

Rank Synopses for Efficient Time Travel on the Web Graph

Klaus Berberich, Srikanta Bedathur, Gerhard Weikum
 Max-Planck Institute for Informatics, Saarbrücken, Germany
 {kberberi, bedathur, weikum}@mpi-inf.mpg.de

Problem

Time-Travel Queries in Web-Archive Search require Retrospective Ranking

Examples:

“Senate Elections Projections” as of 11/02/2006
 “Enron Irregularities” as of 11/15/2001
 “Olympic Games” as of Summer 2004

Providing effective results for such queries requires retrieving historical PageRank scores for the **temporal context** of interest.

Goals:

- **Interpolation**

i.e., capability to estimate a Web page’s PageRank score for any time t in its lifespan

- **Accuracy**

i.e., while tolerating errors in terms of absolute PageRank score, the ranking of Web pages be accurately reconstructed

- **Space Efficiency**

i.e., low total amount of required storage

Solution

Step 1: PageRank Normalization

We normalize PageRank scores computed on $G_t(V_t, E_t)$ (i.e., the graph at time t) dividing by the lower bound PageRank score that would be assigned to a node without incoming edges

$$r_{low} = \frac{1}{|V_t|} (\epsilon + (1 - \epsilon) \sum_{d \in D_t} r(d))$$

It can be shown that the normalized score depends **only on the node’s reachability** but **not on IV_tI or ID_tI** .

Step 2: Rank Synopses Construction

For any Web page u we consider a time series of normalized PageRank scores, computed for each graph snapshots

$$\langle (t_0, r_0), \dots, (t_m, r_m) \rangle$$

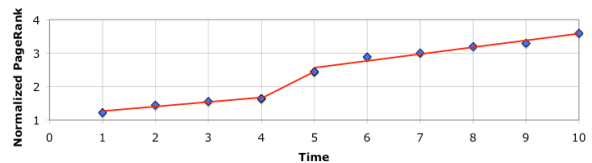
The rank synopsis of u is then a piecewise linear approximation

$$\Phi = \langle ([s_0, e_0], \Phi_0), \dots, ([s_m, e_m], \Phi_m) \rangle$$

for the quality metric (approximation error per segment)

$$error(\Phi_i) = \max_{t_i \in [s_i, e_i]} |1 - (\Phi_i(t_i)/r_i)|$$

Close-to-optimal rank synopses with error threshold θ for n observations can be computed in $O(n^2)$ time.



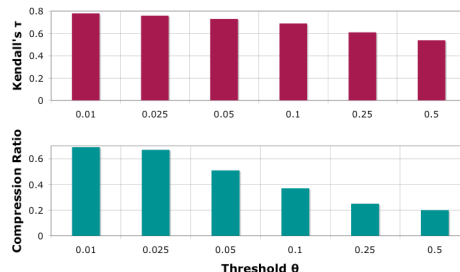
Experiments

Dataset: Wikipedia

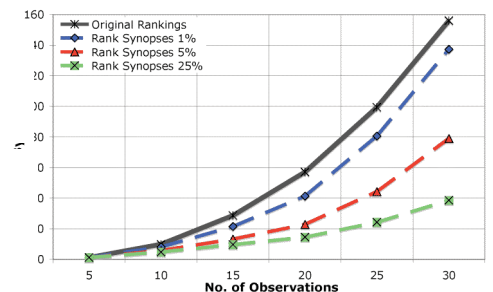
# Nodes	# Edges	Snapshot Period
1,618,650	58,845,136	Bi-monthly

Measures of Effectiveness

Accuracy	Kendall’s τ in $[-1, +1]$
Space Efficiency	Compression Ratio



Accuracy and Space Efficiency for varying θ



Storage required for increasing number of observations



max planck institut
informatik