Link-based Authority Ranking in Evolving Networks

Klaus Berberich
kberberi@mpi-inf.mpg.de

Introduction

Link-based authority ranking techniques are employed to assess authority in a network of entities (e.g., web pages). In the context of web search these authority assessments have proven valuable when producing ranked result lists. However, existing approaches are based on a static snapshot of the network (e.g., its current state), thus neglecting the network's history and dynamics.

Our Baseline: PageRank [1]

- Graph **G(V,E)** with nodes **V** and edges **E**
- PageRank score **r(y)** of a node **y** defined as

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

- **Intuition**: Random surfer traverses graph either following outgoing edges or jumping to arbitrary nodes. The latter happens with probability ϵ. The score of a node is then proportional to the fraction of time the random surfer spends on the node.
- **Math**: Scores correspond to stationary state probabilities of a Markov chain that are determined by solving an eigenvector problem for a very large matrix.

„Networks highly dynamic"

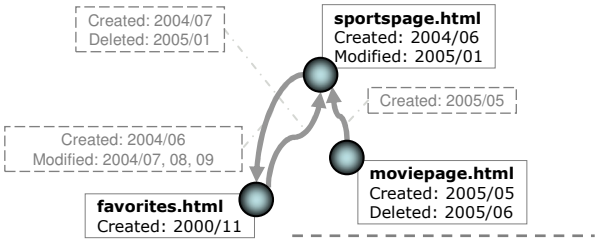

Figure I

Examples of Evolving Networks

- **Web** evolves rapidly: pages and hyperlinks are created and deleted; page contents and anchortexts are modified. (see Figure 1)
- **Bibliographic networks** grow at a high pace: new authors and publications appear, interconnecting citations are added.

Time-aware Authority Ranking [2]

- **Objective**: Assess authority with regard to a user-defined time-window (e.g., year 2004).
- **Our approach *T-Rank*:** Bias PageRank's random walk based on temporal aspects.
- Generalization of PageRank as a baseline

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

- **t** and **s** depend on *freshness* and *activity* that capture the timeliness and the frequency of change of nodes and edges.
- **Results**: User studies show improvements in the quality of rankings on different datasets (web, bibliographic, product data). Table 1 depicts Top-5 lists for the 2000s obtained on bibliographic data (DBLP).

	PageRank	T-Rank
1.	E. F. Codd	Jim Gray
2.	Michael Stonebraker	Michael Stonebraker
3.	Jim Gray	Jeffrey D. Ullman
4.	Jeffrey D. Ullman	Philip A. Bernstein
5.	Philip A. Bernstein	Hector Garcia-Molina

Table I

Trend-based Authority Ranking

- **Motivation:** PageRank favors old nodes. Consequently, it is hard for new nodes to obtain high authority scores.
- **Objective:** Overcome the bias against new nodes while preserving or improving the quality of ranking.
- **Our approach:** Assess the relative change of a node's authority value with regard to a time-window. (This is ongoing work)

References & Publications

[1] Brin S. and Page L.:
The anatomy of a large-scale hypertextual Web search engine. WWW 1998

[2] Berberich K., Vazirgiannis M., Weikum G.:
T-Rank: Time-aware Authority Ranking.
WAW 2004

MAX-PLANCK-GESELLSCHAFT

max planck institut informatik