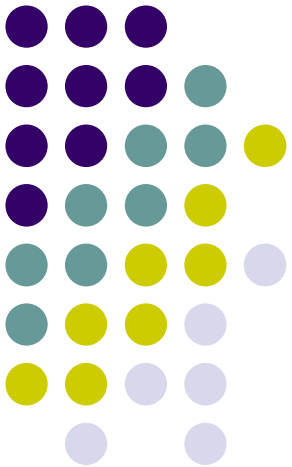
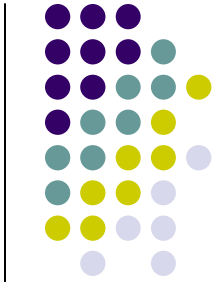
T-Rank: Time-Aware Authority Ranking

Klaus Berberich, Michalis Vazirgiannis, Gerhard Weikum
Max-Planck Institute for Computer Science (Saarbrücken)



WAW 2004, Rome (Italy), 10/16/2004

Outline



- Motivation
- Objectives
- Basics
- T-Rank: Time-aware Authority Ranking
- Experiments
- Conclusions
- Ongoing and future work

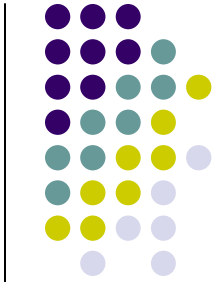
Motivation I

- Structure of the Web evolves at high pace
  - 25% new links, 8% new pages per week [Nto04]
- Page contents change frequently
  - 15% of pages weekly updated [Fet03]
- Temporal aspects of the Web's evolution
  - *How recent is a Web page or a link?*
  - *How frequently is a Web page or a link modified?*

[Nto04] A. Ntoulas, J. Cho and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective, Proceedings of the 13th conference on World Wide Web, pp. 1-12, 2004. ACM Press

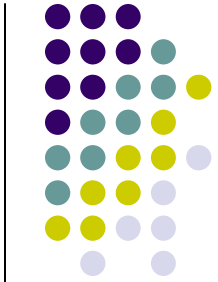
[Fet03] D. Fetterly, M. Manasse, M. Najork and J. Wiener. A large-scale study of the evolution of web pages, Proceedings of the 12th conference on World Wide Web, pp. 669-678, 2003. ACM Press

Motivation II



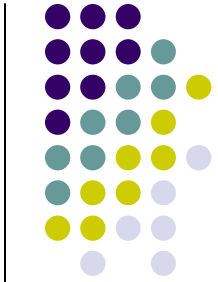
- Link-analysis techniques do not address evolution and temporal aspects
  - VLDB04 and VLDB05 websites not among top-*5* for query *"VLDB Conference"* in Sep. 04
- User's interest has a temporal dimension (e.g., most recent, last year…)

Objectives



- Integration of temporal aspects into link-based authority ranking
- Time-aware rankings that reflect
  - the user's demand for recent information
  - bring up authorities regarding a temporal interest

Basics – PageRank



- *PageRank* as a baseline

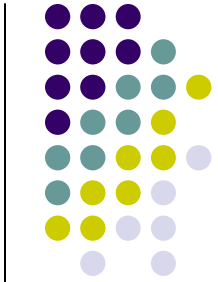
$$r(y) = \sum_{(x,y)\in E} (1-\varepsilon) \cdot \frac{r(x)}{\text{outdegree}(x)} + \frac{\varepsilon}{n}$$

- Generalization of *PageRank*

$$r(y) = \sum_{(x,y)\in E} (1-\varepsilon) \cdot t(x,y) \cdot r(x) + \varepsilon \cdot s(y)$$

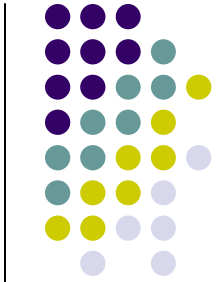
  - ***t(x,y)*** describes transition probabilities
  - ***s(y)*** describes random jump probabilities

Basics – Evolving graph



- Model of **evolving graph** *G(V,E)*
  - Nodes and edges temporally annotated
    - **TS$_{Creation}$:** creation time
    - **TS$_{Deletion}$:** deletion time
    - **TS$_{Modifications}$:** set of modification times
    - **TS$_{Lastmod}$:** last modification time
- Time represented by **integers** (e.g., days since reference date)

Basics – Temporal interest

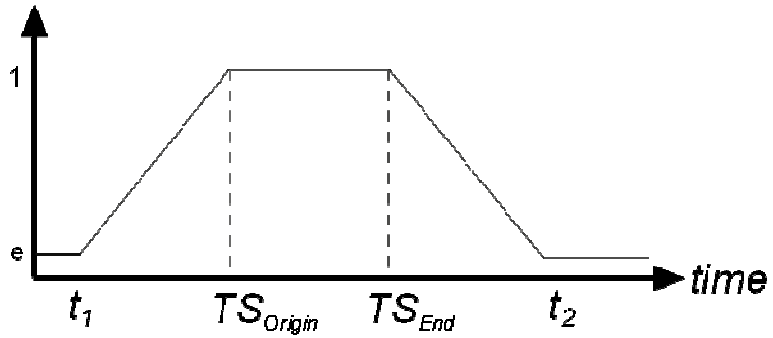


- **Temporal interest** defined by
  - **temporal window of interest** $[TS_{Origin}, TS_{End}]$
  - **tolerance interval** $[t_1,t_2] : t_1 \leq TS_{Origin} \leq TS_{End} \leq t_2$
- Graph with respect to the temporal interest $G_{ti}(V,E)$ contains nodes and edges with $TS_{Deletion} \geq t_1 \wedge TS_{Creation} \leq t_2$

Basics – Freshness

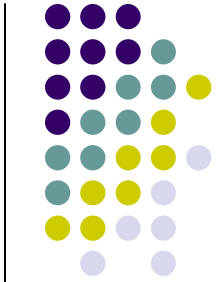
- **Freshness** measures relevance of a timestamp *ts* to a temporal interest

$$f(ts) = \begin{cases} \text{if } TS_{Origin} \le ts \le TS_{End}: & 1 \\ \text{if } t_1 \le ts < TS_{Origin}: & \dfrac{1-e}{TS_{Origin} - t_1} \cdot (ts - t_1) + e \\ \text{if } TS_{End} < ts \le t_2: & \dfrac{e-1}{t_2 - TS_{End}} \cdot (ts - TS_{End}) + 1 \\ \text{otherwise}: & e \end{cases}$$



- **Freshness of node *x*:** $f(x) = f(TS_{Lastmod}(x))$
- **Freshness of edge *(x,y)*:** $f(x,y) = f(TS_{Lastmod}(x,y))$

Basics – Activity

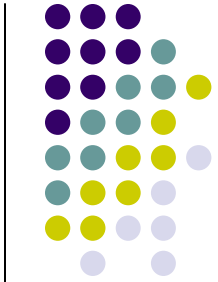


- **Activity** measures frequency of change with respect to a temporal interest

$$a(TS) = \begin{cases} if\ \mathrm{TS} \cap [t_1, t_2] \neq \varnothing: & \sum_{t1}^{t2} \{f(ts) \mid ts \in TS\} \\ otherwise: & e \end{cases}$$

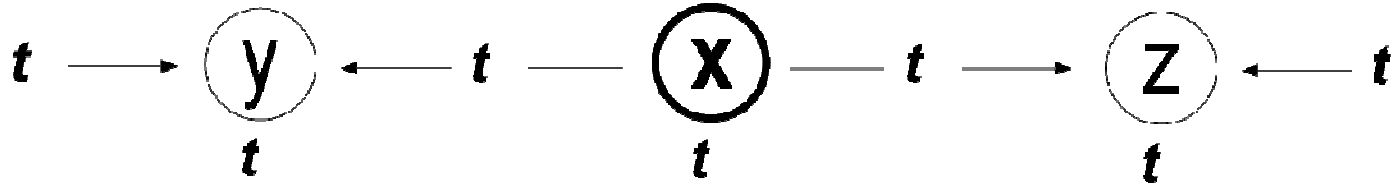
- **Activity of node *x***: $a(x) = a(TS_{Modifications}(x))$
- **Activity of edge (*x,y*)**: $a(x,y) = a(TS_{Modifications}(x,y))$

T-Rank – Overview



- **Modified *PageRank*** on $G_{ti}(V,E)$
- Transition probabilities $t(x,y)$ depend on **freshness** of nodes and edges
- Random jump probabilities depend on **freshness and activity** of nodes and edges

T-Rank – Transitions



$$t(x,y) = w_{t1} \cdot \frac{f(y)}{\sum_{(x,z)\in E} f(z)} + w_{t2} \cdot \frac{f(x,y)}{\sum_{(x,z)\in E} f(x,z)} + w_{t3} \cdot \frac{avg\{f(v,y) \mid (v,y) \in E\}}{\sum_{(x,w)\in E} avg\{f(v,w) \mid (v,w) \in E\}}$$

- Transitions favor **fresh** nodes/edges
- Coefficients w_{ti} probabilities that random surfer follows *(x,y)* with probabilities proportional to
  - freshness of node *y*
  - freshness of edge *(x,y)*
  - average (mean) freshness of incoming edges of node *y*

T-Rank – Random jumps

$$s(y) = w_{s1} \cdot \frac{f(y)}{\sum_{z \in V} f(z)} + w_{s2} \cdot \frac{a(y)}{\sum_{z \in V} a(z)} +$$

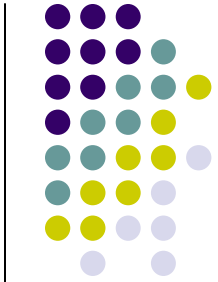
$$w_{s3} \cdot \frac{avg\{f(v,y) \mid (v,y) \in E\}}{\sum_{z \in V} avg\{f(w,z) \mid (w,z) \in E\}} + w_{s4} \cdot \frac{avg\{a(v,y) \mid (v,y) \in E\}}{\sum_{z \in V} avg\{a(w,z) \mid (w,z) \in E\}}$$

- Random jumps favor **fresh and active** nodes/edges
- Coefficients w_{si} probabilities that random surfer jumps to node y with probabilities proportional to
  - freshness and activity of node y
  - average (mean) freshness and activity of incoming edges of node y

Experiments – DBLP I

- **Digital Bibliography & Library Project** (DBLP) freely available bibliographic dataset (as XML)
- **Evolving graph** derived from DBLP
  - Authors as nodes, citations as edges
  - ~350K (~16K) nodes, ~350K edges
- ***T-Rank*** and ***PageRank*** applied for temporal interests on **decades** (70s to 00s)

Experiments – DBLP II

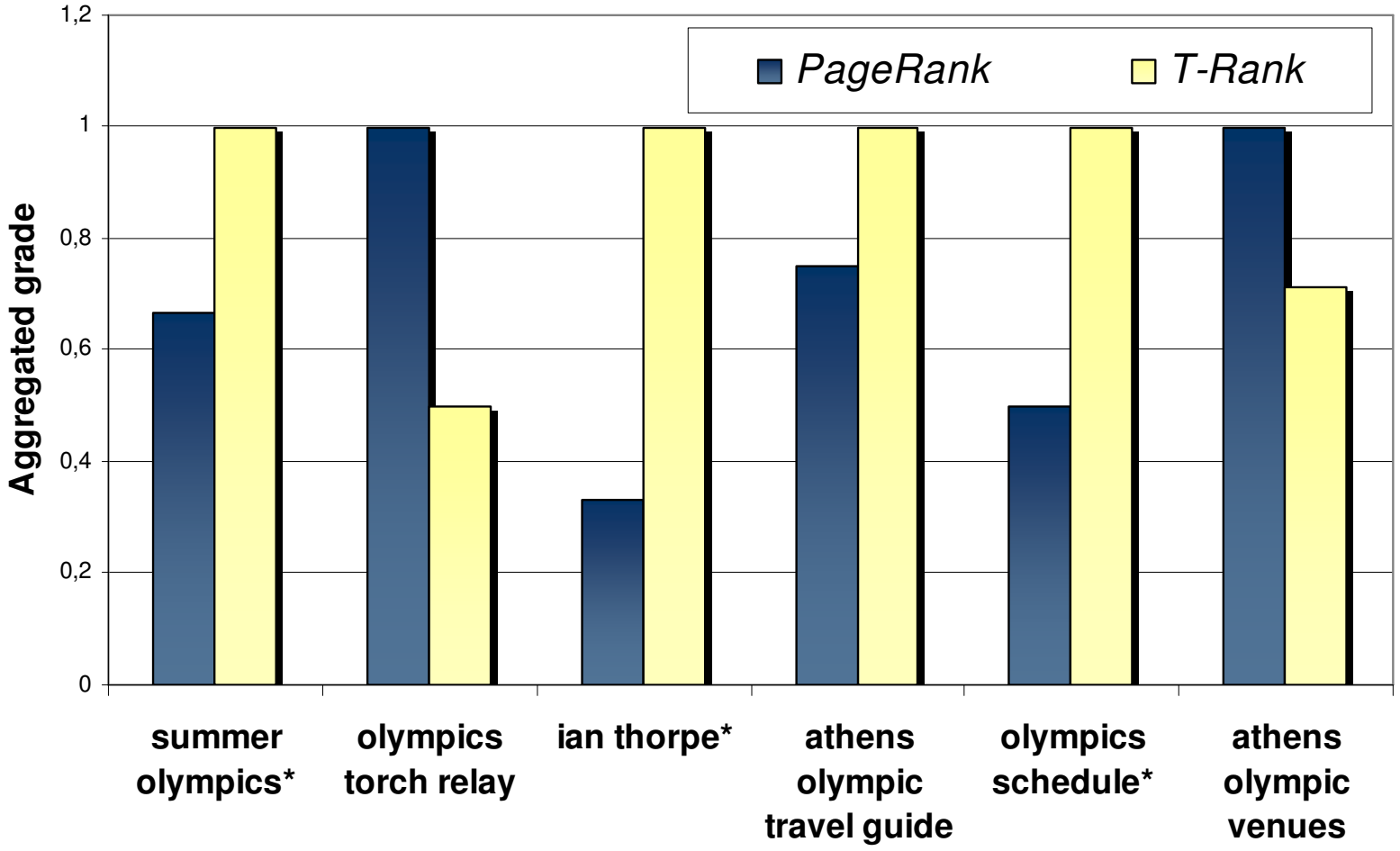


	***PageRank** 2000s*	***T-Rank** 2000s*
1	E. F. Codd	Jim Gray
2	Michael Stonebraker	Michael Stonebraker
3	Jim Gray	Jeffrey D. Ullman
4	Donald D. Chamberlin	Philip A. Bernstein
5	Jeffrey D. Ullman	Hector Garcia-Molina
6	Philip A. Bernstein	Jeffrey F. Naughton
7	Raymond A. Lorie	Donald D. Chamberlin
8	Morton M. Astrahan	David J. DeWitt
9	Kapali P. Eswaran	Jennifer Widom
10	John Miles Smith	Rakesh Agrawal

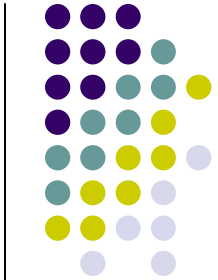
Experiments – Web I

- **Olympic Games 2004**
  - ~200K thematically related Web pages
  - 9 crawls in period July 26th to September 1st
- **Blind test** comparing *PageRank* and *T-Rank*
  - Users asked to **grade quality** of given top-*10* lists
  - Half of the queries drawn from Google Zeitgeist

Experiments – Web II



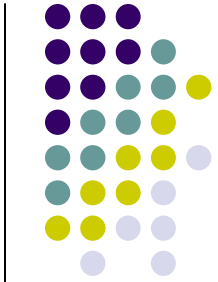
Conclusions



- Integration of temporal aspects into link-based authority ranking produces time-aware rankings
- Time-aware authority rankings bring up authorities with respect to a temporal interest
- Experimental results show that users prefer time-aware authority rankings for most queries

Ongoing and future work

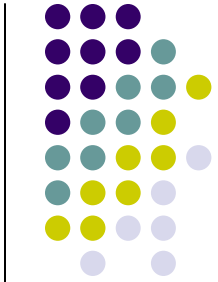
- Lightweight version of *T-Rank*
- Trend-based Authority Ranking techniques
  - *Which pages had the largest relative gains in authority?*
- Automatic parameter tuning based on properties of the dataset and user input
- Online computation of time-aware rankings



Thank you for your attention!

Questions and feedback are welcome!

Prototype implementation



- **Java** implementation (J2SE 1.4.3)
- **Oracle 9i** used for storage of data
- **Application-independent**, since based on database relations capturing **evolving graph**
- **Bingo!** focused crawler collects Web data
- **Power method** (multi-threaded)
- **Compressed Row Storage**