BuzzRank...

Klaus Berberich, Srikanta Bedathur, Gerhard Weikum
Max-Planck-Institut für Informatik, Saarbrücken, Germany
{kberberi, bedathur, weikum}@mpi-inf.mpg.de

Michalis Vazirgiannis
Athens University of Economics and Business, Athens, Greece
mvazirg@aueb.gr

Motivation

Link-based ranking methods like PageRank fail for information needs like the following, since the history of the (e.g., web) graph overshadows its recent evolution.

1. *Highest rated Movies to watch (in cinema)*
   What we expect: *Mission Impossible III*
   What we get: *The Godfather*

2. ***Top-importance-gaining publications in database research***

Idea

Importance (e.g., PageRank) scores of a node v co-evolve with the underlying graph thus forming a **time series** with importance scores r_i at observation times t_i

$$\langle (t_0, r_0), \ldots, (t_n, r_n) \rangle$$

BuzzRank analyzes such *time series of importance scores* and **quantifies contained trends** based on a *growth model* of importance scores.

Thus, BuzzRank identifies items that increased their importance significantly in a period of interest, or, caused significant ***BUZZ***.

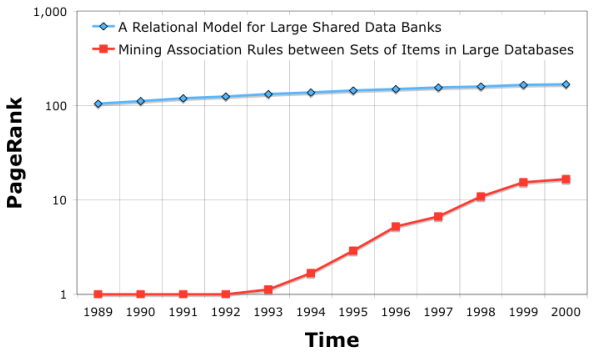
"...and the Trend is Your Friend"

Use Case -
Identify Database Research Papers of Interest

Assume a user wants to identify recent publications in database research that are *becoming* important.

PageRank
The top-ranked publication upto '00 – ***"A Relational Model of Data for Large Shared Data Banks"*** by E. F. Codd (published 1970).

But...,
Agrawal et al.'s ***"Mining Association Rules between Sets of Items in Large Databases"*** would be a better result if the given information need arises at any time between '93 and '00.



As can be seen from the above figure, the association rules paper gained relatively more importance at any point since '93.

BuzzRank identifies this and ranks the paper by Agrawal et al., ahead of its opponent.

BuzzRank

BuzzRank quantifies the ***buzz*** created within a time-interval of interest $[t_{begin}, t_{end}]$.

Method

The growth of importance scores is modeled by the following generic growth model with $\alpha_v(t)$ being the rate of importance growth of node v at time t.

$$\hat{r}(v,t) = e^{\int_0^t \alpha_v(t)dt}$$

We assume that the growth rate is time-invariant within the time-interval $[t_{begin}, t_{end}]$ giving us the following simplified model with parameters α_v and $A_{v,t_{begin}}$

$$\hat{r}(v,t) = A_{v,t_{begin}}\, e^{\alpha_v\,(t-t_{begin})}$$
$$: \quad t_{begin} \leq t \leq t_{end}$$

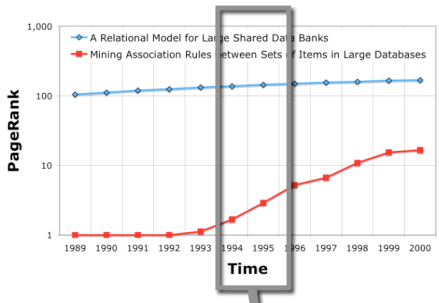
Optimal parameter values α^*_v and $A^*_{v,t_{begin}}$ are estimated using the method of least squares, i.e., minimizing

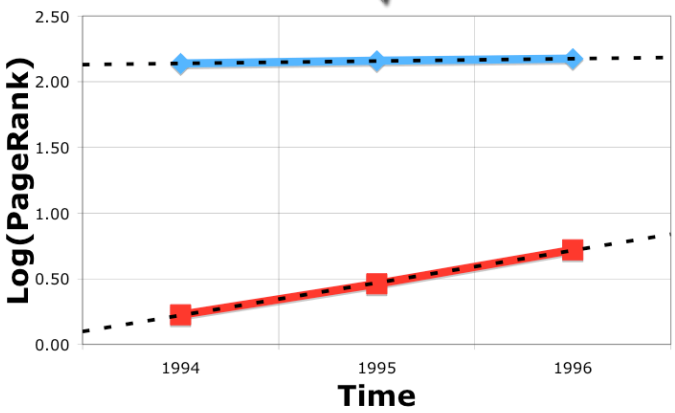
$$\sum_{t_{begin} \leq t_i \leq t_{end}} \left(r(v,t_i) - \hat{r}(v,t_i)\right)^2$$

Applying a log-transformation to the observed importance scores, this is equivalent to fitting a straight line, so that closed-form solutions exist.

Finally, BuzzRank ranks nodes based on the parameter α^*_v which is an estimate of the growth rate of a node's importance in the time-interval $[t_{begin}, t_{end}]$.

$[t_{begin}, t_{end}] = [\text{'94, 95'}]$





$$(a^* = 1.04,\ A^*_{t_{begin}} = 131.86)$$
$$(a^* = 1.77,\ A^*_{t_{begin}} = 1.06)$$

	Top-2@['94,'95]
1.	Mining Association Rules between Sets of Items in Large Databases
2.	A Relational Model of Data for Large Shared Data Banks

PageRank Normalization

We use PageRank to assess importance as an input to BuzzRank. Plain PageRank scores, however, are not comparable across graphs.

Therefore, 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

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

It can be shown that the normalized score depends only on the node's reachability but not on the graph size or the number of dangling nodes D_t.

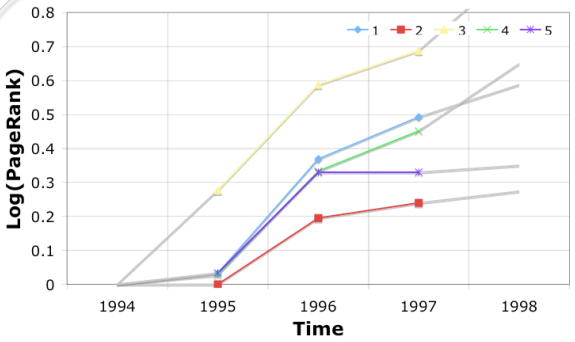
Experiments

Experiments were conducted on bibliographic **DBLP** dataset.

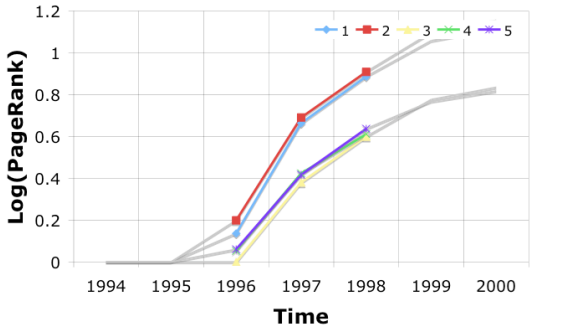
Input: PageRank rankings computed for years '89–'99
Output: BuzzRank rankings for two year time intervals ***[t, t+1]***

Top Buzzing Publications

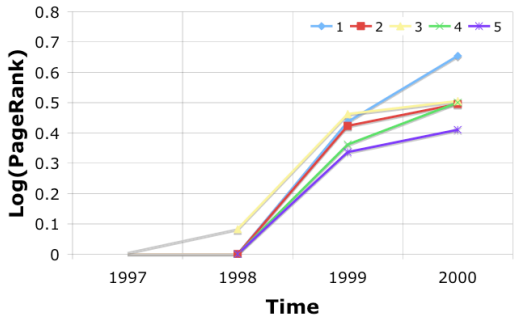
Years	Title
['89,'90]	The Object-Oriented Database System Manifesto
['90,'91]	CYC: Toward Programs With Common Sense
['91,'92]	ARIES: A Transaction Recovery Method Supporting Fine-Granularity…
['92,'93]	Simplifying Decision Trees
['93,'94]	World-Wide Web: The Information Universe
['94,'95]	The Power of Languages for the Manipulation of Complex Values
['95,'96]	**Towards Heterogeneous Multimedia Information Systems: The Garlic Approach**
['96,'97]	**Implementing Data Cubes Efficiently**
['97,'98]	Modeling Multidimensional Databases
['98,'99]	**XML-QL: A Query Language for XML**



	Top-5@['95,'96]
1.	Towards Heterogeneous Multimedia Information Systems…
2.	Database Research: Achievements and Opportunities…
3.	Fast Algorithms for Mining **Association Rules** in Large…
4.	Discovery of Multiple-Level **Association Rules** from…
5.	A High Performance Configurable Storage Manager



	Top-5@['96,'97]
1.	Implementing **Data Cubes** Efficiently
2.	**Data Cube**: A Relational Aggregation Operator…
3.	Index Selection for **OLAP**
4.	W3QS: A Query System for the World-Wide Web
5.	On the Computation of Multidimensional Aggregates



	Top-5@['98,'99]
1.	XML-QL: A Query Language for **XML**
2.	Join Synopses for Approximate Query Answering
3.	Selectivity Estimation Without the Attribute Value…
4.	Hash Joins and Hash Teams in Microsoft SQL Server
5.	Relational Databases for Querying **XML** Documents…

Conclusions and Future Work

BuzzRank…

- identifies *"hot"* authoritative items in given time period
- is based on time series of precomputed PageRank scores
- is complementary to PageRank

In the Future...

- Experiments on a variety of datasets: Wikipedia, Web graph,…
- Extension of the underlying model (e.g., time-varying growth rate)
- Improvement of scalability

max planck institut informatik