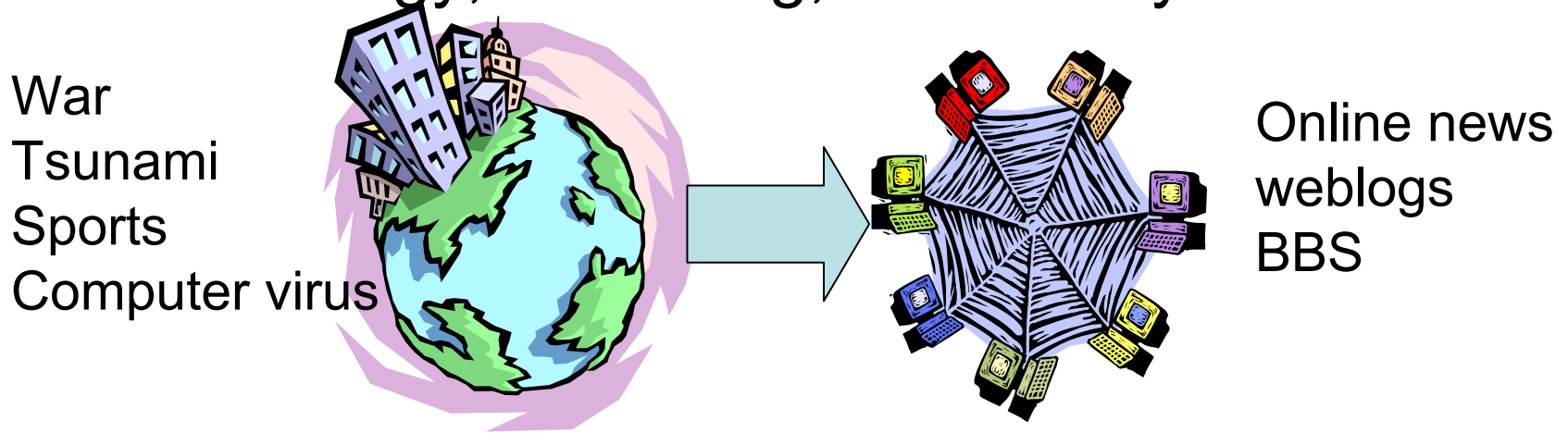


What's Really New on the Web? Identifying New Pages from a Series of Unstable Web Snapshots

Masashi Toyoda and Masaru Kitsuregawa
IIS, University of Tokyo

Web as a projection of the world

- Web is now reflecting various events in the real and virtual world
- Evolution of past topics can be tracked by observing the Web
- Identifying and tracking **new information** is important for observing **new trends**
 - Sociology, marketing, and survey research



Observing Trends on the Web (1/2)

- Recall (Internet Archive) [Patterson 2003]
 - # pages including query keywords

Recall BETA "Terrorism" go After January 1996 before May 2003

Searched For: "Terrorism" from January 1996 to May 2003 Approximately 3,912,164 URL months

[About Recall](#) [Feedback](#)

Jan 1996 Jan 1998 Jan 2000 Jan 2002 Apr 2003 Jan 1996 1998 2000 2002/2003

Counter Terrorism
Prevention of Terrorism
Response to Terrorism
War Against Terrorism
International Terrorism
War on Terrorism

Percentage of Returned Pages Mentioning X by Date

Percentage of Peak by Month

Narrow results by:

Instead of Terrorism
Do you Mean?

Capitalization Variants

Categories

People
Terrorism
International
Security

Topics

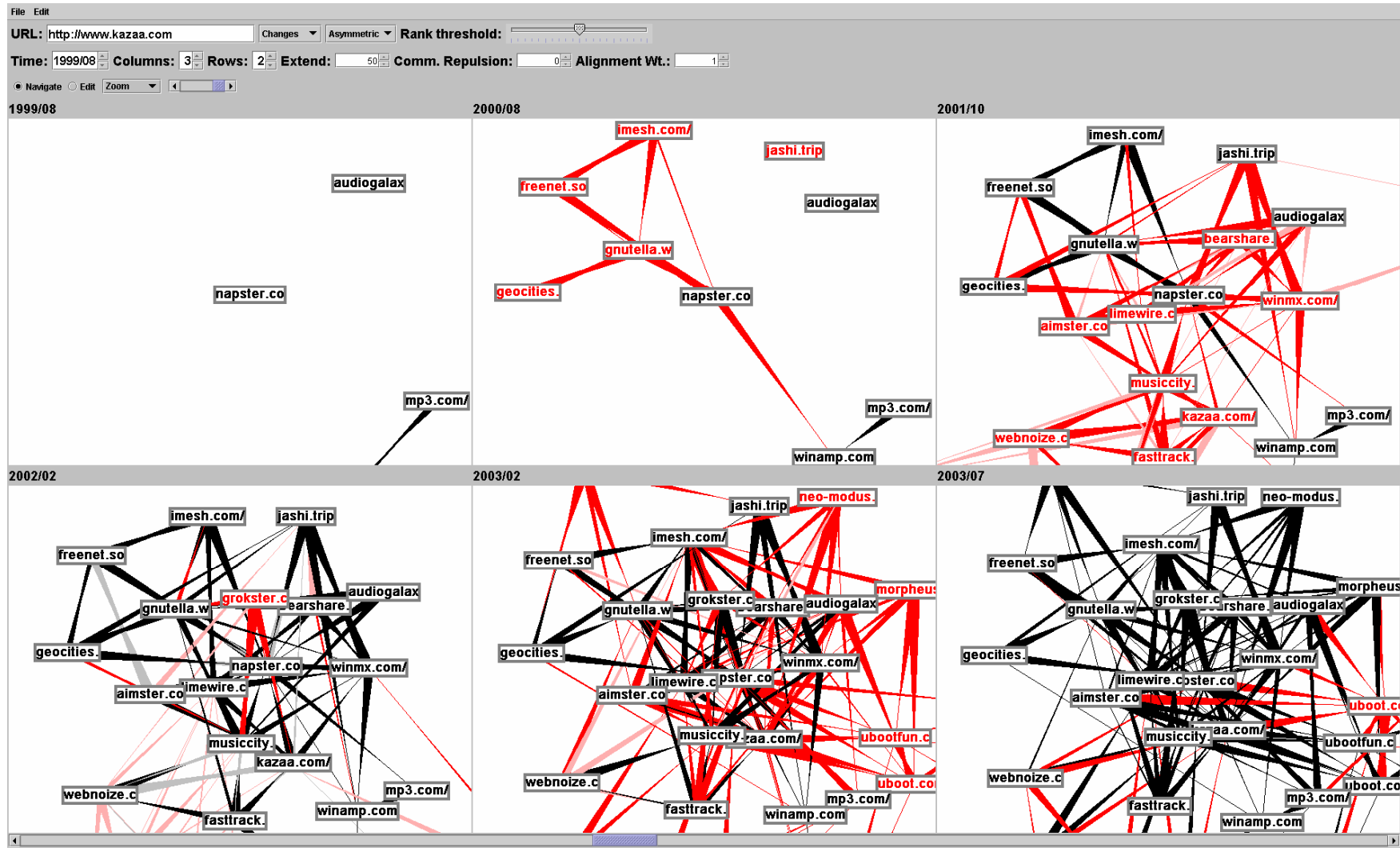
War on Terrorism
Nuclear Terrorism
International Terrorism

[Removing Syria from the List of State Sponsors of Terrorism](#) December 2000
Home Spotlight International **Terrorism** Counter-**Terrorism** Arab-Israeli Conflict Search Products Services Forum ...
According to the 1998 edition of the State Department's annual report, Patterns of Global **Terrorism** , Syria provides safe haven and logistical support to ...
Removing Syria from the List of State Sponsors of **Terrorism** Between Peace and Counter**Terrorism**
<http://www.ict.org.il:80/articles/schenker.htm>

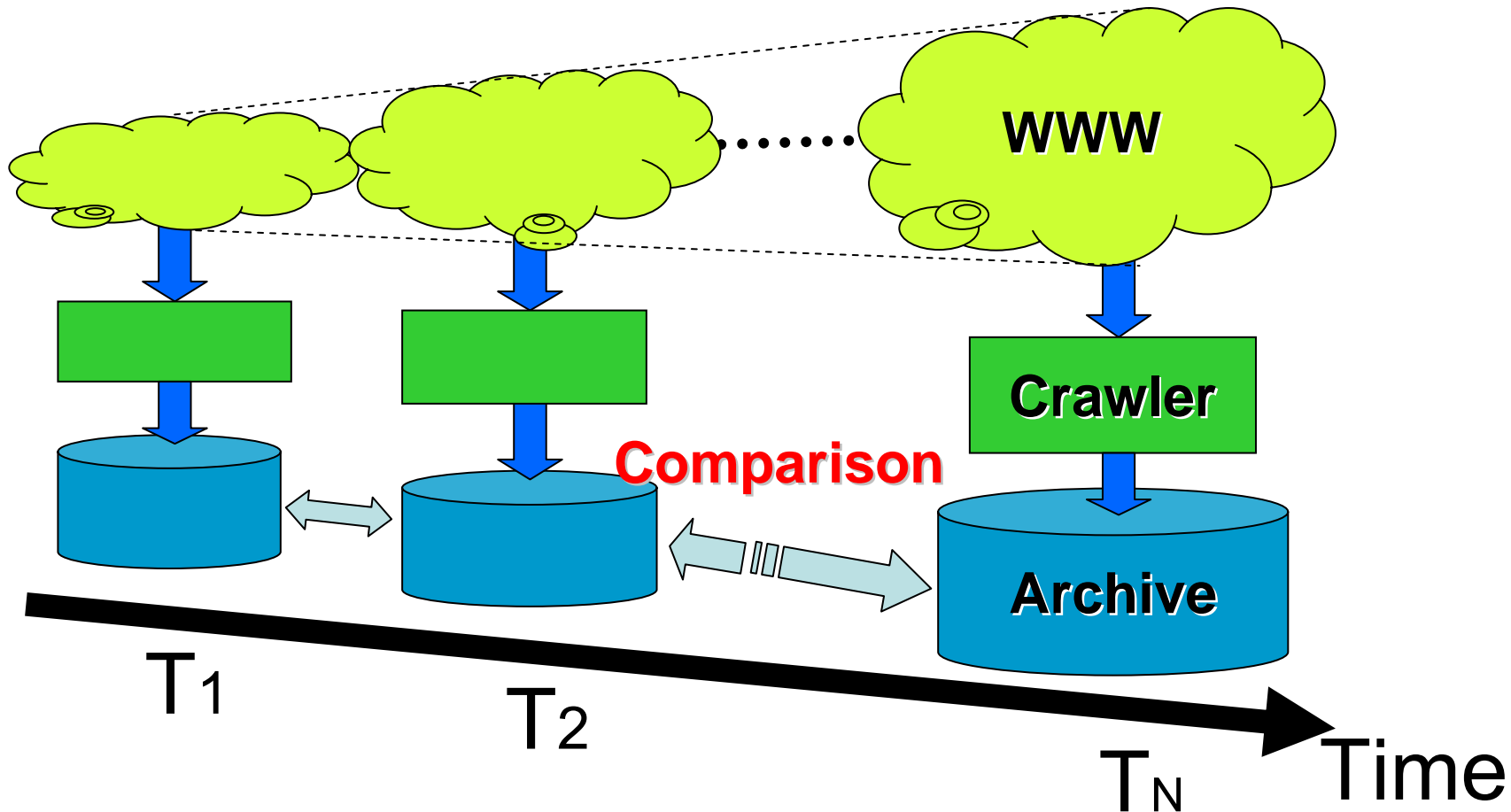
[State's Taylor Summarizes Annual Global Terrorism Report, May 21, 2002](#) June 2002
" President Bush called on all nations to unite in a coalition and to use every element of national power to fight this threat, " Ambassador Francis Taylor said May 21 during a briefing on the new " Patterns of Global **Terrorism**: 2001" report. ...
The entire " Patterns of Global **Terrorism**: 2001" report can be found on the State Department's Internet website at <http://www...>
Patterns of Global **Terrorism** 2001 describes a complex web of nations, ethnicities, financial networks and arms shipments that constitutes today's terrorist threat
http://www.usembassy.it:80/file2002_05/alia/a2052103.htm

Observing Trends on the Web (2/2)

- WebRelievo [Toyoda 2005]
 - Evolution of link structure



Periodic Crawling for Observing Trends on the Web



Difficulties in Periodic Crawling (1/2)

- **Stable crawls** miss new information
 - Crawling a fixed set of pages [Fetterly et al 2003]
 - ↑ Can identify changes in the pages
 - ↓ Overlook new pages
 - Crawling all the pages in a fixed set of sites [Ntoulas et al 2004]
 - ↑ Can identify new pages in these sites
 - ↓ Overlook new sites
 - ↓ Possible only on a small subset of sites
- **Massive crawls** are necessary for discovering new pages and new sites

Difficulties in Periodic Crawling (2/2)

- Massive crawls make snapshots **unstable**
 - Cannot crawl the whole of the Web
 - # of uncrawled pages overwhelms # of crawled pages even after crawling 1B pages [Eiron et al 2004]
 - **Novelty of a page crawled for the first time remains uncertain**
 - The page might exist at the previous time
 - “Last-Modified” time guarantees only that the page is older than that time

Our Contribution

- Propose a *novelty measure* for estimating the certainty that a newly crawled page is really new
 - New pages can be extracted from a series of unstable snapshots
- Evaluate the precision, recall, and miss rate of the novelty measure
- Apply the novelty measure to our Web archive search engine

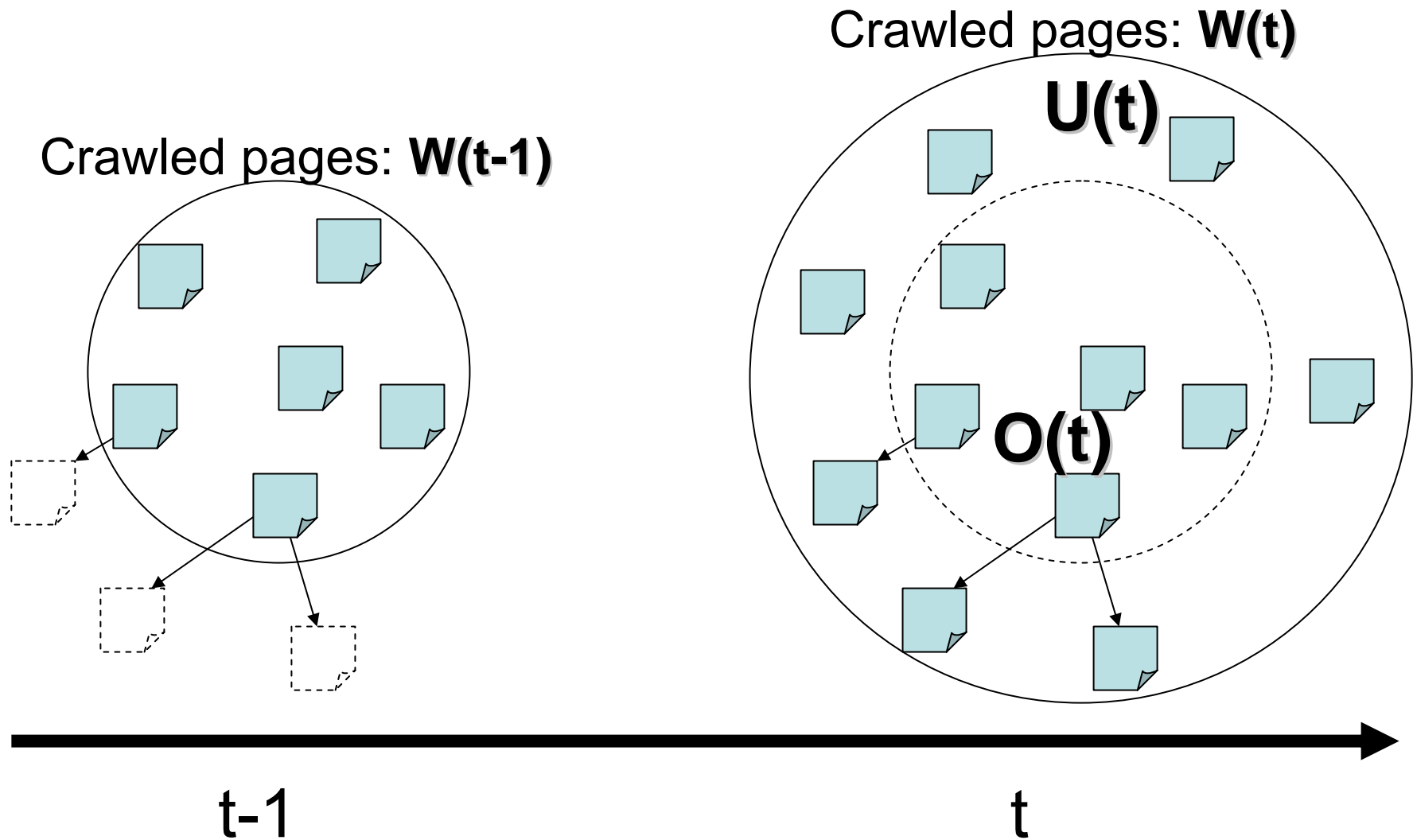
Basic Ideas

- The novelty of a page p is the certainty that p appeared between $t-1$ and t
 - p appears when it can first be crawled and indexed
 - p is new when it is pointed to only by new links
 - If only new pages and links point to p , p may also be novel
- The novelty measure can be defined recursively and can be calculated in a similar way to PageRank [Brin and Page 1998]
- Reverse of the decay measure [Bar-Yossef et al 2004]
 - p is decayed if p points to dead or decayed pages

Novelty Measure

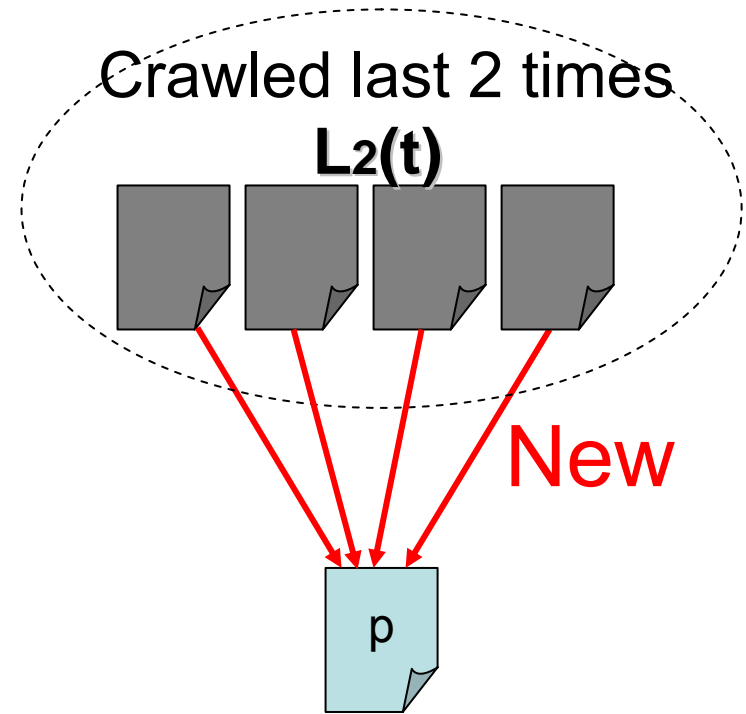
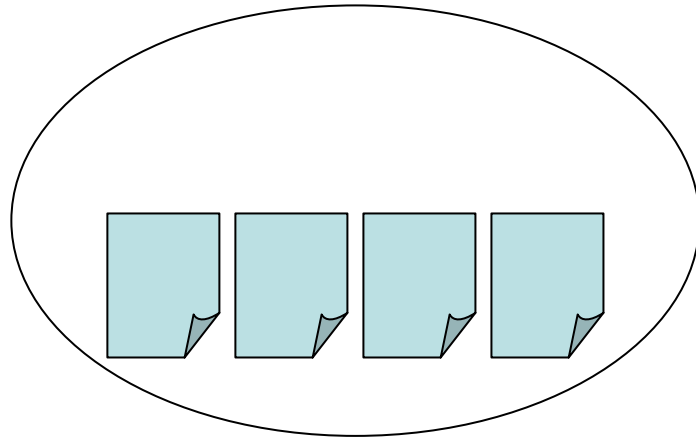
- **$N(p)$** : The novelty of page **p** ($0 \leq 1$)
 - 1: The highest certainty that **p** is novel
 - 0: The novelty of **p** is totally **unknown** (not old)
- Pages in a snapshot **$W(t)$** are classified into old pages **$O(t)$** and unknown pages **$U(t)$**
- Each page p in **$U(t)$** is assigned **$N(p)$**

Old and Unknown Pages



How to Define Novelty Measure

If all in-links come from pages crawled last 2 times ($L_2(t)$)



$N(p) = 1$

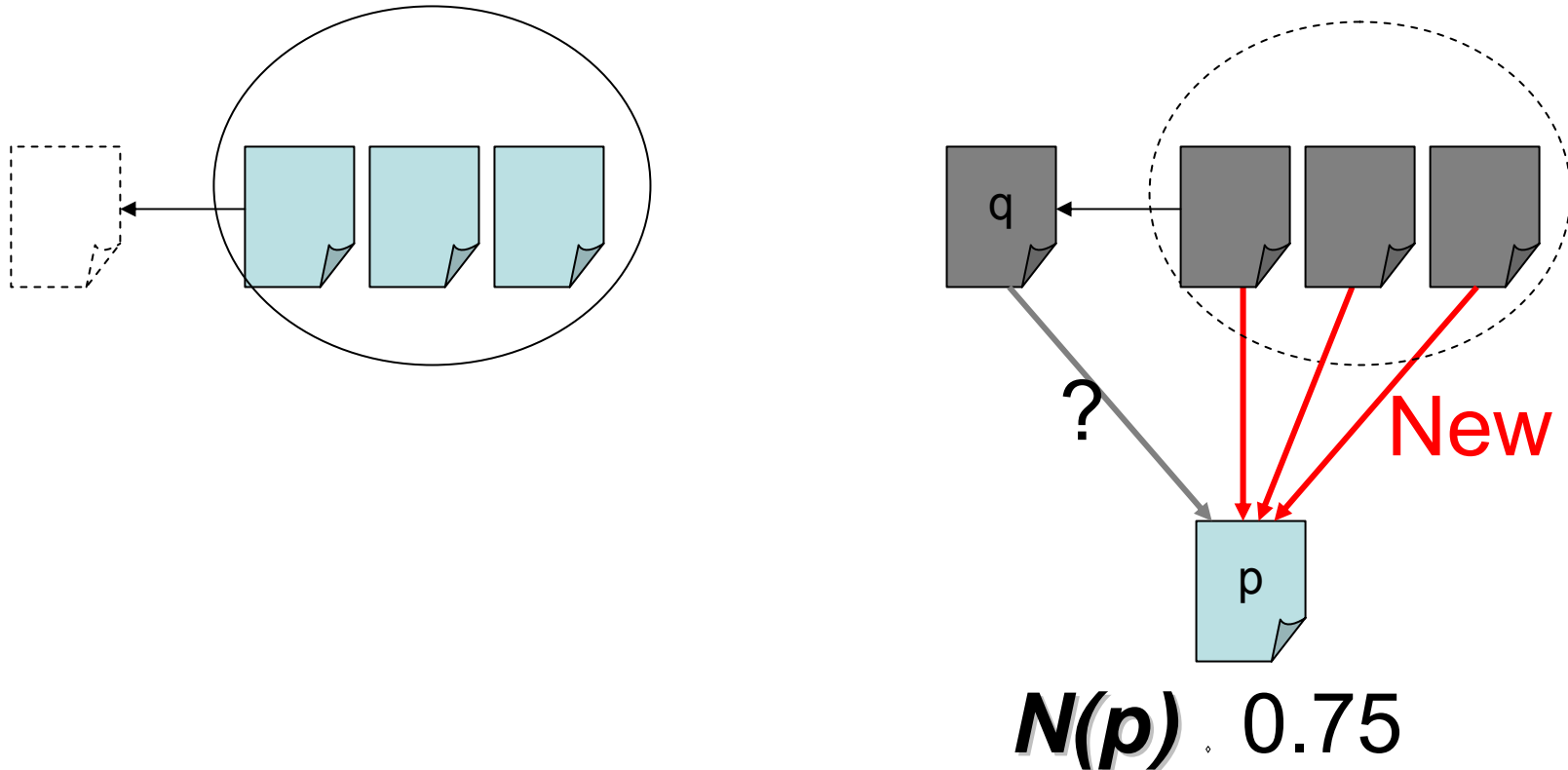


t-1

t

How to Define Novelty Measure

If some in-links come from $\mathbf{O}(t)\text{-L}_2(t)$

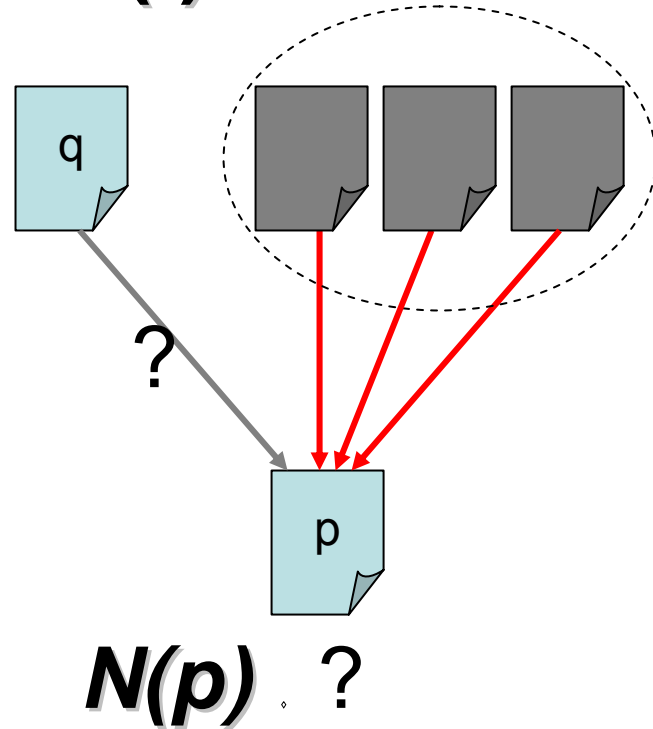
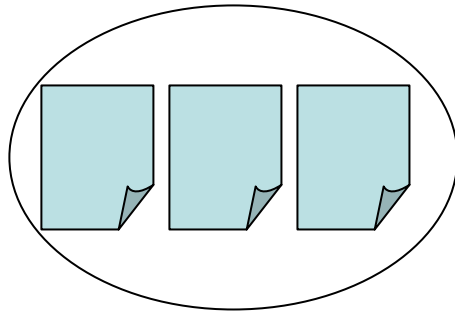


$t-1$

t

How to Define Novelty Measure

If some in-links come from $\mathbf{U}(t)$?

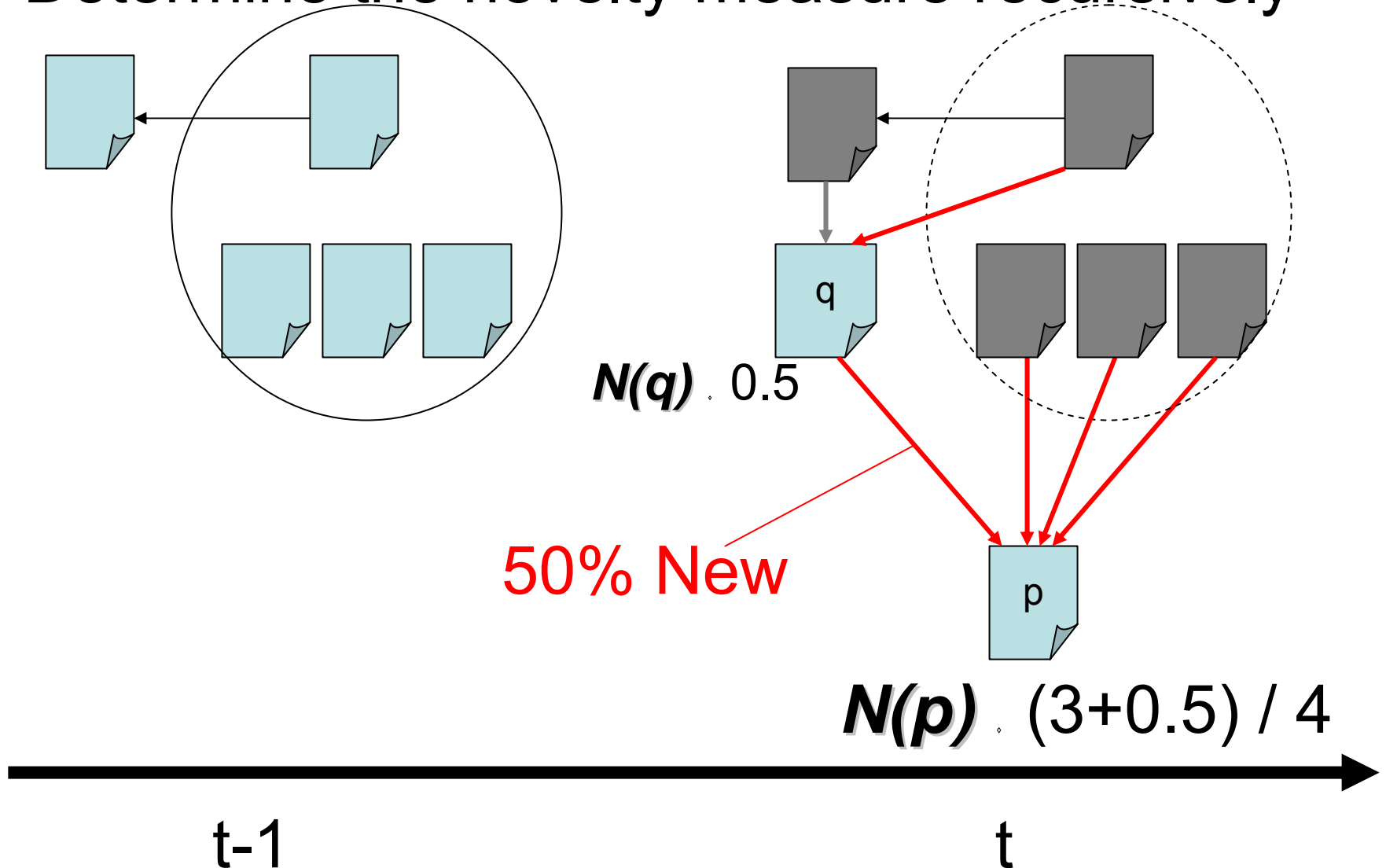


$t-1$

t

How to Define Novelty Measure

Determine the novelty measure recursively



Definition of Novelty Measure

$$\mathcal{N}(p) = (1 - \delta) \frac{\sum_{(q,p) \in I(p)} n(q,p)}{|I(p)|}$$

$$n(q,p) = \begin{cases} 1 & q \in L_2(t_k) \\ 0 & q \in O(t_k) \setminus L_2(t_k) \\ \mathcal{N}(q) & q \in U(t_k) \end{cases} \quad (1)$$

- δ : damping factor
 - probability that there were links to p before $t-1$

Experiments

- Data set
- Convergence of calculation
- Distribution of the novelty measure
- Precision and recall
- Miss rate

Data Set

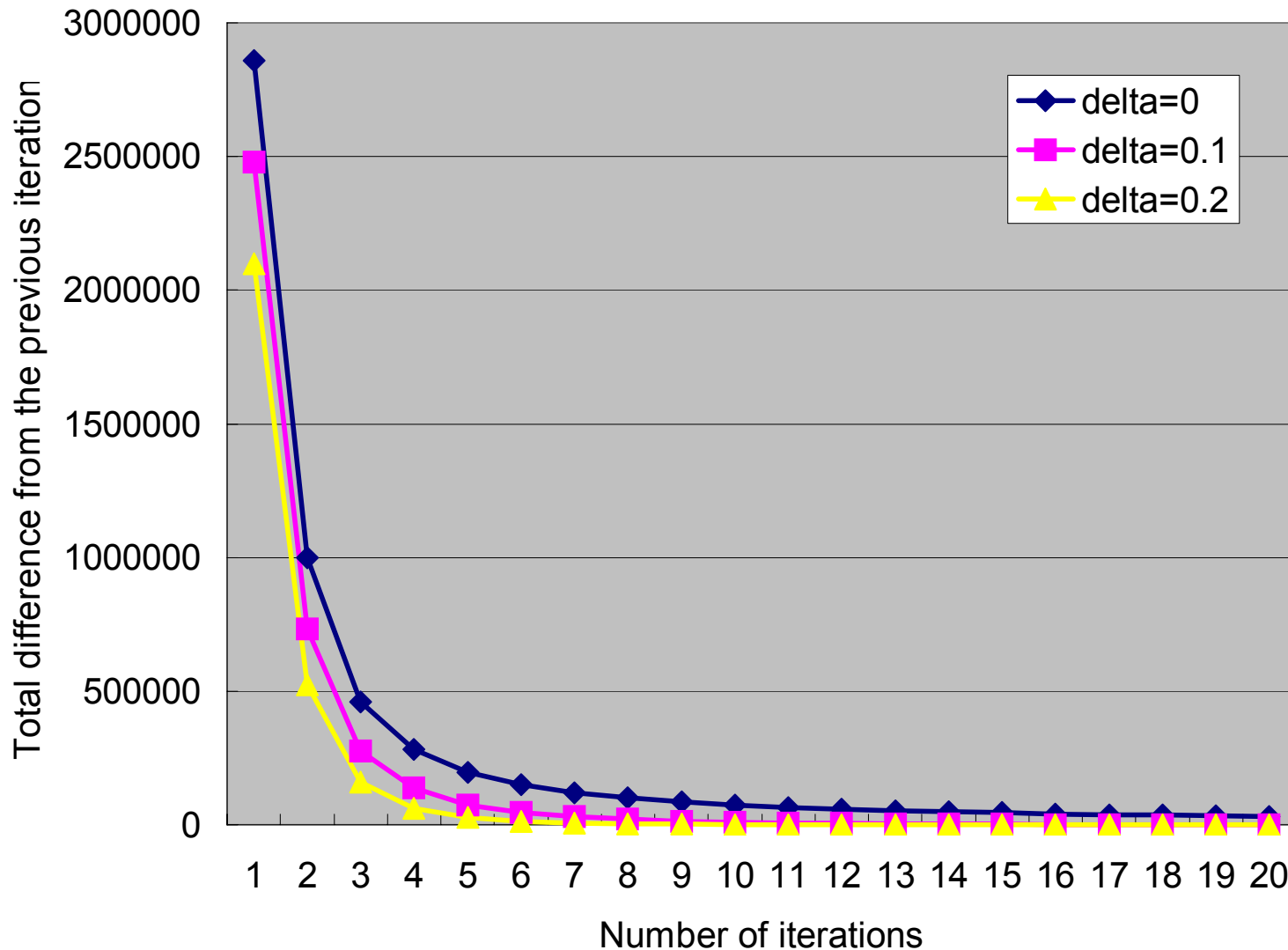
- A massively crawled Japanese web archive
 - 2002: .jp only
 - 2003.: Japanese pages in any domain

Time	Period	Crawled pages	Links
1999	Jul to Aug	17M	120M
2000	Jun to Aug	17M	112M
2001	Oct	40M	331M
2002	Feb	45M	375M
2003	Feb	66M	1058M
2003	Jul	97M	1589M
2004	Jan	81M	3452M
2004	May	96M	4505M

Time	Jul 2003	Jan 2004	May 2004
$ L2(t) $	49M	61M	46M
$ O(t) - L2(t) $	23M	14M	20M
$ U(t) $	25M	6M	30M
$ W(t) $	97M	81M	96M

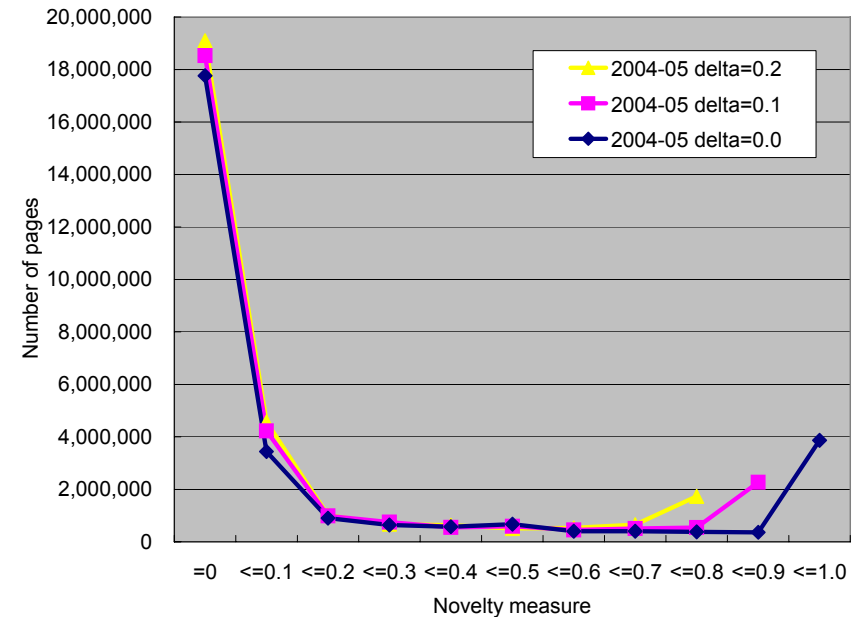
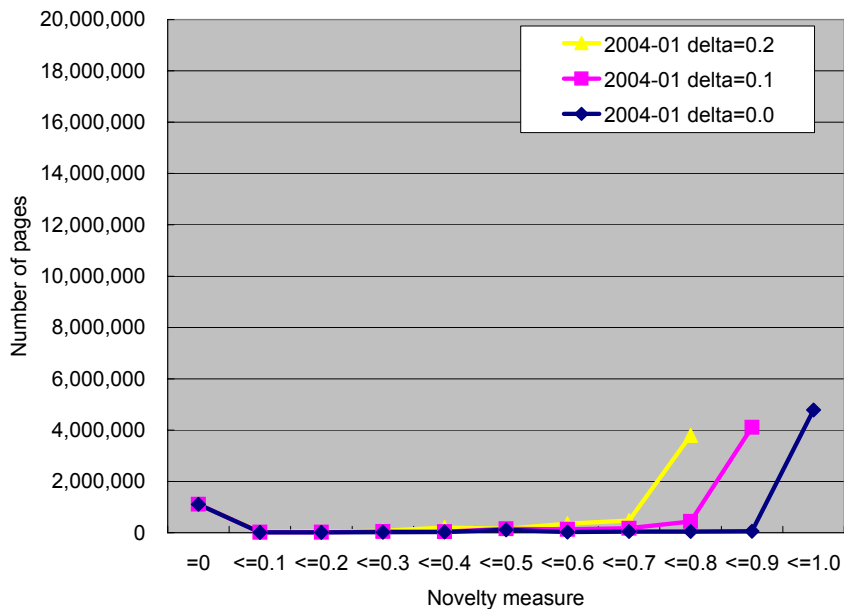
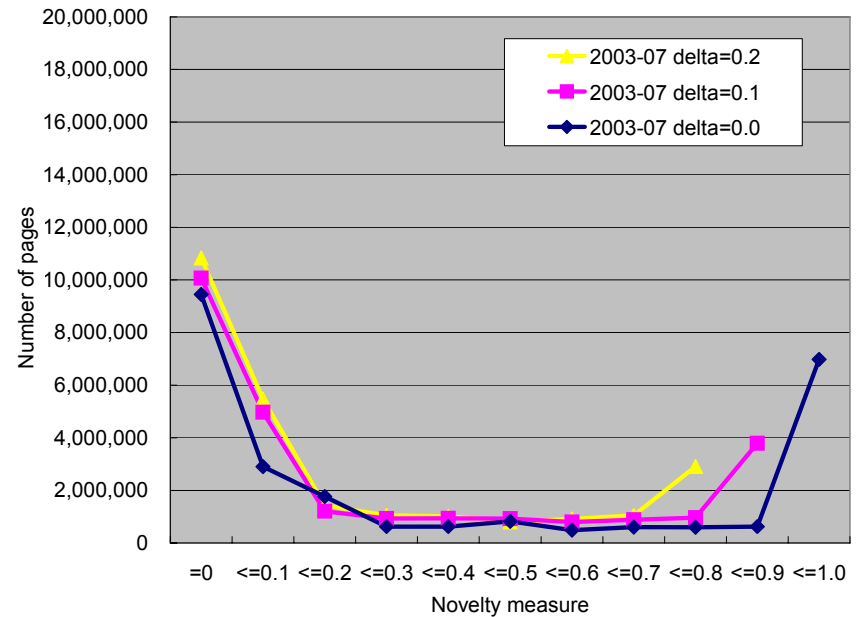
Convergence of Calculation

- 10 iterations are sufficient for $0 < \dots$



Distributions of the Novelty Measure

- Have 2 peaks on 0 and MAX
 - cf. Power-law of in-link distribution
- Depend on the fraction of $L_2(t)$ and $U(t)$
- Not change drastically by delta except for the maximum value



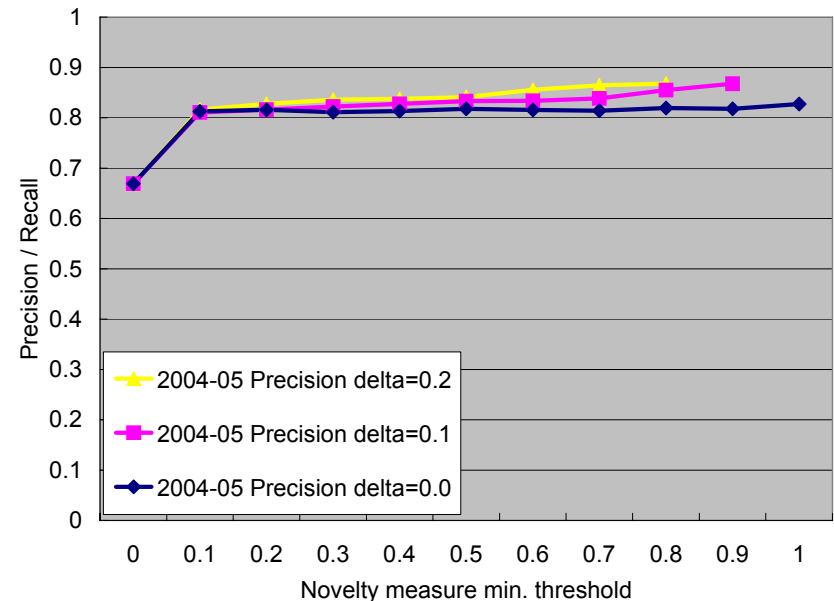
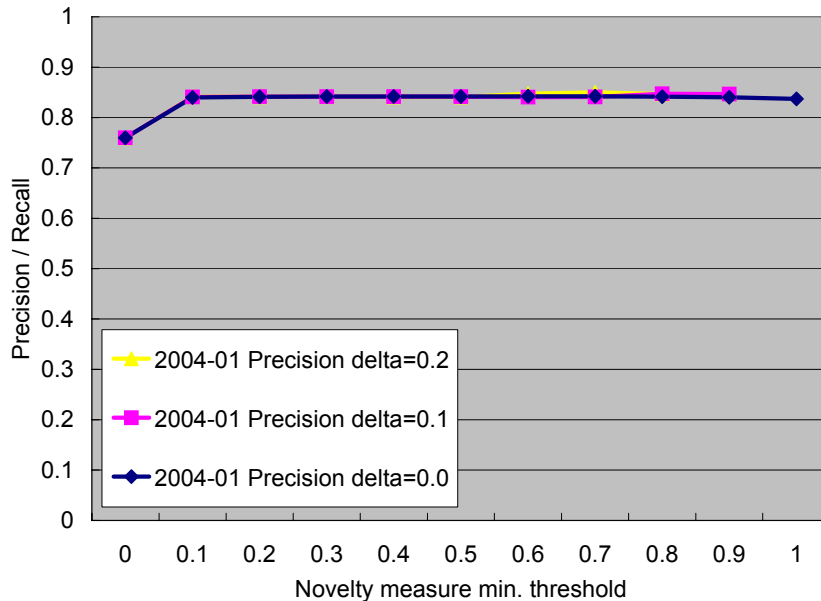
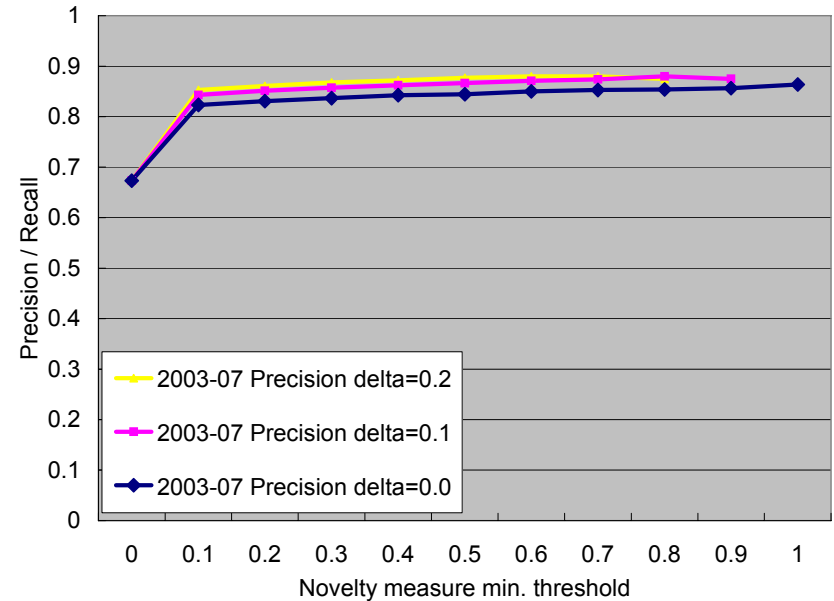
Precision and Recall

- Given threshold τ ,
 p is judged to be novel when $\tau < N(p)$
 - Precision: $\#(\text{correctly judged}) / \#(\text{judged to be novel})$
 - Recall: $\#(\text{correctly judged}) / \#(\text{all novel pages})$
- Use URLs including dates as a golden set
 - Assume that they appeared at their including time
 - E.g. <http://foo.com/2004/05>
 - Patterns: YYYYMM, YYYY/MM, YYYY-DD

	Jul 2003	Jan 2004	May 2004
With old date (before t-1)	299,591 (33%)	87,878 (24%)	402,365 (33%)
With new date (t-1 to t)	593,317 (65%)	270,355 (74%)	776,360 (64%)
With future date (after t)	24,286 (2%)	7,679 (2%)	36,476 (3%)
Total	917,194 (100%)	365,912 (100%)	1,215,201 (100%)

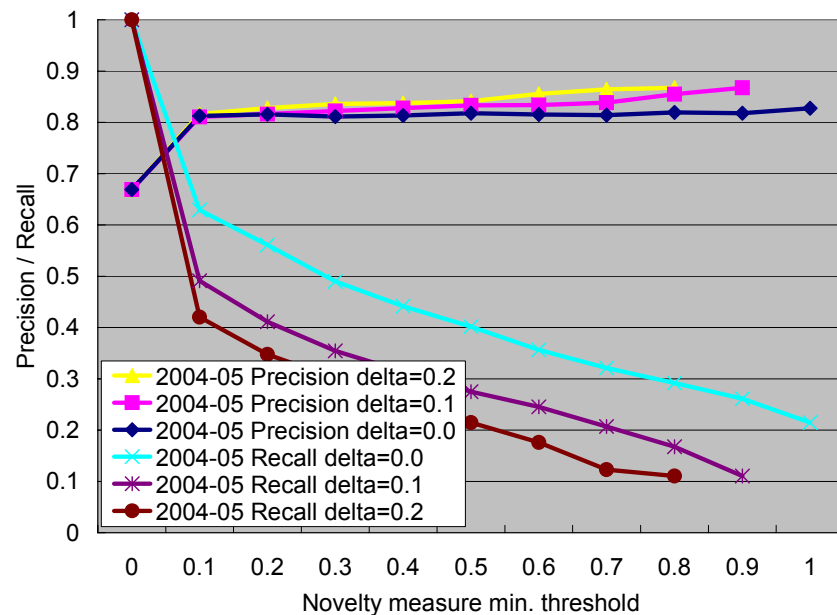
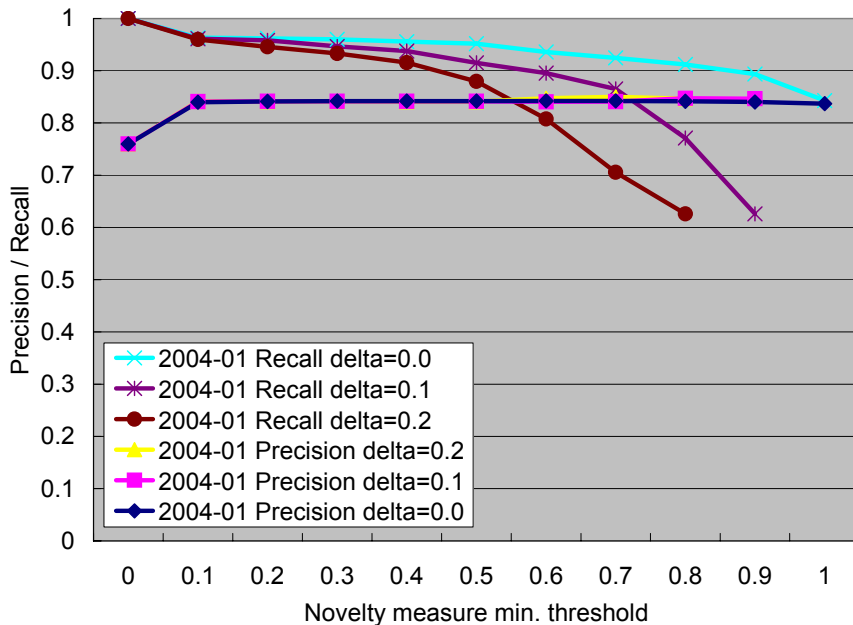
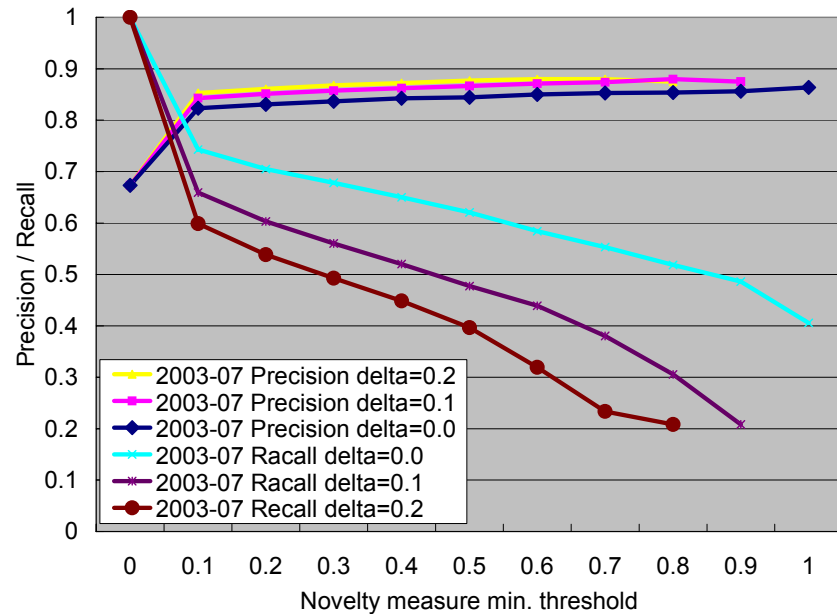
Precision and Recall (1/2)

- Positive δ gives 80% to 90% precision in all snapshots
- Precision jumps from the baseline when δ becomes positive, then gradually increases
- Positive delta values give slightly better precision



Precision and Recall (2/2)

- Recall drops according to the distribution of novelty measure
- Positive delta values decrease the recall



Guideline for Selecting Parameters

- When higher precision is required
 - $0 < \beta < 0.2$
 - Higher β
- When higher recall is required
 - $\beta = 0$
 - Small positive β

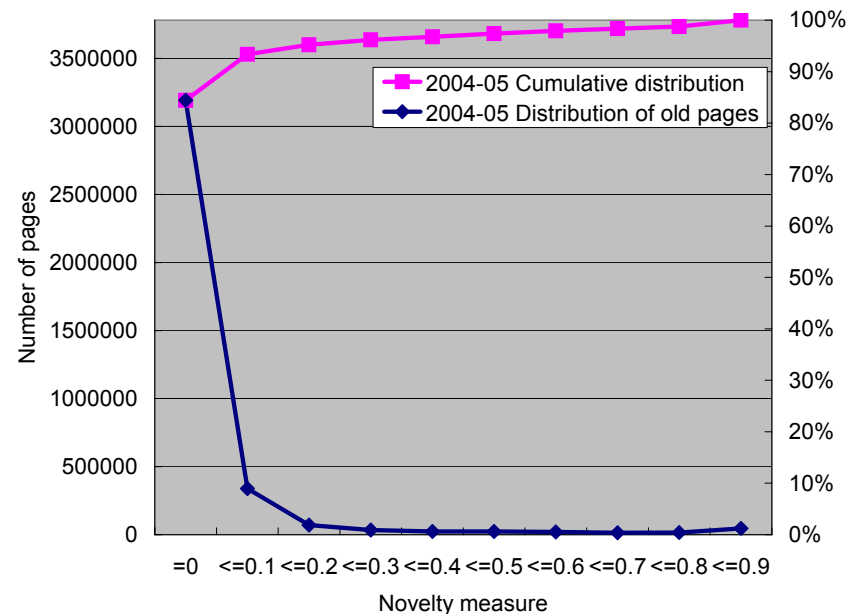
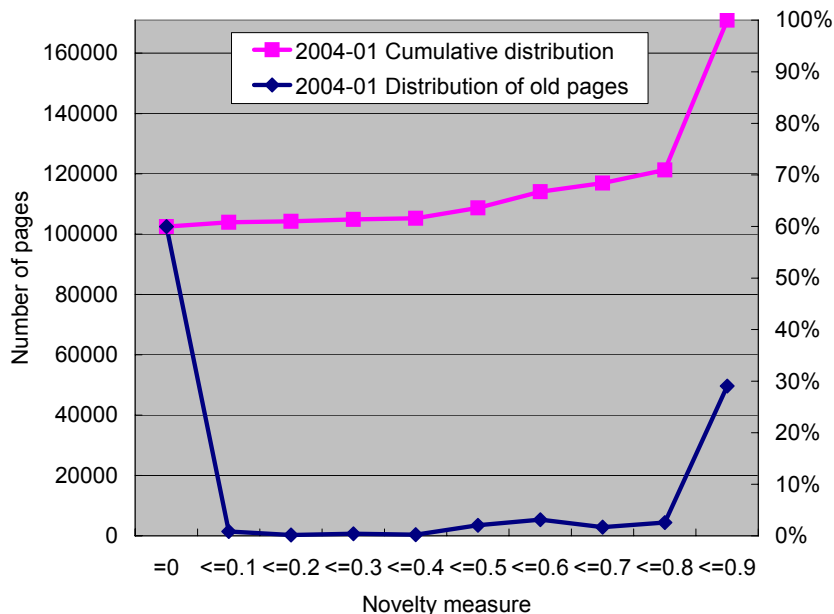
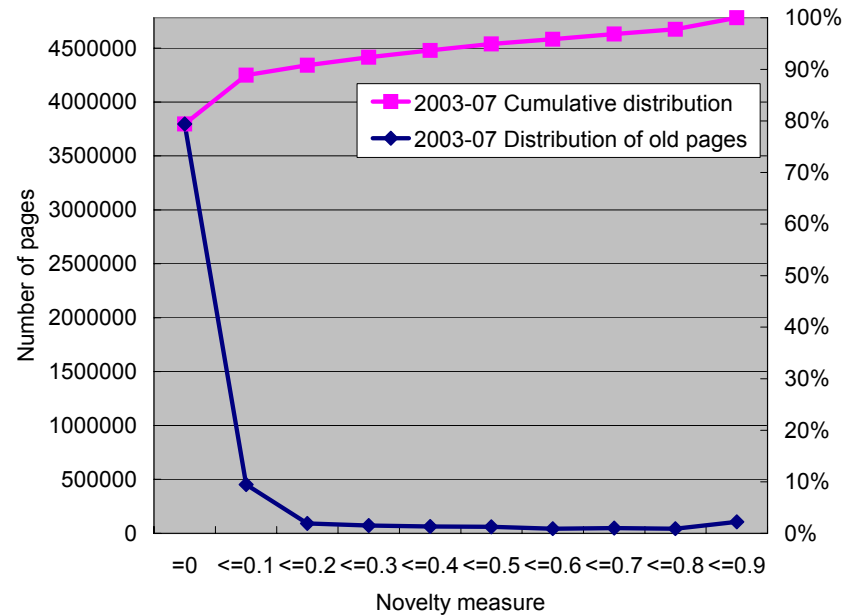
Miss Rate

- Fraction of pages miss-judged to be novel
 - Use a set of old pages as a golden set
 - Last-Modified time $< t - 1$
 - Check how many pages are assigned positive ***N*** values

Time	# old pages in $U(t)$	$ U(t) $
Jul 2003	4.8M	25M
Jan 2004	0.17M	6M
May 2004	3.8M	30M

Miss Rate

- Old pages tend to be assigned low N values
- In Jul 2003 and May 2004
 - Miss rate . 20% ($0 < N$)
 - Miss rate . 10% ($0.1 < N$)
- In 2004, Miss rate . 40%
 - # old pages is only 3% of $U(t)$ in Jan 2004



Application

Web Archive Search Engine

- Text search on all archived pages
 - Results in each snapshot can be sorted by their relevancy and novelty
- Changes in the number of novel pages are shown as a graph
 - Old pages but include the keyword first at t
 - Newly crawled pages judged to be novel ($\leq N(p)$)
 - Uncertain pages ($N(p) = 0$)

Conclusions

- Novelty measure
 - The certainty that a newly crawled page is really new
- Novel pages can be extracted from a series of unstable snapshots
- Precision, recall, and miss rate are evaluated with a large Japanese Web archive
- Novelty measure can be applied to search engines for web archives