

Web Dynamics

Part 5 – Searching the Past

5.1 Time-travel problems

5.2 Efficient Time-Travel Search

5.3 Temporal measures of page importance

Time Travel Problems on the Web

Search engines index only the *current* Web

But: Many interesting aspects on the *historical* Web:

5.2 → Search the Web as of a specific time in the past
(„opinions of major US politicians on the Iraq War in 2002“)

5.3 → Analyze the Web as of a specific time in the past
(„most authoritative news page in 2002“)

- Analyze temporal development of the Web
(„since when have political blogs been around?“)

Web Archives don't provide these functionalities
(at least not publicly)

Rare example: Google@2001

The screenshot shows the Google search engine interface from 2001. The search bar contains the text "paris hilton". The search results are displayed in a list format. The first result is "Hilton Paris - France" with a description of 3 prestigious hotels. The second result is "Hotel Discounts : Hotel In" with a description of listing discounts. The third result is "Travelocity : Message Bo" with a description of a message board. The fourth result is "Paris Las Vegas - A Hilt" with a description of a Hilton hotel in Las Vegas. The search results are displayed in a list format. The first result is "Hilton Paris - France" with a description of 3 prestigious hotels. The second result is "Hotel Discounts : Hotel In" with a description of listing discounts. The third result is "Travelocity : Message Bo" with a description of a message board. The fourth result is "Paris Las Vegas - A Hilt" with a description of a Hilton hotel in Las Vegas.

Google! In honor of our 10th birthday, we've brought back our oldest available index. Take a look back at Google in January 2001.

2001 Web

[Hilton Paris - France](#)
3 prestigious hotels to discover attendees or enjoy typical fren
...
<http://www.hilton-paris.com/> -

[Hotel Discounts : Hotel In](#)
Listing of Discounts of Hotels:
<http://www.vacationweb.com/>

[Travelocity : Message Bo](#)
#2 of 3: **Paris Hilton**. katherin
November of 1997. It is right b
<http://dest.travelocity.com/Con>

[Paris Las Vegas - A Hilt](#)
Paris Las Vegas - A Hilton C
Las Vegas.
<http://www.lasvegasstriphotels>


Web Images Maps News Shopping Gmail more ▾

Google [Advanced Search](#)
[Preferences](#)

Web News Images Video Blogs Result

Related searches: [paris hilton movie](#) [1 night in paris](#) [paris exposed](#)

News results for [paris hilton](#)

 [Windsor Star](#) **Paris Hilton** debuts new single, "My BFF" - 8 hours ago
By Jill Serjeant LOS ANGELES (Reuters) - Leave it to **Paris Hilton** to ratchet up the hype for her new TV show's debut, releasing a new single, ...
[Reuters](#) - 30 related articles »

Paris Hilton finds new BFF Ellen - [Oneindia](#) - 9 related articles »
Paris Hilton porno art goes on display - [Stuff.co.nz](#) - 4 related articles »

[Paris Hilton](#) - Wikipedia, the free encyclopedia
[19] In March 2008, it was reported that **Hilton** would star in a new MTV reality series tentatively titled **Paris Hilton's My New BFF**, about her looking for a ...
en.wikipedia.org/wiki/Paris_Hilton - 141k - [Cached](#) - [Similar pages](#)

[Paris Hilton](#) | The Official Website
ParisHilton.com **Paris Hilton**, Nicky **Hilton** Fashion, Pictures, Apparel, Jewellery , Film, and Fun.
www.parishilton.com/ - 8k - [Cached](#) - [Similar pages](#)

[Paris Hilton](#) Zone | [Paris Hilton](#) Pictures, Pics, Photos
4000+ new **Paris Hilton** pictures, **Paris** wallpaper, sex tape, lyrics, audio, video , daily **Paris** pics & news.
www.parishiltonzone.com/ - 49k - [Cached](#) - [Similar pages](#)

Web Dynamics

Part 5 – Searching the Past

5.1 Time-travel problems

5.2 Efficient Time-Travel Search

5.3 Temporal measures of page importance

(Some of the slides were contributed by Klaus Berberich)

The Need for Time-Travel Search

- Historical information needs, e.g.,
 - Contemporary (~2001) articles about the movie “Harry Potter and the Sorcerer’s Stone”
 - Search for prior art for a patent submitted 2005
 - Links to some illegal content before Feb 2009
- Relevant pages disappeared in the current Web, but preserved by Web archives (e.g., archive.org)
- Search in existing Web archives limited and ignores the time-axis

The Need for Time-Travel Search

Result on current Web

1 result from the Web arch

Internet Archive Search: (harry potter and the sorcerer's stone) AND date:[2001-11-01 TO 2001-1 - Windows

http://www.archive.org/search.php?query=%28harry%20potter%20and%20the%20sorcerer%27s%20stor

File Edit View Favorites Tools Help

☆ Favorites ☆ HRS - HOTEL RESERVATION... MMCI Wiki WebHome WisNetGrid Fos...

Internet Archive Search: (harry potter and the sorcer...

INTERNET ARCHIVE

Web Moving Images Texts Audio Software

Home Donate | Forums | FAQs | Contributions | Terms, Pri

Search: (harry potter and the sorcerer's All Media Types

Search

Search Results

Results: 1 through 1 of 1 (0.042 secs)

You searched for: (harry potter and the sorcerer's stone) AND date:[2001-11-01 TO 2

Harry Potter and the Sorcerer's Stone

This **Harry Potter** theme has wallpapers, icons, pointers, so

Keywords: [software downloads](#); [software](#); [Downloads](#)

Downloads: 45

www.boxofficemojo.com/movies/?id=harrypotter... - [Im Cache](#) - [Ahnlich](#)

Summer Term 2010

Relevant (but unfound) result

http://web.archive.org/web/20020209030855/www.nytimes.com/2001/11/16/movies/16POTT.html

Wizard School Without the ...

Movies

The New York Times

HOME JOB MARKET REAL ESTATE AUTOMOBILES

NEWS

International National Nation Challenged Politics Business Technology Science Health Sports New York Region Education Weather Obituaries NYT Front Page Corrections Special Winter Olympics

OPINION

Editorials/Op-Ed Readers' Opinions

Scotland, PA. Opens Today

FEATURES

Arts Books Movies Travel Dining & Wine Home & Garden Fashion & Style New York Today Crossword/Games Cartoons Magazine Week in Review Photos College Learning Network

SERVICES

Archive Classifieds Help Center NYT Mobile NYT Store E-Cards & More About NYTDigital Jobs at NYTDigital Online Media Kit Our Advertisers

NEWSPAPER

Home Delivery Customer Service Electronic Edition Media Kit

Search Past 30 Days

Welcome, [ia_archiver](#)

[Sign Up for Newsletters](#) [Log Out](#)

[E-Mail This Article](#) [Printer-Friendly Format](#)

[Most E-Mailed Articles](#) [Single-Page View](#)

November 16, 2001

MOVIE REVIEW / 'HARRY POTTER AND THE SORCERER'S STONE'

Wizard School Without the Magic

By ELVIS MITCHELL

THE world may not be ready yet for the film equivalent of books on tape, but this peculiar phenomenon has arrived in the form of the film adaptation of J. K. Rowling's "Harry Potter and the Sorcerer's Stone." The most highly awaited movie of the year has a dreary, literal-minded competence, following the letter of the law as laid down by the author. But it's all muted flourish, with momentary pleasures, like Gringott's, the bank staffed by trolls that looks like a Gaudí throwaway. The picture is so careful that even the tape wrapped around the bridge of Harry's glasses seems to have come out of the set design. (It never occurred to anyone to show him taping the frame together.)

The movie comes across as a covers act by an extremely competent tribute band not the real thing but an incredible simulation and there's an audience for this sort of thing. But watching "Harry Potter" is like seeing "Beatlemania" staged in the Hollywood Bowl, where the cheers and screams

ADVERTISEMENT

The New York Times

[Check Delivery Options | 50% Off-Click Here!](#)

Weekend Box Office: Ocean Robs Potter - December 11, 2001. Steven Soderbergh's Ocean's Eleven stole the top spot from Harry Potter and the Sorcerer's Stone ...

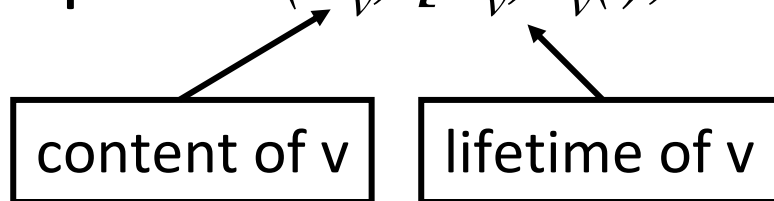
Time-Travel Search Beyond the Web

More versioned document collections:

- Wikis (like Wikipedia)
- Repositories (e.g., controlled by CVS, Subversion)
- Your Desktop

Formal Model: Document Versions

Assume continuous time dimension $T=[0...\infty($.
For each document (=url) d , maintain set of different **versions** $V(d)$, where each $v \in V(d)$ is a tuple $v=(c_v, [s_v, e_v($, with $e_v=\infty$ for current versions.



Different versions of the same document have **disjoint lifetimes** $\Rightarrow (d, s_v)$ identifies version

Archive can only estimate versions of a document

Time-Travel Keyword Queries

Time-travel keyword query $q=(k,I)$ combination of

- standard keyword query $k=(k_1, \dots, k_n)$
- time-of-interest interval $I=[s_I, e_I]$

Two important subclasses:

- **Point-in-time** queries: $s_I=e_I$
- **Interval** queries: $e_I>s_I$



Example:

“harry potter” @ 2001/11/14

This is a point-in-time query if the granularity of time is 1 day!

Scoring Point-in-Time Time-Travel Queries

Reminder: score in standard text retrieval:

$$s(d, q) \propto \sum_{k \in q} \underset{\substack{\text{frequency of } k \text{ in } d}}{tf(d, k)} \cdot \underset{\substack{\text{importance of } k}}{idf(k)}$$

$$idf(k) \propto \frac{N}{df(k)}$$

score of version $v = (c_v, [s_v, e_v])$ for $q = (\{k_1 \dots k_n\}, t)$

$$s_T(v, q) \propto \begin{cases} 0 & \text{if } t \notin [s_v, e_v] \\ \sum_{k_i} \underset{\substack{\text{frequency of } k_i \text{ in } c_v}}{tf(c_v, k_i)} \cdot \underset{\substack{\text{importance of } k_i \text{ at query time } t}}{idf(k_i, t)} & \text{if } t \in [s_v, e_v] \end{cases}$$

$$idf(k, t) \propto \frac{N(t)}{df(k, t)}$$

N : # docs; $N(t)$: #docs at time t
 $df(k)$: # docs with term k
 $df(k, t)$: # docs with term k at time t

Inverted Lists in Text IR

Reminder: Inverted Lists in text retrieval

For each term k , keep list $(d, \text{score}(d, k))$ of documents containing term k and their score, in some order

List for term k
in score order


d1,0.9
d7,0.85
d2,0.84763
d119,0.79
...

List for term k
in document order

d1,0.9
d2,0.84763
d4, 0.27
d7,0.85
...

Query processing using merge joins of these lists
(plus optional top- n for efficiency)

Extension for time-travel: SOPT

1. Split score in tf and idf component
(idf is query-dependent!)  store this somewhere else
2. For each term k , keep list $(v, tf(v, k), (s_v, e_v))$ of document versions containing term k , their tf value, and their lifetime, in some order

List for term k in score order

~~d1,90,(2001/jan/01,2001/jan/15)~~

~~d1,90,(2001/jan/16,2001/feb/28)~~

d7,85,(2004/aug/14,2004/aug/16) ✓

d1,84,(2001/mar/01,∞) ✓

...

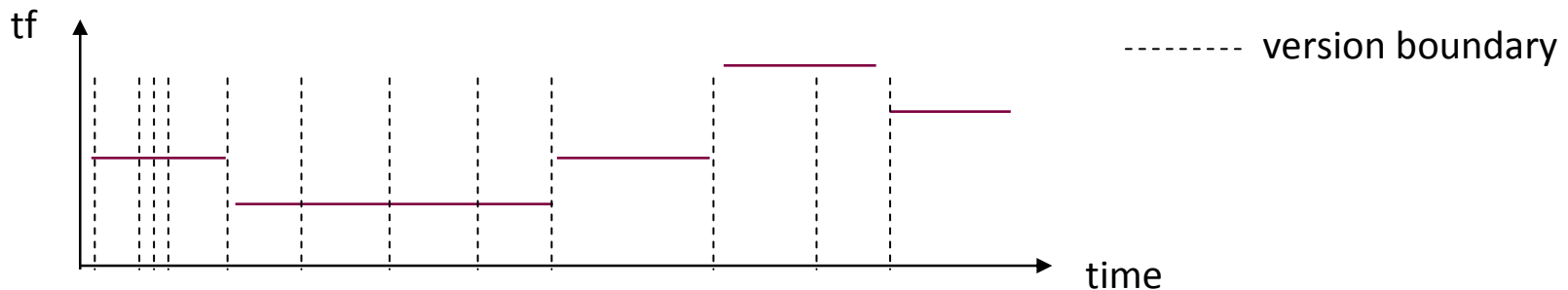
Example:
 $k@2004/aug/15$

Query processing using merge joins of these lists
plus ignoring versions where lifetime does not match query

This is not good enough

Major problems of this simple approach:

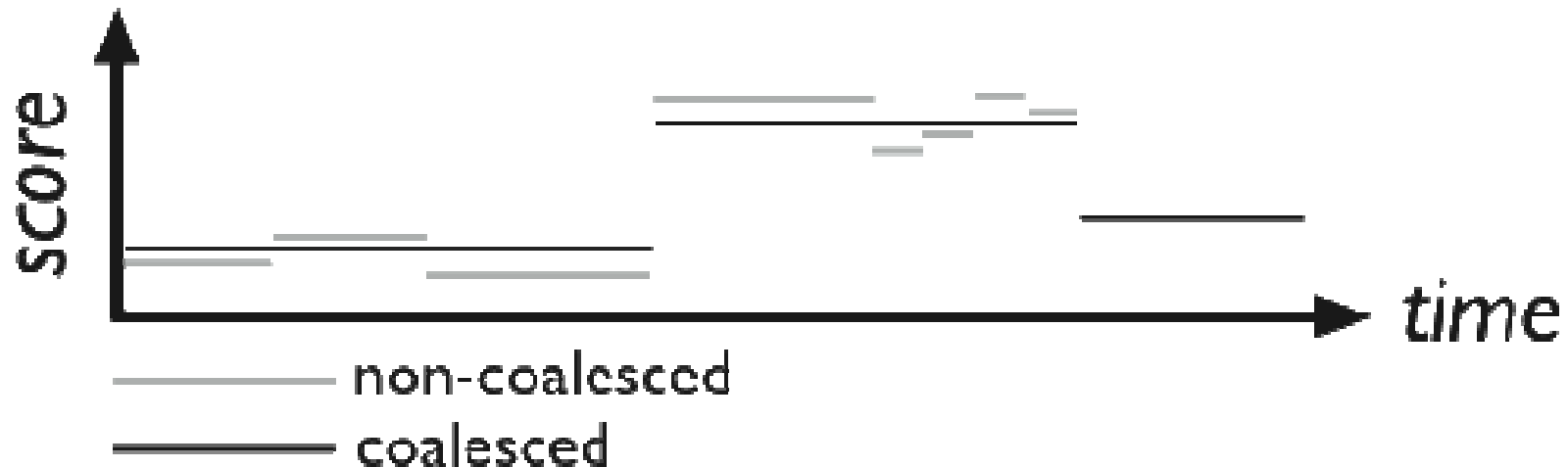
- index size explodes (*one index entry per version per term*)
⇒ for Wikipedia alone: $9 \cdot 10^9$ entries!
- Many entries
 - differ only in their lifetimes
 - have almost identical tf values (hardly matters for ranking)



Reducing Index Size: Coalescing

Idea:

Coalesce sequences of temporally adjacent postings having similar scores

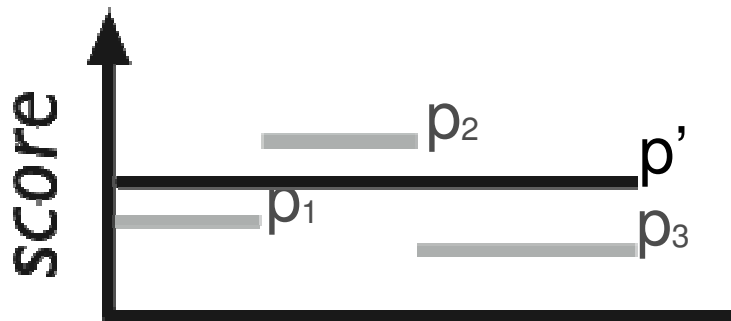


Can drastically reduce index size
But: what happens to result quality?

Formal Optimization Problem

Problem Statement:

Given input sequence I find a *minimal length output sequence* O with approximation errors bounded by a threshold ε



Guarantee:

$$|p' - p_i| / |p_i| \leq \varepsilon$$

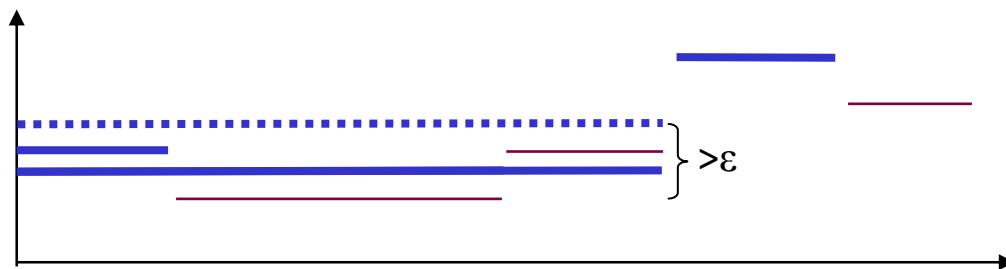
Approximate Temporal Coalescing (ATC):

finds an optimal output sequence using a greedy linear time algorithm

Approximate Temporal Coalescing (ATC)

General approach:

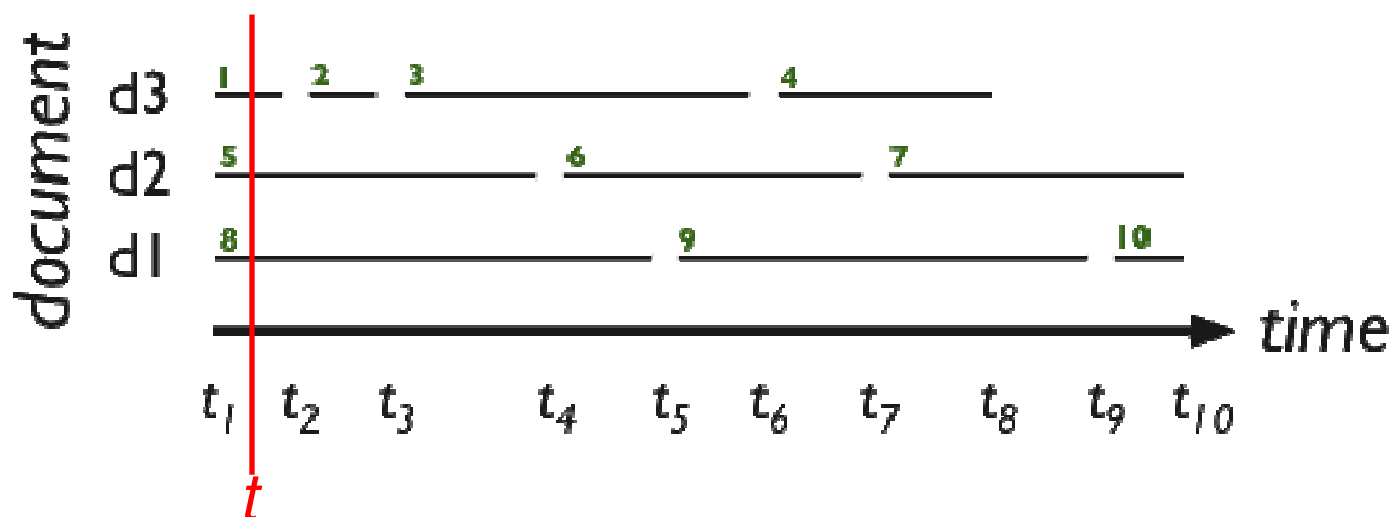
- Scan from left to right
- Maintain current estimate for representative p'
- When next value is encountered, check if it can be represented within the error margin
 - If not, close current subsequence



Tuning query performance

Problem:

Many postings are *ignored* during query processing

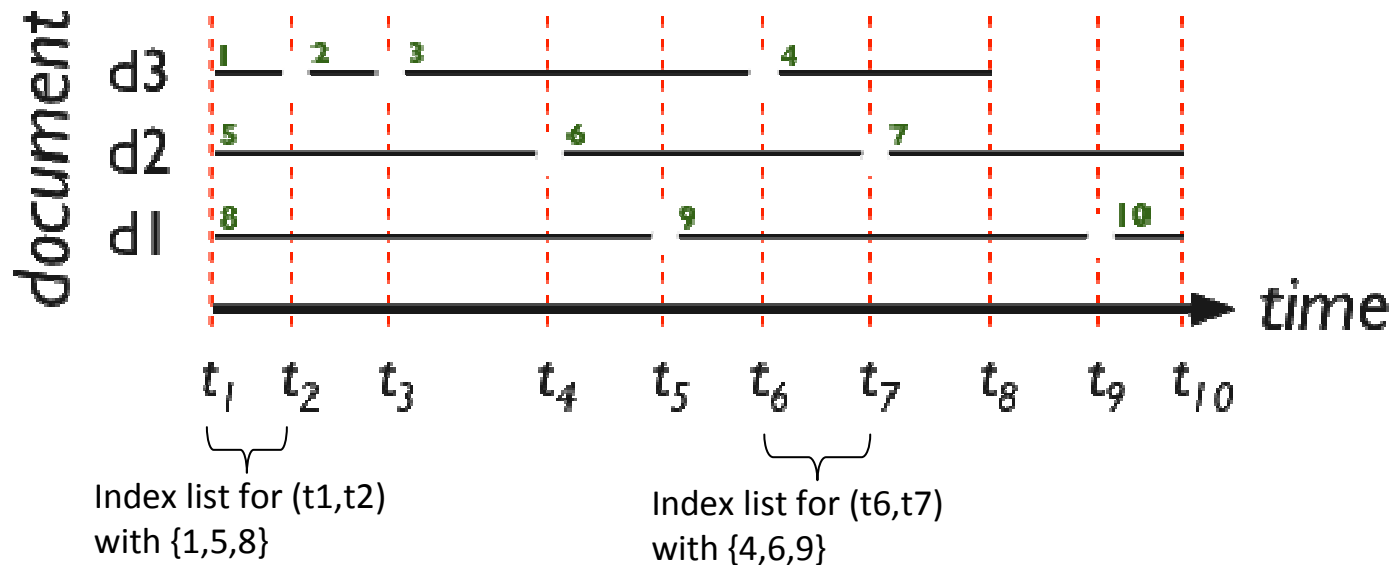


We read 10 postings,
but only {1, 5, 8} are needed

Tuning Query Performance: POPT

Idea:

Materialize smaller sublists containing only postings that **overlap** with a smaller interval



Maintaining a sublist for each elementary interval yields optimal query performance

Tuning Index Performance

Two extreme solutions up to now:

- ***space-optimal***: keep only a single list (SOPT)
- ***performance-optimal***: keep one list per elementary time-interval (POPT)

Now: two systematic techniques to trade-off space and performance

- ***performance-guarantee***: consumes minimal space while retaining a performance guarantee (PG)
- ***space-bound***: achieves best performance while not exceeding a space limit (SB)

Performance Guarantee (PG)

- consumes *minimal* space
- *guarantees* that for any t *at most $\gamma \cdot n_t$ postings are read* where n_t is the number of postings that exist at time t

Optimal solution computable for discrete time by means of induction (on the number of time points) *in $O(T^2)$ time and $O(T^2)$ space* (where T is the number of distinct timestamps in the list)

- start with elementary intervals (length 1)
- compute optimal solution for intervals of length $k+1$ from solutions for intervals of length $\leq k$

Space Bound (SB)

- achieves *minimal expected processing cost* (i.e., expected length of the list that is scanned)
- consumes *at most $\kappa \cdot n$ space* where n is the length of the original list

Optimal solution computable using dynamic programming in *$O(n^4)$ time* and $O(n^3)$ space

Approximate solution computable in *$O(T^2)$ time* and $O(T)$ space using simulated annealing

Experimental Evaluation: Setup

Implementation:

Java, Oracle 10g

Datasets:

- WIKI: Revision history of English Wikipedia (2001-2005)
892K documents / 13,976K versions / 0.7 TBytes
- UKGOV: Weekly crawls of 11 .gov.uk sites (2004-2005)
502K documents / 8,687K versions / 0.4 TBytes

Queries:

- 300 keyword queries from AOL query log that most frequently produced a result click on en.wikipedia.org / .gov.uk
- Each keyword query is assigned one time point per month in the collection's lifespan (18K / 7.2K time-travel queries in total)

Experimental Evaluation: Setup

Implementation:

Java, Oracle 10g

Datasets:

- WIKI: Revision history of English Wikipedia (2001-2005)
892K documents / 13,976K versions / 0.7 TBytes
- UKGOV: Weekly crawls of 11 .gov.uk sites (2004-2005)

WIKI:

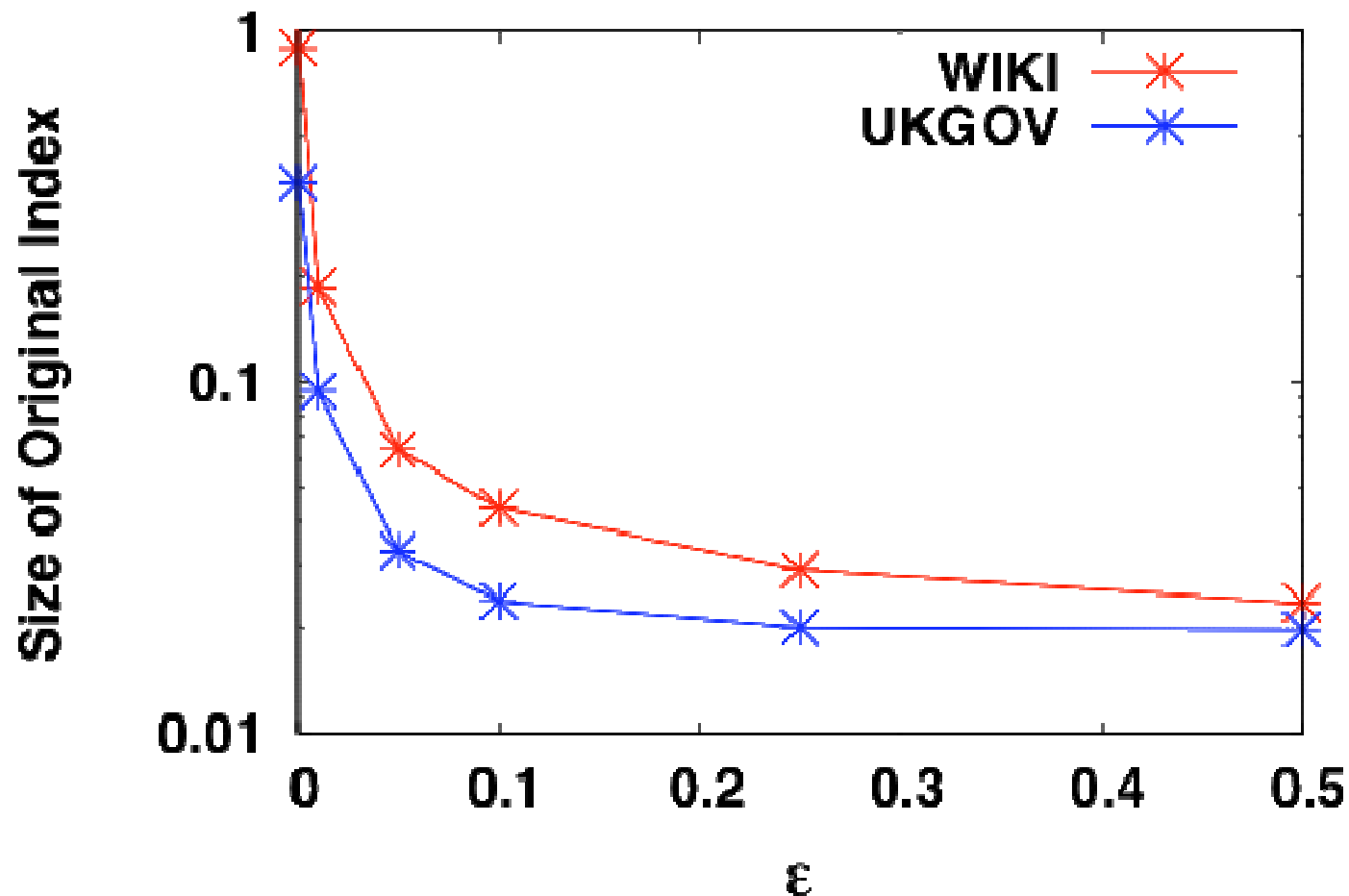
ten commandments, abraham lincoln, da vinci code, harlem renaissance...

UKGOV:

1901 uk census, british royal family, migrant worker statistics, witness intimidation...

Approximate Temporal Coalescing

Indexes computed for different values of threshold ε



At the same time provides excellent result quality

Sublist Materialization - Setup

Start with index created by ATC for $\varepsilon = 0.10$

For terms in query workloads (422/522) apply

- SOPT and POPT
- PG for γ varying between 1.10 and 3.00
- SB for κ varying between 1.10 and 3.00

Report

- Space, i.e., total number of postings in materialized sublists
- Expected Processing Cost (EPC), i.e., expected length of scanned list for random term and time

Performance Guarantee

	WIKI		UKGOV	
	Space	EPC	Space	EPC
P_{OPT}	14,428%	100%	11,406%	100%
S_{OPT}	100%	963%	100%	147%

Performance Guarantee

	WIKI		UKGOV	
	Space	EPC	Space	EPC
$\gamma = 1.10$	1,004%	106%	616%	103%
$\gamma = 1.50$	295%	132%	233%	117%
$\gamma = 2.00$	195%	160%	163%	125%
$\gamma = 3.00$	145%	207%	132%	133%

EPC = Expected Processing Cost

Space Bound

	WIKI		UKGOV	
	Space	EPC	Space	EPC
P_{OPT}	14,428%	100%	11,406%	100%
S_{OPT}	100%	963%	100%	147%

Space Bound

	WIKI		UKGOV	
	Space	EPC	Space	EPC
$\kappa = 3.00$	288%	139%	273%	107%
$\kappa = 2.00$	194%	171%	180%	119%
$\kappa = 1.50$	146%	214%	131%	131%
$\kappa = 1.10$	109%	406%	104%	145%

EPC = Expected Processing Cost

Web Dynamics

Part 5 – Searching the Past

5.1 Time-travel problems

5.2 Efficient Time-Travel Search

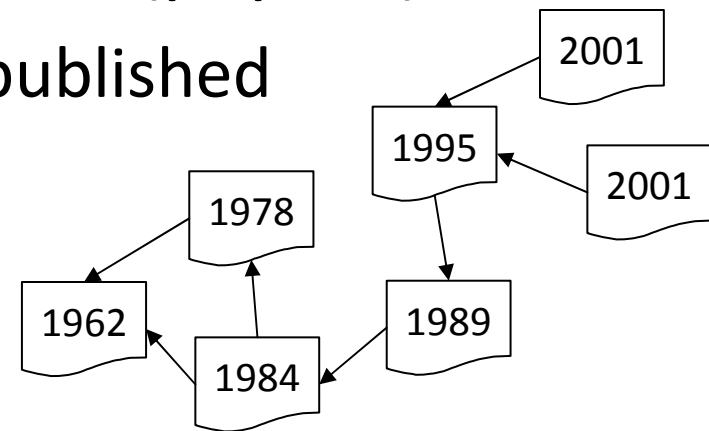
5.3 Temporal measures of page importance

Differences between Citations and Links

- Citations in printed documents (papers)

- never change once paper is published
- mostly to recent documents

⇒ Old papers hardly cited,
negative authority bias

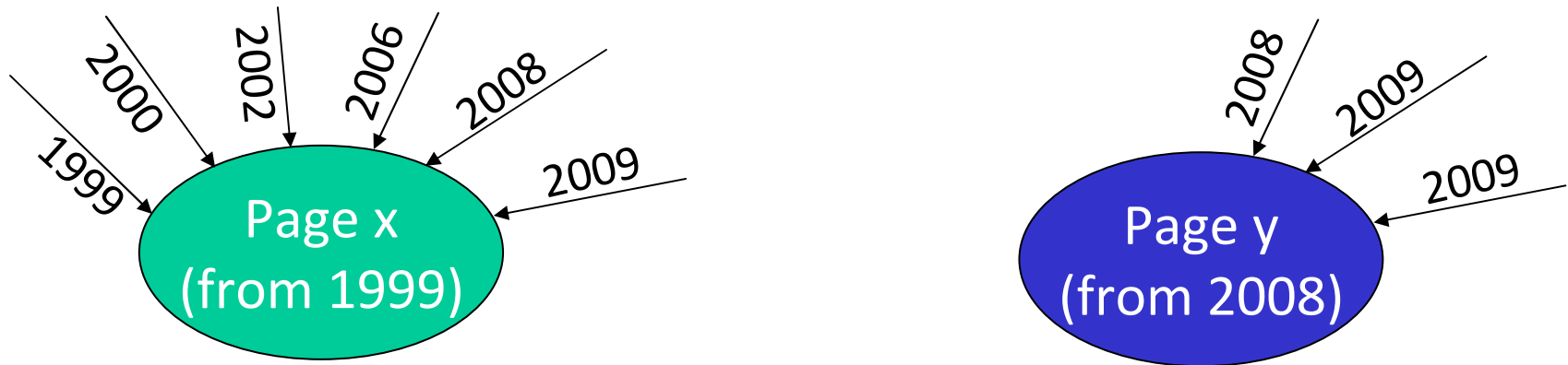


- Links on the Web

- frequently change after page is published
- old (but updated!) pages still get many new links

⇒ Old pages have **positive authority bias**

Temporal Development of Links



- PageRank (HITS, ...): **x** more authoritative than **y**
- But:
 - **x** has 6 links in 10 years
 - **y** has 3 links in 2 years
 - ⇒ **y** a lot more dynamic and up-to-date than **x**,
but difficult to beat **x**'s “temporal advantage”
- Why **Temporal notions of authority required!** 1999?

Example: Search for SIGMOD conference

Old pages
dominate over
page for 2009
conference



Web Bilder Video Maps News Shopping E-Mail Mehr ▼

Google [Erweiterte Suche](#)
[Einstellungen](#)

Suche: ☒ Das Web ☐ Seiten auf Deutsch ☐ Seiten aus Deutschland

Web

[ACM SIGMOD Conference](#) - [[Diese Seite übersetzen](#)]
Michael Stonebraker (Ed.): Proceedings of the 1992 ACM SIGMOD International Conference on Management of Data, San Diego, California, June 2-5, 1992. ...
www.informatik.uni-trier.de/.../sigmod/index.html - [Im Cache](#) - [Ähnlich](#)

[ACM SIGMOD Conference 1999: Philadelphia, Pennsylvania, USA](#) - [[Diese Seite übersetzen](#)]
Alex Delis, Christos Faloutsos, Shahram Ghandeharizadeh (Eds.): SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, ...
www.informatik.uni-trier.de/.../sigmod/sigmod99.html - [Im Cache](#) - [Ähnlich](#)
[Weitere Ergebnisse von www.informatik.uni-trier.de »](#)

[The ACM SIGMOD/PODS Conference: Vancouver, 2008 - Welcome](#) - [[Diese Seite übersetzen](#)]
28th ACM SIGMOD/PODS International Conference on Management of Data / Principles of Database Systems, Vancouver, BC, Canada.
www.sigmod08.org/ - [Im Cache](#) - [Ähnlich](#)

[Conferences — Association for Computing Machinery](#) - [[Diese Seite übersetzen](#)]
Online registration, calendar, and links to ACM sponsored conferences.
www.acm.org/conferences/ - [Im Cache](#) - [Ähnlich](#)

[ACM SIGMOD Conference 2007: Beijing, China](#) - [[Diese Seite übersetzen](#)]
Chee Yong Chan, Beng Chin Ooi, Aoying Zhou (Eds.): Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, June 12-14, ...
www.sigmod.org/dblp/.../sigmod/sigmod2007.html - [Im Cache](#) - [Ähnlich](#)

[ACM SIGMOD Conference 1996: Montreal, Canada](#) - [[Diese Seite übersetzen](#)]
@proceedings[DBLP:conf/sigmod/96, editor = {H. V. Jagadish and Inderpal Singh Mumick}, title = {Proceedings of the 1996 ACM SIGMOD International Conference ...
www.sigmod.org/dblp/db/.../sigmod/sigmod96.html - [Im Cache](#) - [Ähnlich](#)
[Weitere Ergebnisse von www.sigmod.org »](#)

[SIGMOD/PODS 2003 Conference](#) - [[Diese Seite übersetzen](#)]
22th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems, San Diego, California, June 9-12, 2003.
db.ucsd.edu/sigmodpods03/ - [Im Cache](#) - [Ähnlich](#)

Modelling Temporal Changes

For each page p , maintain

- timestamp of **creation** $TS_C(p)$
- timestamp of **deletion** $TS_D(p)$
- set of timestamps of **modifications** $TS_M(p)$

(timestamp: amount of time units since time 0)

Analogous definitions for link (x,y) :

- timestamp of **creation** $TS_C(x,y)$: time when (x,y) added
- timestamp of **deletion** $TS_D(x,y)$: time when (x,y) del'ed
- set of timestamps of **modifications** $TS_M(x,y)$
- timestamp $TS(x,y)$: last modification time of page x

Timestamped Link Profile (TLP)

Goal: Measure the „activity“ of a topic on the Web

⇒ Construction of *Timestamped Link Profile*:

- Collect set of Web pages for the topic (e.g., by collecting results of keyword queries)
- Collect set of inlinks (x,y) to these pages (provided by search engines: `link:url`)
- Compute temporal distribution of timestamps of inlinks (partitioning time range into intervals)

Based on *limited sample* of the inlinks

Timestamps usually available for some inlinks only (last-modified timestamp of page)

Example TLP

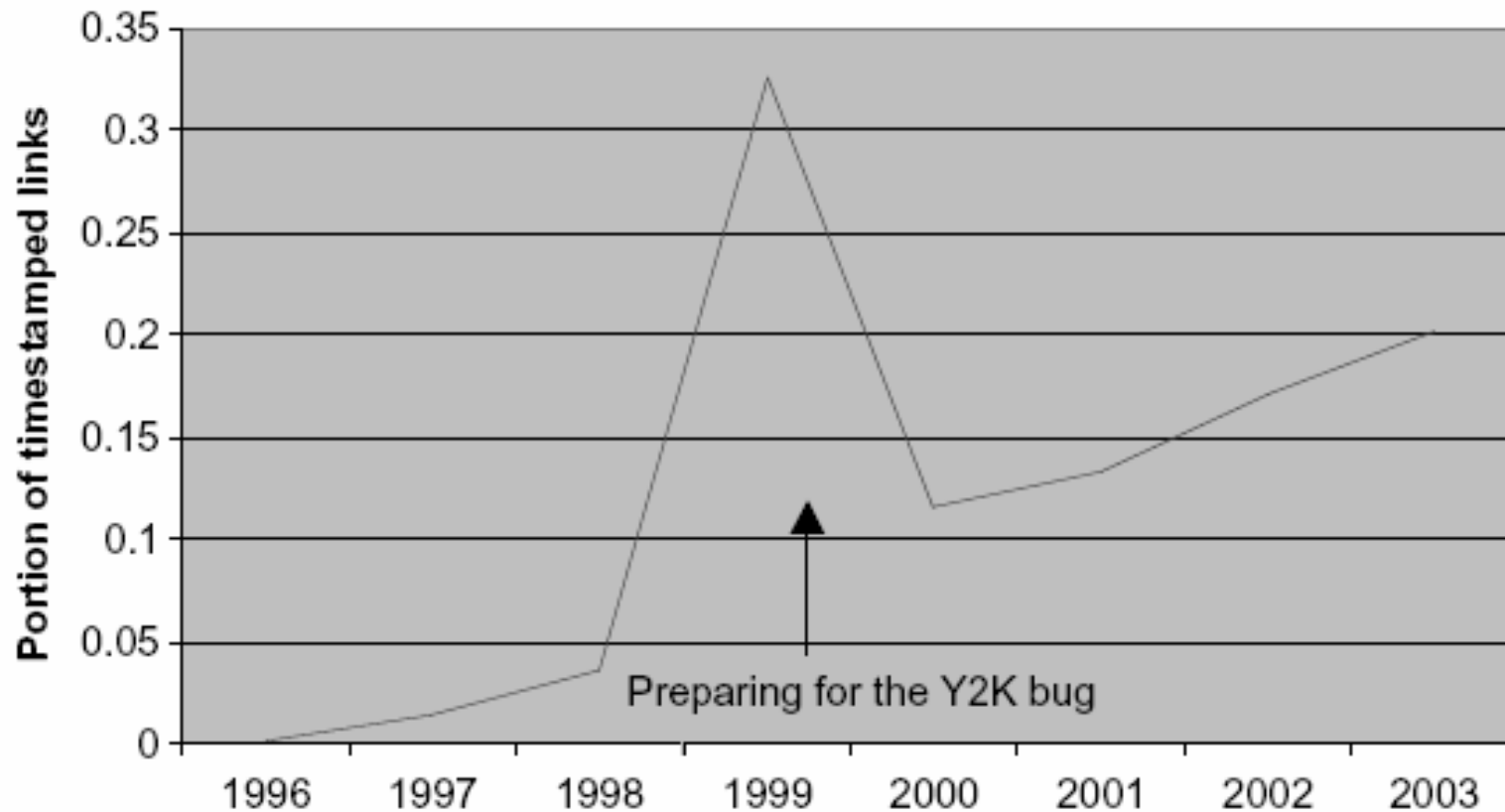


FIG. 2. TLP for the theme “Year 2000 Bug” (~750 timestamped links).

Amitay et al., JASIST 2004

Towards Timely Authorities

Goal: Determine *currently* authoritative pages
(opposed to those authoritative years ago,
but still around)

Intuition of [Amitay et al.]:

- Deviate from uniform link weight in HITS etc
- Give more weight to recent links:

$$\begin{aligned} \textit{weight}(x,y) &\propto 1/\textit{age}(x,y) \\ &= 1/(\textit{currentTime} - \textit{TS}(x,y)) \end{aligned}$$

(with linear or exponential decay)

Authoritative Pages in the Past

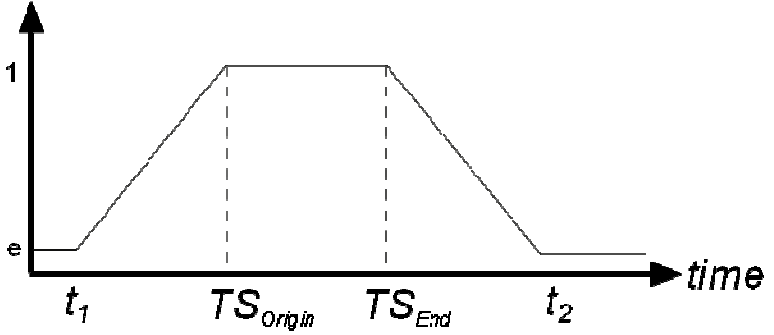
Goal: extend this approach towards

- finding important pages at *any interval* in the past
- including *page activity* as quality measure

Consider *interval of interest* $ti=[TS_{Origin}, TS_{End}]$ with additional *tolerance interval* $[t1, t2]$ where pages are less interesting, but still relevant to user
($t1 \leq TS_{Origin}$, $t2 \geq TS_{End}$)

Freshness

Freshness measures relevance of timestamp to interval of interest:

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


Freshness of node x : $f(x) = f(TS(x))$

Freshness of edge (x,y) : $f(x,y) = f(TS(x,y))$

Activity

Activity of set TS of timestamps measures frequency of change with respect to interval of interest:

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

Activity of node x : $a(x) = a(TS_M(x))$

Activity of edge (x,y) : $a(x,y) = a(TS_M(x,y))$

Restricting the Graph to an Interval

For graph G and interval of interest $ti=[ts,te]$ with tolerance interval $[t1,t2]$, consider **time projection**

$G_{ti}=(V_{ti},E_{ti})$ of $G=(V,E)$:

$$V_{ti}=\{v \in V \mid TS_C(v) \leq t_2 \wedge TS_D(v) \geq t_1\}$$

$$E_{ti}=\{(x,y) \in E \mid (x,y) \in V_{ti} \times V_{ti} \wedge TS_C(x,y) \leq t_2 \wedge TS_D(x,y) \geq t_1\}$$

Special case $t_1=t_2$: G_{ti} snapshot of G as of time t_1

Towards Temporal PageRank

Standard definition of PageRank:

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

Generalized version allowing for **non-uniform** transition and random jump probabilities:

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

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

Temporal Pagerank (T-Rank)

- **Modified *PageRank*** on G_{ti}
- 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

- 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

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

T-Rank – Random Jumps

- Random jumps favor *fresh and active* nodes/edges
- Coefficients w_{s_i} 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

$$s(y) = w_{s1} \cdot \frac{f(y)}{\sum_{z \in V} f(z)} + w_{s2} \cdot \frac{a(y)}{\sum_{z \in V} a(z)} + \\ w_{s3} \cdot \frac{\text{avg}\{f(v, y) \mid (v, y) \in E\}}{\sum_{z \in V} \text{avg}\{f(w, z) \mid (w, z) \in E\}} + w_{s4} \cdot \frac{\text{avg}\{a(v, y) \mid (v, y) \in E\}}{\sum_{z \in V} \text{avg}\{a(w, z) \mid (w, z) \in E\}}$$

T-Rank Experiment: DBLP

Digital Bibliography & Library Project (DBLP) freely available
bibliographic dataset (as XML)

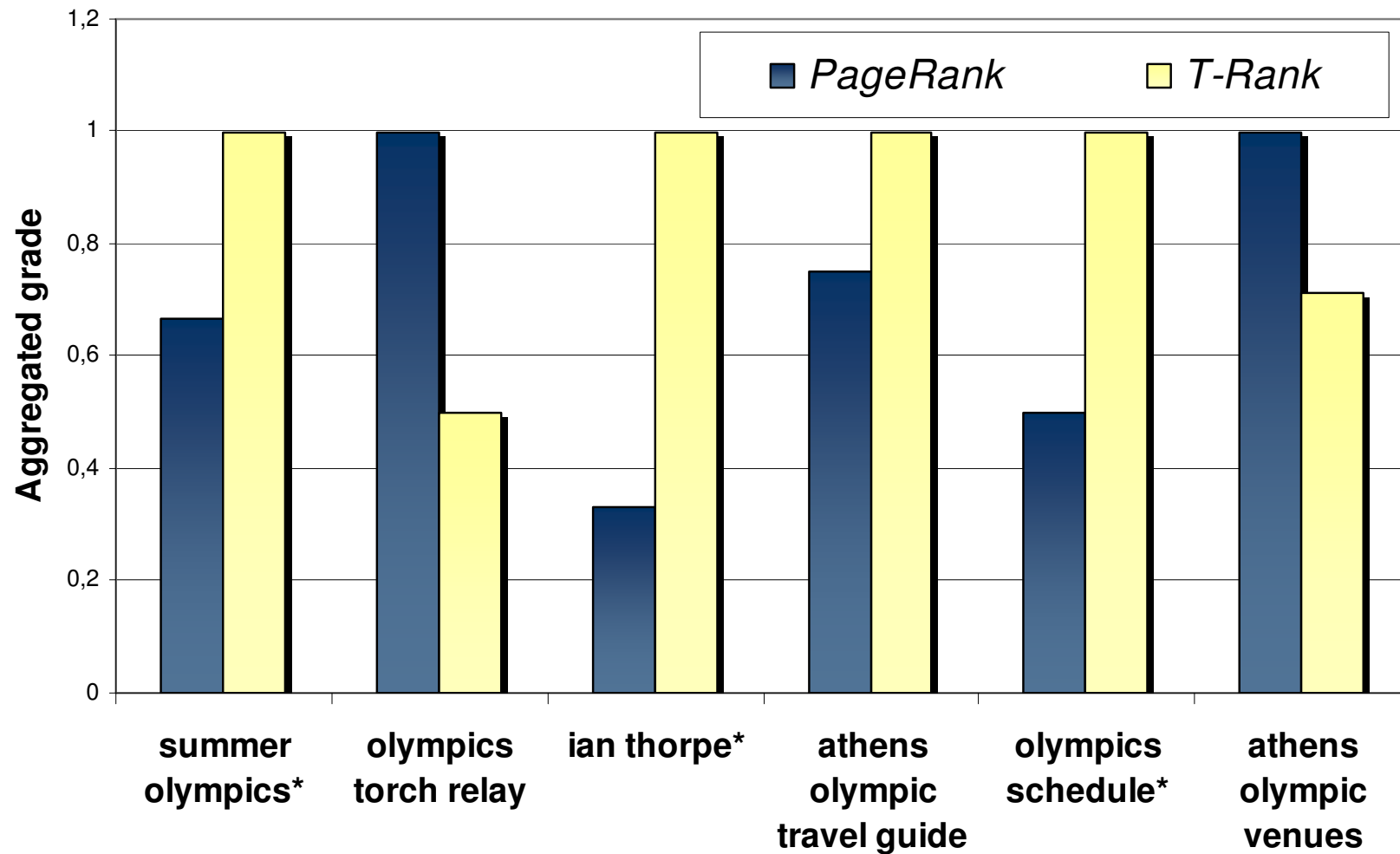
Evolving graph derived from DBLP: Authors as nodes, citations as edges

	<i>PageRank 2000s</i>	<i>T-Rank 2000s</i>
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

T-Rank Experiment: Web

- **Theme: 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

T-Rank Experiment: Web



Berberich et al, Internet
Mathematics 2006

References

Time-Travel Search:

- Klaus Berberich et al.: *A Time Machine for Text Search*, SIGIR Conference, 2007
- Klaus Berberich et al.: *FluxCapacitor: Efficient Time-Travel Text Search*, VLDB Conference, 2007

Temporal Link Analysis:

- L. Adamic & B.A. Huberman: *The Web's hidden order*, CACM 44(9), 2001
- Einat Amitay et al.: *Trend Detection Through Temporal Link Analysis*, Journal of the American Society for Information Science and Technology 55, pp. 1-12, 2004
- Ricardo Baeza-Yates et al.: *Web Structure, Dynamics and Page Quality*, SPIRE Conference, 2002
- Klaus Berberich et al.: *Time-Aware Authority Ranking*, Internet Mathematics 2(3), 2006
- Klaus Berberich et al.: *A Pocket Guide to Web History*, SPIRE Conference, 2007
- Philip S. Yu et al.: *On the Temporal Dimension of Search*, WWW Conference, 2004