source with indexing and retrieval modules.
- Depending on user requirements, version 6.0 might necessitate the following third party software packages.
  - If one intends to use MySQL functionality:
    - You will need access to the MATLAB Database Toolbox
    - You need to download MySQL and Java Connector
- If one wishes to apply the packages below for SVD and low-rank approximation, the following packages are already preloaded:
  - NMF/ANLS

More than 4000 requests worldwide . . .

Caltech, Maryland, Purdue, Carnegie Mellon, Tennessee, Berkeley, Texas, Minnesota, Stanford, MIT, Columbia Renault, Leuven, Max-Planck, Michigan, Oxford, Philips, Princeton, Illinois, ETH, RPI, Los Alamos, Toronto, Queen Mary, St Andrews, Colorado, Texas, Livermore, Mathworks, Yahoo!, . . .

Outline

Retrieval

Latent Semantic Analysis

Data Explosion

$\Downarrow$

Big Data Collections $\Longrightarrow$ **Parsing & Processing** $\Longrightarrow$ Large & Sparse Term-Document Matrices

$\Downarrow$

Dimensionality Reduction $\Longleftarrow$ Difficult Management

$\Downarrow$

Low Rank Approximation

Low Rank Approximation

Given an $m \times n$ matrix A and a rank parameter $k \ll \min\{m, n\}$, the Low-Rank Approximation problem is to find a matrix Z of rank k such that $\|A - Z\|_{2,F}$ is sufficient small.

Eckart-Young Theorem

The minimization problem:

$$\min_{rank(Z)=k} \|A - Z\|_{2,F}$$

has a solution given by the truncated SVD:

$$Z = A_k = U_k S_k V_k^\top$$

Truncated Versions of SVD

+ reveals *latent semantic structure*
+ construct *orthogonal bases* for the terms (rows) and documents (columns)

Low Rank Approximation

Given an $m \times n$ matrix A and a rank parameter $k \ll \min\{m, n\}$, the Low-Rank Approximation problem is to find a matrix Z of rank k such that $\|A - Z\|_{2,F}$ is sufficient small.

Eckart-Young Theorem

The minimization problem:

$$\min_{rank(Z)=k} \|A - Z\|_{2,F}$$

has a solution given by the truncated SVD:

$$Z = A_k = U_k S_k V_k^\top$$

Truncated Versions of SVD

+ reveals *latent semantic structure*
+ construct *orthogonal bases* for the terms (rows) and documents (columns)

- requires the *entire matrix in RAM*
- lacks *interpretability*
- results in *dense factors*
- is *time* inefficient

1. Time Efficient Algorithms

$\Downarrow$

SPQR
(BPS05)

2. Storage Efficient Algorithms

$\Downarrow$

SDD
(KO00)

3. Interpretable Algorithms

"Matrices are about their columns and rows" (G. Strang)

Concerving

NMF

Deterministic

Pseudo-Skeleton
(GTZ97)

SCRA
(BPS05)

Randomized

CUR
(DMM08)

CX
(DMM08)

CUR/CX Algorithms (DMM08)

Basic Citation:

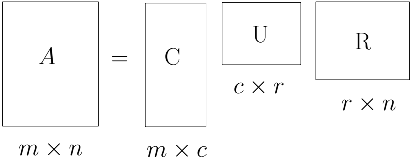
SIAM J. MATRIX ANAL. APPL.
Vol. 30, No. 2, pp. 844–881

RELATIVE-ERROR *CUR* MATRIX DECOMPOSITIONS*

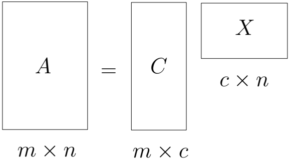
PETROS DRINEAS†, MICHAEL W. MAHONEY‡, AND S. MUTHUKRISHNAN§

Basic Idea:
Randomly select columns (and) rows of *A* based on probability vectors constructed by dominant right/left singular vectors (2nd Generation).

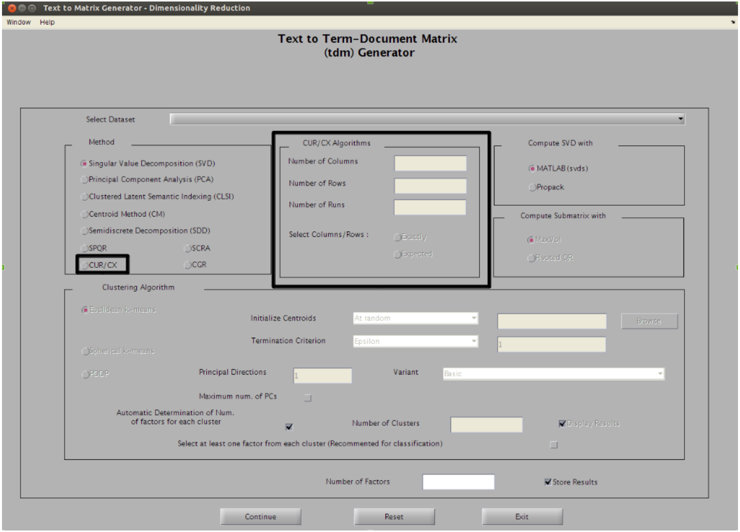
CUR



CX



CUR/CX Algorithms in TMG
Graphical Interface



GUI usage

1. Select `CUR/CX` radio button
2. Block CUR/CX Algorithm is activated
3. `Number of Columns` & `Number of Rows` fields define the algorithm

Algorithm	Number of Columns	Number of Rows
CX	✓	×
CUR	✓	✓
Error	×	✓
Error	×	×

4. `Number of Runs` field defines the repetitions of the algorithm
5. Random selection policies:

Selection Method	Description
`Exactly`	select the exact desired number of rows/columns
`Expected`	select in expectation the desired number of rows/columns

Data Explosion

$\Downarrow$

Big Data Collections $\Longrightarrow$ Parsing & Processing $\Longrightarrow$ Large & Sparse Term-Document Matrices

$\Downarrow$

Difficult Management & Storage

Consider a document-term matrix A and a query vector q. We seek documents similar to the query.

Similarity Measures

1. Euclidean Distance

$$Aq$$

2. Cosine

$$\frac{Aq}{||A||\,||q||}$$

$$\Downarrow$$

Aq

Data Intensive

$$\Downarrow$$

Reduction of the size of A and q

BMM/PIP Algorithms

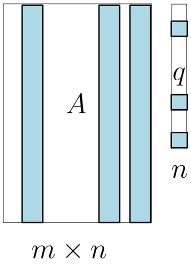
Basic Matrix Multiplication (DKM04)

SIAM J. COMPUT.
Vol. 36, No. 1, pp. 132–157

FAST MONTE CARLO ALGORITHMS FOR MATRICES I: APPROXIMATING MATRIX MULTIPLICATION*

PETROS DRINEAS†, RAVI KANNAN‡, AND MICHAEL W. MAHONEY§

Basic Idea:
Randomly select columns of *A* and elements of *q* based on one probability vector constructed by the euclidean norms of the columns of *A* and *q*.



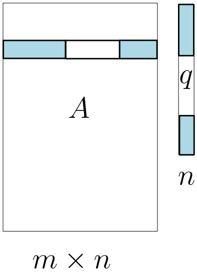
Probabilistic Inner Product (EB+11)

SIAM J. SCI. COMPUT.
Vol. 33, No. 4, pp. 1689–1706

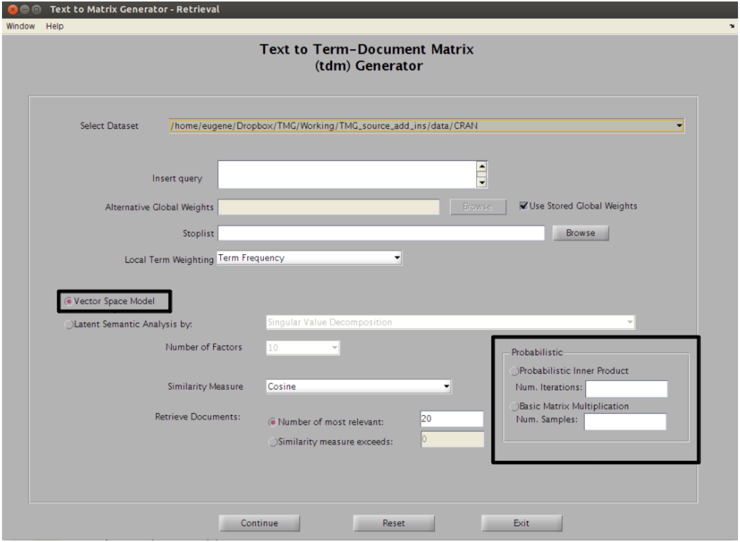
IMPORTANCE SAMPLING FOR A MONTE CARLO MATRIX MULTIPLICATION ALGORITHM, WITH APPLICATION TO INFORMATION RETRIEVAL*

SYLVESTER ERIKSSON-BIQUE†, MARY SOLBRIG‡, MICHAEL STEFANELLI§, SARAH WARKENTIN¶, RALPH ABBEY‖, AND ILSE C. F. IPSEN‖

Basic Idea:
Randomly select columns of *A* and elements of *q* based on *n* probability vectors constructed by the euclidean norms of the columns of *A* and *q*.



BMM/PIP Algorithms in TMG
Graphical Interface



GUI usage

1. Select `Vector Space Model` radio button
2. Block Probabilistic is activated
3. Algorithm Selection

Algorithm	Fields	Characteristics
Probabilistic Inner Product	`Num. Iterations`	repetitions
		sampling 1% of elements
Basic Matrix Multiplication	`Num. Samples`	samples

Outline

CRANFIELD Collection (CRAN) (2568×1398)

Cranfield collection of 1398 documents, 225 queries

Parsing Options			
Terms		**Min Global Frequency**	**Weights (NGL)**
numerics removal	stemming	2	`cae`

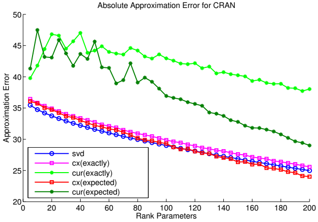
DR Parameters			
Rank	$k = [5 : 5 : 200]$		
	$k = [200 : 50 : 1300]$		
CUR/CX	**Num. Runs**	**Num. Rows**	**Num. Columns**
	5	$4k$ (2500 if $2k > 2568$)	$2k$ (1300 if $2k > 1398$)

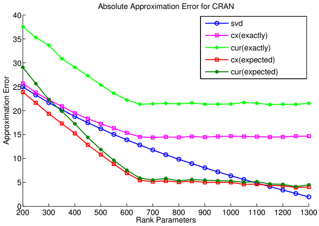
VSM Parameters	
PIP	**Num. Iterations**
	10
BMM	**Num. Samples**
	$[10 : 50 : 2500]$

System Specifications

System Specifications			
Processor	**RAM**	**MATLAB**	**OS**
Intel Core i5 2500 (4Cores) @3.3GHz	16 GB	R2012b	Debian 3.2.12-1

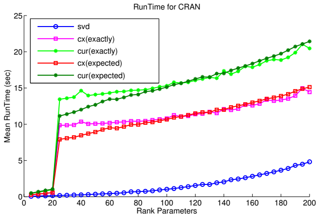
DR Approximation Error

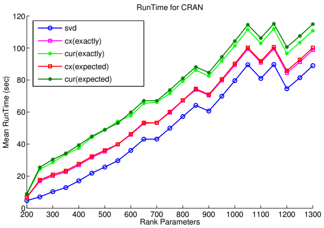




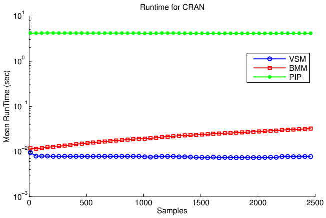
Running Time

Latent Semantic Indexing



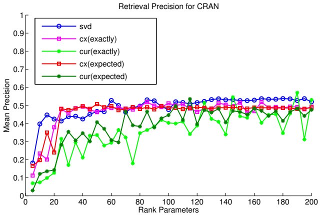


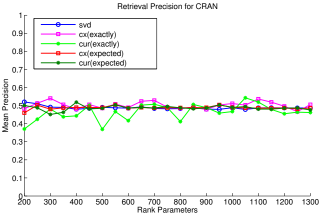
Vector Space Model



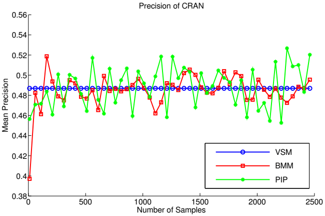
Retrieval Accuracy

Latent Semantic Indexing



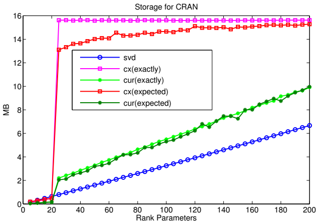


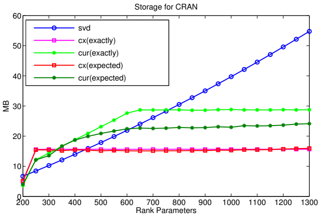
Vector Space Model



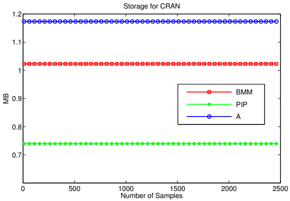
Storage

Latent Semantic Indexing





Vector Space Model



Goals:

- ☺ rapid familiarization of randomized techniques
- ☺ prototyping and testing new algorithms

Summary

Goals:

- ☺ rapid familiarization of randomized techniques
- ☺ prototyping and testing new algorithms

Landscape:

	Tasks	Categories
Term Document Matrix	**Dimensionality Reduction**	**General DRM**
		NMF
	Retrieval	**Vector Space Model**
		Latent Semantic Analysis
	Clustering	
	Classification	

■ Present Work

■ Future Work/ Work in Progress

Work in Progress:

- Parallel implementations
- Construction of a complete & extensible Randomization Module
  - ✓ Randomized NMF
  - ✓ Randomized Clustering
  - ✓ Randomized Classification
- Incorporating methods for handling efficiently the coupling

Text Data + **Algorithms** + **Friendly MATLAB Interface**

Work in Progress:

- Parallel implementations
- Construction of a complete & extensible Randomization Module
  - ✓ Randomized NMF
  - ✓ Randomized Clustering
  - ✓ Randomized Classification
- Incorporating methods for handling efficiently the coupling

Text Data + **Algorithms** + **Friendly MATLAB Interface**

Your Algorithms are welcome!

☺

Thank you!

Questions ?

[BPS05] M. W. Berry, S. A. Pulatova, and G. W. Stewart. ``Algorithm 844: Computing sparse reduced-rank approximations to sparse matrices''. In: *ACM TOMS* 31.2 (June 2005), 252–269.

[DKM04] P. Drineas, R. Kannan, and M. W. Mahoney. ``Fast Monte Carlo algorithms for matrices I: Approximating Matrix Multiplication''. In: *SISC* 36.1 (2004), 132–157.

[DMM08] P. Drineas, M. Mahoney, and S. Muthukrishnan. ``Relative-Error CUR Matrix Decompositions''. In: *SIMAX* 30.2 (Sept. 2008), 844–881.

[EB+11] S. Eriksson-Bique et al. ``Importance Sampling for a Monte Carlo Matrix Multiplication Algorithm, with Application to Information Retrieval''. In: *SISC* 33.4 (2011), 1689–1706.

[GTZ97] S. Goreinov, E. Tyrtyshnikov, and N. Zamarashkin. ``A theory of pseudoskeleton approximations''. In: *Linear Algebra and its Applications* 261.1-3 (1997), 1–21.

[KO00] T. Kolda and D. O'Leary. ``Algorithm 805: Computation and Uses of the Semidiscrete Matrix Decomposition''. In: *ACM TOMS* 26.3 (Sept. 2000), 415–435.

[ZG06] D. Zeimpekis and E. Gallopoulos. ``TMG: A MATLAB toolbox for generating term document matrices from text collections''. In: *Grouping Multidimensional Data: Recent Advances in Clustering*. Ed. by J. Kogan, C. Nicholas, and M. Teboulle. Berlin: Springer, 2006, 187–210.