Searching the Web with Low Space Approximations

András A. Benczúr,
Computer and Automation Institute,
Hungarian Academy of Sciences

Joint work with Károly Csalogány, Dániel Fogaras,
Balázs Rácz, and Tamás Sarlós

May 19, 2006

Contents

- Efficient algorithms for Personalized PageRank
  [Fogaras–Rácz WAW 2004, SBCsFR WWW 2006]
- Similarity Search
  [Fogaras–Rácz WWW 2005 and SBCsFR]
- Relative Error Low-rank Matrix Approximation
  [Sarlós, manuscript]

What's Common?

- Relevance in Web IR
  - ranking, similarity, personalization
  - spam hunting [BCsSU AIRWeb 2005]
  - LSI, HITS
- Infeasible even to store: $n \times n$ matrices
  - Web Pages: n in order of ten billions
  - Experiments: Stanford WebBase $n = 80M$
- Randomized approximation
  - Typical solution: sampling
  - Our new results based on sketching and random projections

Personalized PageRank – Preliminaries and Sampling

Definition: random surfer with *teleportation distr.* r

$$\mathrm{PPR}_r(u) = c \cdot r(u) + (1-c) \sum_{vu \in E} \mathrm{PPR}_r(v)$$

Linearity: single page teleportation suffices

$$\mathrm{PPR}_r(u) = \sum_v r(u) \cdot \mathrm{PPR}_v(u)$$

Path summation:

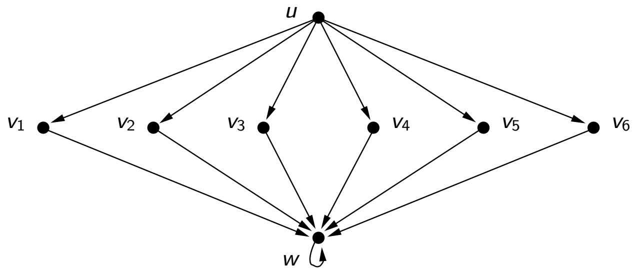
$$\mathrm{PPR}_u^{[k]}(v) = \sum_{k \geq 0} c \cdot (1-c)^k \sum_{v_0=u, v_1, \ldots, v_k=v} \frac{1}{(d^+(v_0) \cdots d^+(v_{k-1}))}$$

Sampling: random paths as above [Fogaras–Rácz, WAW 2004]

- First algorithm with no restriction on u
- Relative approx $(1 \pm \epsilon)$; out of bounds prob δ
- Uses $O(\epsilon^{-2} \log 1/\delta \log n)$ space

Personalized PageRank – Rounding and Sketching

Example



Power iteration propagates large variance downwards
Dynamic programming [Jeh–Widom WWW 2003] averages the error upward

$$\mathrm{PPR}_u = c\chi_u + (1-c) \cdot \sum_{v:(uv)\in E} \mathrm{PPR}_v / d^+(u).$$

Problem: small world, nonzeroes quickly grow in u's neighborhood

New results – Rounding and Sketching

Sloppy Attendant: round change down to nearest ϵuro

- Requires space $1/\epsilon \cdot \log n$ to store a sparse PPR_u vector
- Matching communication complexity lower bound for a top list query database

Drunken Surfer: mix up memories by random hash of pages

- Use $\log 1/\delta$ surfers and use minimum vote: Count-Min Sketch
- Dynamic programming over sketches by their linearity
- Space $1/\epsilon \log 1/\delta$ per page optimal for value queries

SimRank – Preliminaries and Sampling

"Two pages are similar if pointed to by similar pages" [Jeh–Widom KDD 2002]:

$$\mathrm{Sim}^{(k)}(u_1, u_2) = \begin{cases} (1-c) \cdot \frac{\sum \mathrm{Sim}^{(k-1)}(v_1,v_2)}{d^-(u_1) \cdot d^-(u_2)} & \text{if } u_1 \neq u_2 \\ 1 & \text{if } u_1 = u_2. \end{cases} \quad (1)$$

Path pair summation (incl. sampling [Fogaras–Rácz WWW 2005]) over

$$\begin{aligned} u = w_0, w_1, \ldots, w_{k'-1}, w_{k'} &= v_2 \\ u = w'_0, w'_1, \ldots, w'_{k'-1}, w'_{k'} &= v_1 \end{aligned}$$

Fully Personalized PageRank
Similarity Search
Relative Error Low-rank Matrix Approximation

SimRank – Reduction to Personalized PageRank

Version 0 reduction: count path pairs from v_1 and v_2 that may meet several times

$$\mathrm{Sim}^{(0)}_{v_1,v_2} = \sum_{k>0} (1-c)^k \sum_u \mathrm{RP}^{[k]}_{v_1}(u) \mathrm{RP}^{[k]}_{v_2}(u)$$

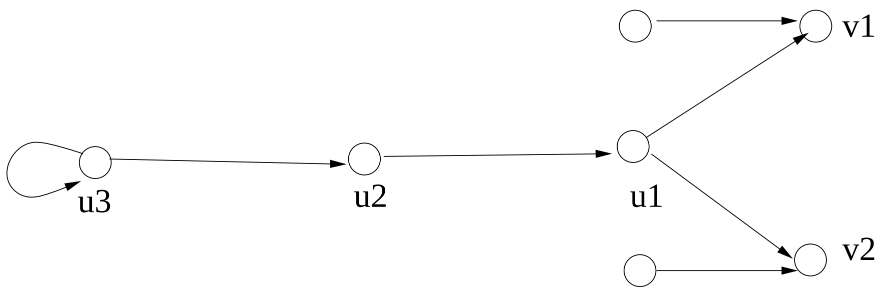
Self-similarity SimRank of *at least* $t+1$ meeting points

$$\mathrm{SSim}^{(t+1)}(v) = \sum_u \sum_{k>0} (1-c)^k \mathrm{RP}^{[k]}_v(u) \mathrm{RP}^{[k]}_v(u) \cdot \mathrm{SSim}^{(t)}(u)$$

Obtain SimRank by inclusion-exclusion of self-similarities
Converges for $1-c < 1/2$, technicalities to carry through approximation

András A. Benczúr, Hungarian Academy of Sciences | Searching the Web with Low Space Approximations

SimRank Example



$$\sum_{k>0} \frac{1}{3^k} \sum_u \mathrm{RP}^{[k]}_{v_1}(u)\mathrm{RP}^{[k]}_{v_2}(u) = \frac{1}{4} \cdot \frac{1}{3}\left(1 + \frac{1}{3} + \frac{1}{3^2} + \dots\right) = \frac{1}{12} \cdot \frac{3}{2}$$

$$\mathrm{SSim}^{(0)}(u_i) = \frac{1}{3} + \frac{1}{3^2} + \dots = \frac{1}{2} \qquad \mathrm{SSim}^{(1)}(u_i) = \frac{1}{4}$$

$$\mathrm{SSim}(u_i) = 1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \dots = \frac{2}{3} \checkmark$$

Singular Value Decomposition

- Fundamental tool in data mining (e.g. clustering) and web IR (e.g. HITS, LSI)
- Task: Given $A \in \mathbb{R}^{m \times n}$ find rank-k matrix A_k such that $\|A - A_k\|_F$ is minimal, where $\|X\|_F^2 = \sum_{ij} x_{ij}^2$
- Solution: Singular Value Decomposition, slow as e.g. $O(min\{mn^2, nm^2\})$
- Several results based on sampling of the form
  $$\|A - \widehat{A}_k\|_F \le \|A - A_k\|_F + \epsilon^t \|A\|_F$$
- $\|A\|_F$ might be a significantly larger than $\|A - A_k\|_F$!

Fast Relative Error SVD via Random Projections

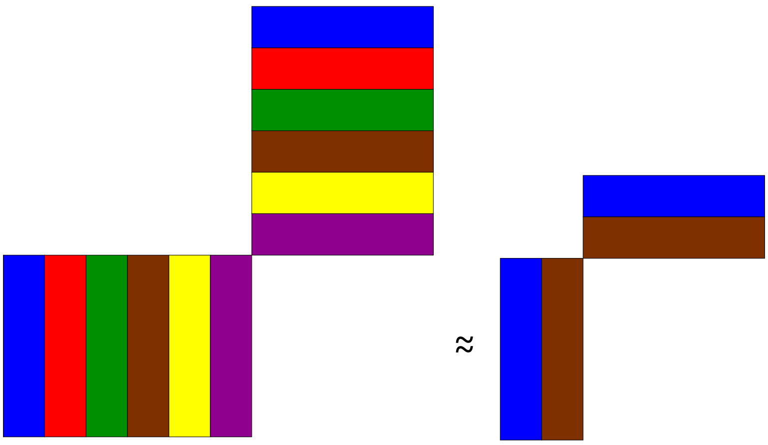
- Three recent independent results [HP06, DV06, Sar06] on $\|A - \widehat{A}_k\|_F \leq (1 + \epsilon)\,\|A - A_k\|_F$
- [DV06, Sar06] both project input to r-dim subspace, and run SVD on projection. Total time $O(Mr + (n+m)r^2)$ with M non-zeroes

	Ours	[DV06]
Fewer passes:	2	$O(k \log k)$
Faster in k:	$r = \frac{k}{\epsilon} + k \log k$	$r = \frac{k}{\epsilon} + k^2 \log k$
Subspace:	random linear combination of rows	non-uniform random sample of rows

- Heavily builds on [AMS99, Ach03, DMM06b, DMM06a, DRVW06]

Fully Personalized PageRank
Similarity Search
Relative Error Low-rank Matrix Approximation

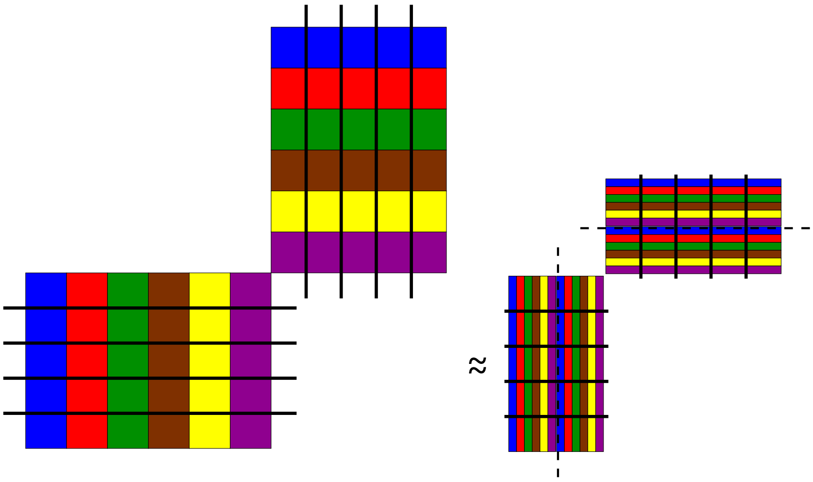
The Core Idea – Approximate Matrix Products



- Reduce $C = A \cdot B$ to smaller $\widehat{A} \cdot \widehat{B}$
- C = sum of dyads, each product of the ith column of A and and the ith row of B $\Longrightarrow$
  - Sample a few (large) dyads to reduce the number of terms in the sum
  - Sampling probabilities need to depend on the data

András A. Benczúr, Hungarian Academy of Sciences
Searching the Web with Low Space Approximations

The Core Idea – Approximate Matrix Products Cont'd



- C_{ij} = dot product of the ith row of A and the jth column of $B \Longrightarrow$
  - Use low-distortion embeddings and compute the dot products of shorter vectors
  - Embeddings are data independent

Conclusion

- Space-optimal summaries for fully personalized PageRank, and for SimRank with decay factor $< 1/2$
- Fast $O(1)$-pass relative error SVD algorithm
- At the heart of it: low space approximation of large vectors in the $\|\ldots\|_\infty$ and $\|\ldots\|_2$ norms

References

András A. Benczúr, Károly Csalogány, Tamás Sarlós, and Máté Uher.
SpamRank – Fully automatic link spam detection.
In *Proc. of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005. To appear in *Information Retrieval*.

Dániel Fogaras and Balázs Rácz.
Scaling link-based similarity search.
In *Proc. of the 14th World Wide Web Conference (WWW)*, pp. 641–650, 2005.

Dániel Fogaras, Balázs Rácz, Károly Csalogány, and Tamás Sarlós.
Towards Scaling Fully Personalized PageRank: Algorithms, Lower Bounds, and Experiments.
Internet Mathematics., 2(3):333-358, 2005. Preliminary version appeared in WAW04.

References Cont'd

Tamás Sarlós, András A. Benczúr, Károly Csalogány, Dániel Fogaras, and Balázs Rácz.
To randomize or not to randomize: Space optimal summaries for hyperlink analysis.
In *Proc. of the 15th International World Wide Web Conference (WWW)*, 2006. Full version available at http://www.ilab.sztaki.hu/websearch/Publications/.

Tamás Sarlós.
Improved approximation algorithms for large matrices via random projections. Manuscript, 2006.

Adam L. Buchsbaum and Raffaele Giancarlo and Balázs Rácz.
New Results for Finding Common Neighborhoods in Data Streams. Submitted to *SIAM Journal on Computing*, 2005.

Thank you!

- http://www.ilab.sztaki.hu/websearch
- Your questions?

Further References

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan.
Polynomial time algorithm for column-row-based relative-error low-rank matrix approximation.
Technical Report 2006-04, DIMACS, 2006.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan.
Sampling algorithms for ℓ_2 regression and applications.
In *Proc. of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1127–1136, 2006.

Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang.
Matrix approximation and projective clustering via volume sampling.
In *Proc. of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.

Further References 2

Amit Deshpande and Santosh Vempala.
Adaptive sampling and fast low-rank matrix approximation, 2006. eccc.hpi-web.de/eccc-reports/2006/TR06-042/.

Sariel Har-Peled.
Low rank matrix approximation in linear time, 2006.
valis.cs.uiuc.edu/~sariel/research/papers/05/lrank/.

Glen Jeh and Jennifer Widom.
SimRank: A measure of structural-context similarity.
In *Proc. of the 8th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 538–543, 2002.

Glen Jeh and Jennifer Widom.
Scaling personalized web search.
In *Proc. of the 12th World Wide Web Conference (WWW)*, pp. 271–279, 2003.

Further References 3

Dimitris Achlioptas.
Database-friendly random projections: Johnson-Lindenstrauss with binary coins.
Journal of Computer and System Sciences, 66(4):671–687, 2003.

Noga Alon, Yossi Matias, and Mario Szegedy.
The space complexity of approximating the frequency moments.
Journal of Computer and System Sciences, 58(1):137–147, 1999.

Graham Cormode and S. Muthukrishnan.
An improved data stream summary: The Count-Min sketch and its applications.
Journal of Algorithms, 55(1):58–75, 2005.