Time-aware Authority Ranking

Klaus Berberich[1] Michalis Vazirgiannis[1,2] Gerhard Weikum[1]

[1]Max Planck Institute for Informatics
{kberberi, mvazirgi, weikum}@mpi-inf.mpg.de

[2]Athens University of Economics & Business
{mvazirg}@aueb.gr

DELIS CCT3 Meeting, Dagstuhl, September 3-4 2005

Motivation

- Web is a very large evolving network
  - 25% new links and 8% new pages per week [NCO04]
  - 15% of pages weekly updated [FMNW03]
- Evolution and associated temporal features like recency and frequency of change ignored in Web search
- For many queries user's interest has a temporal dimension
  - *"new orleans hurricane"* ($\sim$ most recent information)
  - *"www conference"* ($\sim$ upcoming conference)
  - *"tour de france"* ($\sim$ this year's event)

Motivation (cont'd)

Our approach:

- Exploit temporal dimension of user's interest (temporal interest)
- Integrate temporal aspects in link-based authority ranking (in our case PageRank)
- Produce rankings that bring up authorities w.r.t. given temporal interest

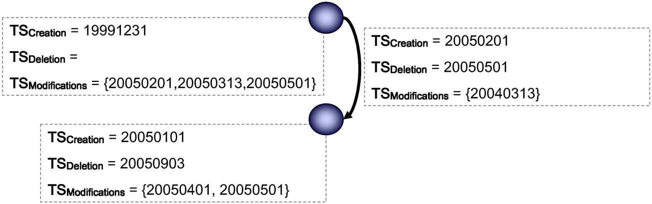
Evolving Graph

- Integers as time representation, s.t. timestamp t gives # of time units that have passed since reference time
- Directed graph $G(V, E)$ with temporal annotations

  $TS_{Creation}$: creation time

  $TS_{Deletion}$: deletion time

  $TS_{Modifications}$: modification times



Temporal Interest

- Models the temporal dimension of user's interest
  - Temporal window of interest $[TS_{Begin}, TS_{End}]$
  - Tolerance interval $[t_1, t_2] \supseteq [TS_{Begin}, TS_{End}]$
- Graph $G_{ti}(V_{ti}, E_{ti})$ with respect to temporal interest

$$
\begin{aligned}
V_{ti} &= \{v \in V \mid TS_{Creation}(v) \leq t_2 \ \wedge \ TS_{Deletion}(v) \geq t_1\} \\
E_{ti} &= \{(x,y) \in E \mid (x,y) \in V_{ti} \times V_{ti} \ \wedge \\
&\quad TS_{Creation}(x,y) \leq t_2 \ \wedge \ TS_{Deletion}(x,y) \geq t_1\}
\end{aligned}
$$

contains nodes and edges that exist at any point of $[t_1, t_2]$

PageRank

- Standard PageRank [PBMW98]

  $$r(y) = (1-\epsilon)\left(\sum_{(x,y)\in E} \frac{r(x)}{outdegree(x)}\right) + \frac{\epsilon}{n}$$

  ϵ probability of making a random jump
- Generalization of PageRank

  $$r(y) = (1-\epsilon)\left(\sum_{(x,y)\in E} t(x,y)\cdot r(x)\right) + \epsilon \cdot s(y)\ ,\ \text{with}$$

  $t(x,y)$ transition probability from node x to node y, and

  $s(y)$ random jump probability of node y
- Definition of t and s such that

  $\sum_y s(y) = 1$ and $\sum_y t(x,y) = 1$

T-Rank Light & T-Rank

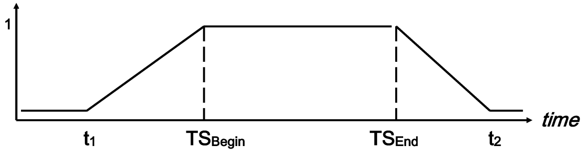
- **Objective:** Ranking that reflects given temporal interest
- Freshness and activity as temporal features of nodes and edges in the graph G_{ti}
- Random jump probabilities s and transition probabilities t defined based on the temporal features

  T-Rank Light retains uniform transition probabilities but relies on time-aware random jump probabilities

  T-Rank relies on both time-aware transition probabilities and time-aware random jump probabilities

Freshness

- Freshness $f(ts) \in]0,1]$ of timestamp ts conveys recency w.r.t. given temporal interest



- Freshness $f(o) \in]0,1]$ of a node or edge o defined as $f(o) = Max\{f(ts) \mid ts \in TS_{Modifications}(o) \cup \{TS_{Creation}(o)\}\}$

Activity

- Activity $a(T) \in]0, \infty]$ of a set of timestamps T conveys frequency of change w.r.t. given temporal interest

$$\sum_{ts \in T} f(ts)$$

- Activity $a(o) \in]0, \infty]$ of a node or edge o defined as

$$a(\,(TS_{Modifications}(o) \;\cap\; [t_1, t_2]) \;\cup\; \{TS_{Creation(o)}\}\,)$$

Random Jump Probabilities s

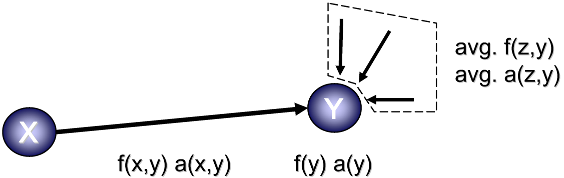
- Random jump probability $s(y)$ depends on
  - freshness and activity of node y
  - avg. freshness and activity of edges pointing to node y

$$
\begin{aligned}
s(y) &= w_{s1} \cdot \frac{f(y)}{\sum_{z \in V_{ti}} f(z)} \\
&+ w_{s2} \cdot \frac{avg\{f(x,y) \mid (x,y) \in E_{ti}\}}{\sum_{z \in V_{ti}} avg\{f(x,z) \mid (x,z) \in E_{ti}\}} \\
&+ w_{s3} \cdot \frac{a(y)}{\sum_{z \in V_{ti}} a(z)} \\
&+ w_{s4} \cdot \frac{avg\{a(x,y) \mid (x,y) \in E_{ti}\}}{\sum_{z \in V_{ti}} avg\{a(x,z) \mid (x,z) \in E_{ti}\}}
\end{aligned}
$$

- Employed in **T-Rank Light** and **T-Rank**

Transition Probabilities t

- Transition probability $t(x, y)$ depends on
  - freshness and activity of node y
  - freshness and activity of edge $x \rightarrow y$
  - avg. freshness and activity of edges pointing to node y



- Employed in **T-Rank**

Experiments

- Three different datasets
  - Digital Bibliography & Library Project (DBLP)
  - Amazon.com
  - Olympic Games 2004
- Experiments focus on
  - quality of ranking (assessed through user studies)
  - parameter sensitivity
- Here we concentrate on
  - “anecdotic evidence” by means of top-10 lists (DBLP)
  - result of user study (Amazon.com)

Digital Bibliography & Library Project

- XML snapshot freely available
- Evolving graph with

  348K (16K) nodes $\sim$ authors

  347K edges $\sim$ citations
- Top-10 lists as obtained for temporal interest on 2000s

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

Amazon.com

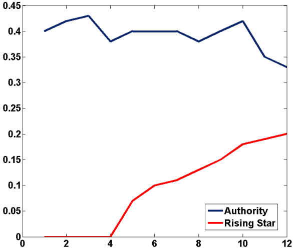
- **Amazon.com** product data obtained through web service
- Evolving graph with

  200K nodes $\sim$ products

  586K edges $\sim$ similar product recommendations
- User study with 5 test persons on 30 queries like *"da vinci code", "che guevara", "mac os x programming",...*
- Top-10 lists for temporal interest 10/1/2004 $\rightarrow$ 3/31/2005
- **T-Rank Light** and **T-Rank** both outperform PageRank on more than 66% of the queries

Ongoing Work: Trend-based Authority Ranking

- Identify trends and exploit them for ranking!



Consider query *"new orleans"*

- Authority could be page about city's history
- Rising Star could be page relating to the recent events

- Related to PageRank/Search Engine bias problem [CR04]
- **First lesson learned:** Hard to judge quality of results

Open Issues

- **Datasets**

  Web data optimal but hard to obtain

  DBLP and Amazon.com, e.g., have their particularities

  Further datasets?

- **Experiments**

  User studies the best one can do?

  What kinds of experiment do you find convincing?

Conclusions

- **T-Rank Light** and **T-Rank** produce time-aware authority rankings that reflect given temporal interest
- Experiments show superiority over non-time-aware methods across different datasets
- More details may be found in

Klaus Berberich, Michalis Vazirgiannis, and Gerhard Weikum.
T-Rank: Time-aware Authority Ranking.
In Stefano Leonardi, editor, *Algorithms and Models for the Web-Graph: Third International Workshop, WAW 2004*, pages 131–141. Springer-Verlag, 2004.

Klaus Berberich, Michalis Vazirgiannis, and Gerhard Weikum.
Time-aware Authority Ranking.
Internet Mathematics, 2005 (accepted for publication).

Thanks for your attention!

Bibliography

Junghoo Cho and Sourashis Roy.
Impact of Search Engines on Page Popularity.
In *Proceedings of the 13th conference on World Wide Web*, pages 20–29. ACM Press, 2004.

Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener.
A Large-Scale Study of the Evolution of Web Pages.
In *Proceedings of the twelfth international conference on World Wide Web*, pages 669–678. ACM Press, 2003.

Alexandros Ntoulas, Junghoo Cho, and Christopher Olston.
What's New on the Web?: The Evolution of the Web from a Search Engine Perspective.
In *Proceedings of the 13th conference on World Wide Web*, pages 1–12. ACM Press, 2004.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd.
The PageRank Citation Ranking: Bringing Order to the Web.
Technical report, Stanford Digital Library Technologies Project, 1998.