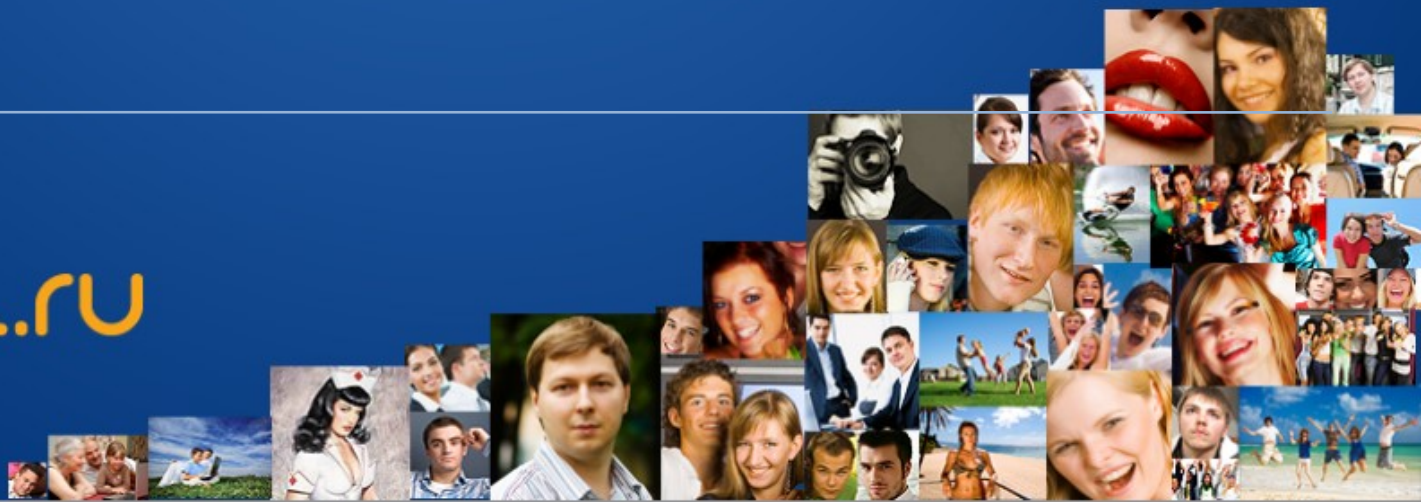
поиск@mail.ru®

The Crawler

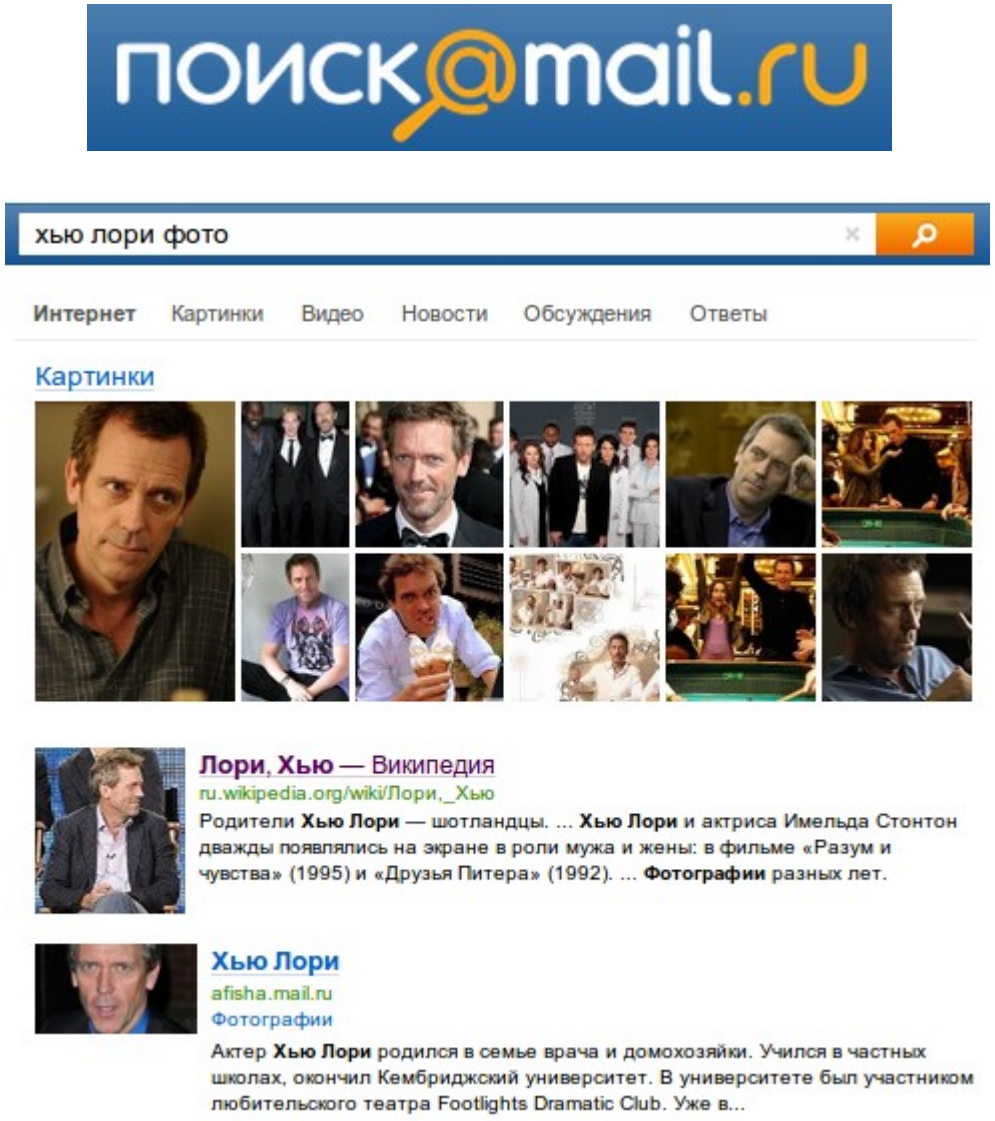
Alexey Voropaev

go.mail.ru

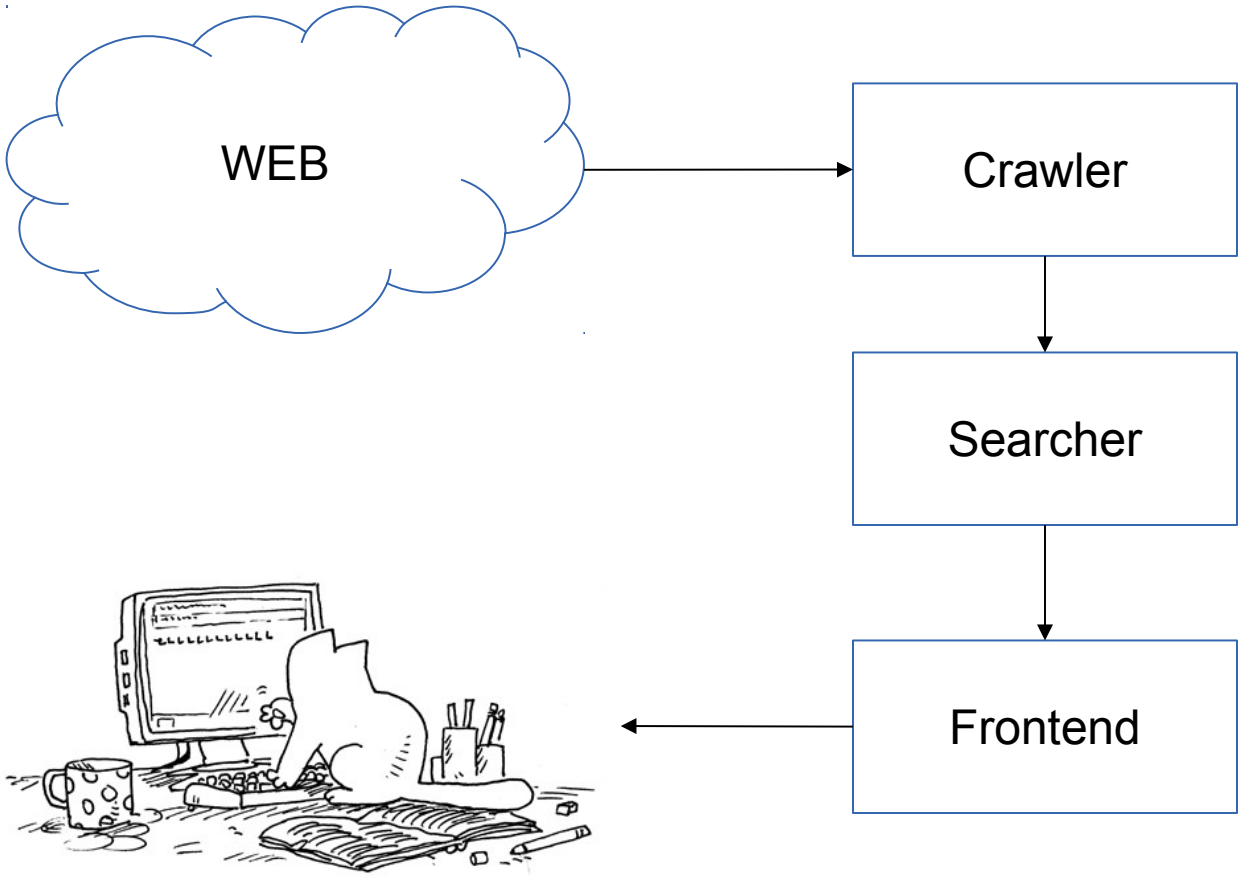


Mail.Ru

- http://go.mail.ru
- Web search for:
  - Russian
  - Ukrainian
  - Kazakh
- 9% of market share
- 100+ developers



SE Architecture



The problem

go.mail.ru

«Running a web crawler is a challenging task»

Sergey Brin and Lawrence Page, 1998

The Russian web

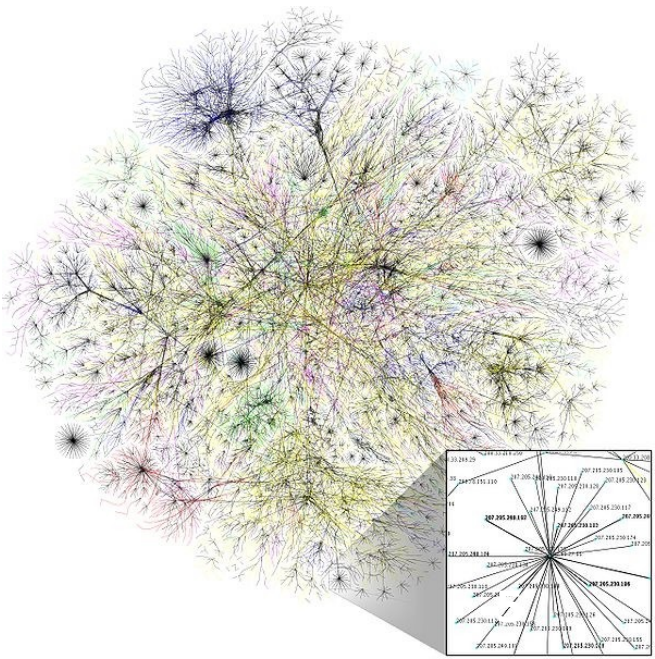
go.mail.ru

The web:

- Web sites: ~ 10^7
- Web pages (sensible): ~ 10^{10}
- Urls: ∞
- 60% of pages are located on 10000 biget sites

Search engine:

- Limited index: ~ 10^9 pages
- Limited bandwith: max ~100 pps
- Limited ranking: ~ 10^5 ppq



Crawler overview

go.mail.ru

Scheduler (aka long-term scheduler)

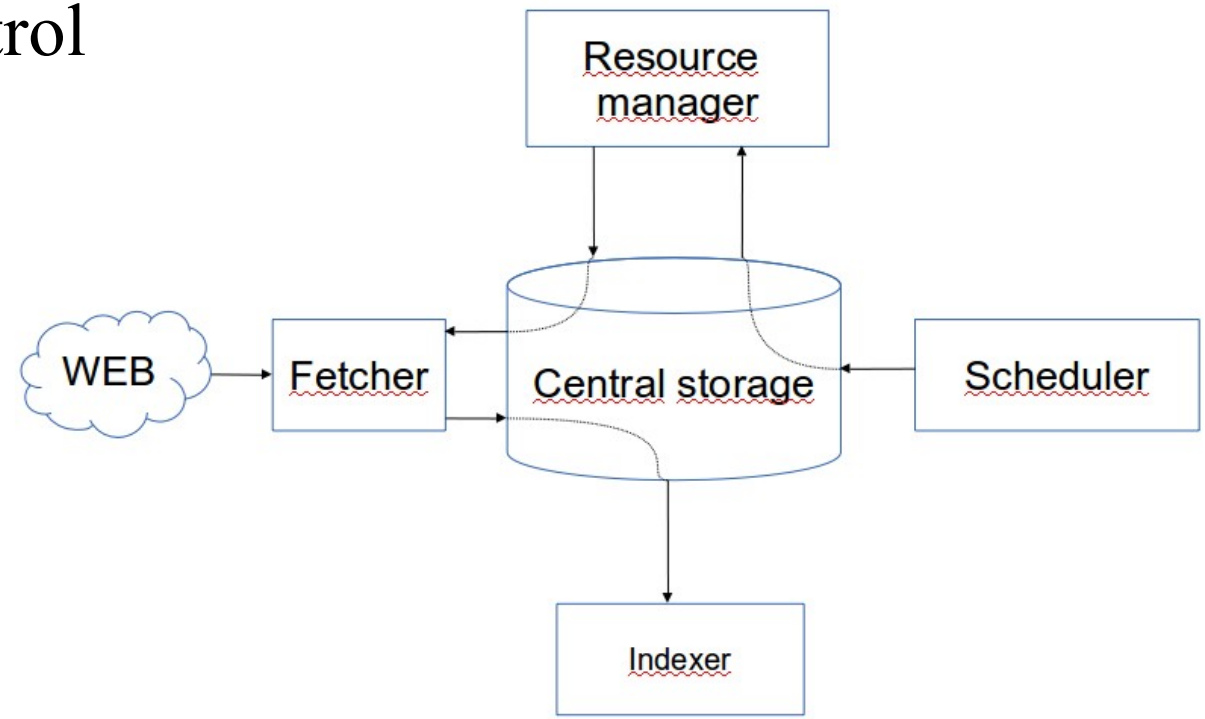
- Selection policy
- Refresh policy

Resource manager (aka short-term scheduler)

- Bandwidth control

Fetcher

- Just WGET



Crawler overview

go.mail.ru

Indexer

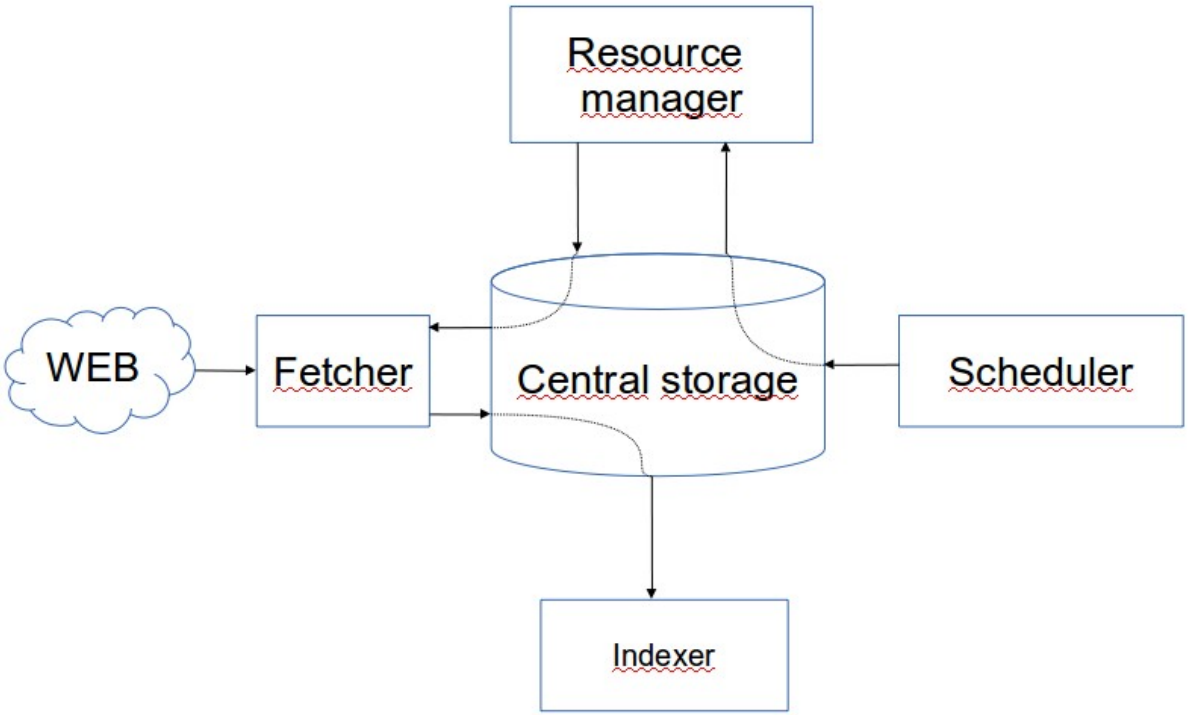
– Analyze of downloaded content

Auxiliary

– Logging

– Statistic

Central storage



Scheduler overview

go.mail.ru

Actually it is set of schedulers

- Web search
  - Download scheduler
  - Indexing scheduler
  - Discovery scheduler
  - Sandbox scheduler
  - Analyzer scheduler
- Image search
  - HTML scheduler
  - Image scheduler
- Experiments

Scheduler overview

go.mail.ru

Focused crawler – download pages with given topic.

We have crawling framework:

1. Define what is *interesting for crawling* pages
2. Find *initial seed*
3. Exploit *topical locality* to find other pages

What is "interesting for crawling" for Web-search?

– Page with Qlinks is interesting
– We can say nothing about page without Qlinks

Topical locality

go.mail.ru

– **Link based**

if A->B then B similar to A

– **URL based**

```
^http://aldebaran.ru/lov/[a-z]+/[a-z]+[0-9]+/$
*
```

```
^http://aldebaran.ru/kid/krapiv/krapiv[0-9]*$
*
```

```
^http://materinstvo.ru/$
+id+module
~module=articles
```

– **Content based**

```
title="Анкета заблокирована"
```

```
query="Что вы думаете об этом товаре?"
```

Small sampling analysis

go.mail.ru

For given:

α	fraction of urls from some group on a site
N	sampling size

Probability to find less then k urls:

$$p_{N,k}(\alpha) = \sum_{i=1}^{k} \binom{i}{N} \alpha^i (1-\alpha)^{(N-i)}$$

$$P_{1000,10}(0.01) \approx 0.58$$
$$P_{1000,10}(0.02) \approx 0.01$$
$$P_{1000,10}(0.03) \approx 2 \times 10^{-5}$$

Sekitei: algorithm

go.mail.ru

1. Sample N urls randomly

2. Generate features for each url
   - number of segments
   - list of parameters
   - <parameters=value>
   - ….

3. Select features with frequency αN

4. Do clustering:
   - Jaccard distance measure $K(a,b) = \frac{|A \cap B|}{|A \cup B|}$
   - Stack clustering

Sekitei: results

go.mail.ru

1. ^/wiki/File:[^/]+\.jpg$
/wiki/File:Spongilla_lacustris.jpg

2. ^/wiki/[^/]+\.jpg$
/wiki/Image:Deve.jpg

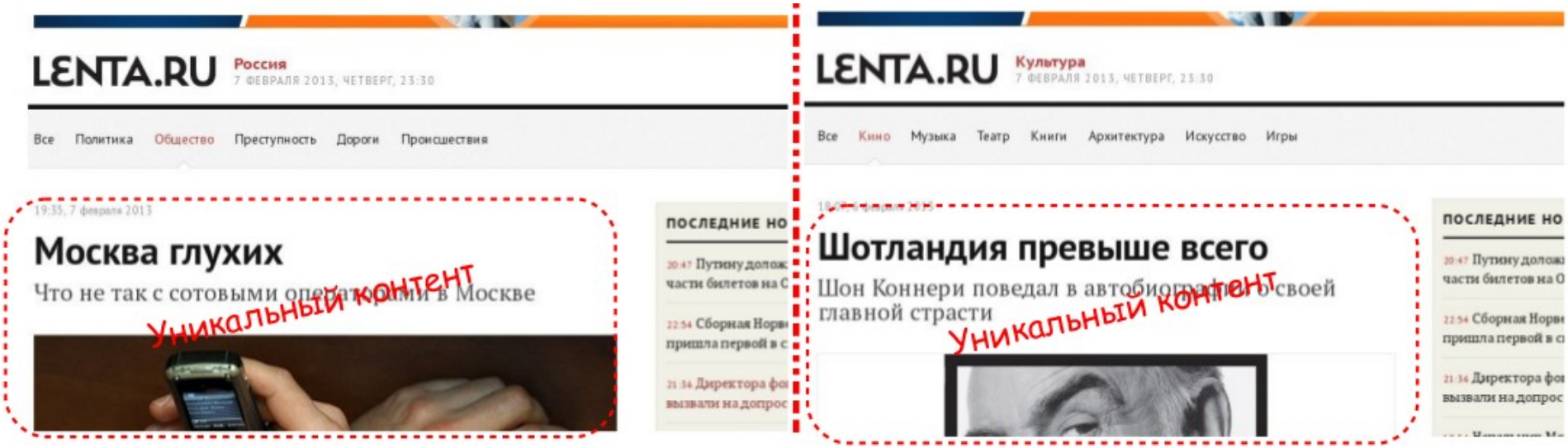
3. ^/wiki/Category:[^/]+$
/wiki/Category:Roman-era_historians

4. ^/wiki/Talk:[^/]+$
/wiki/Talk:North_Light

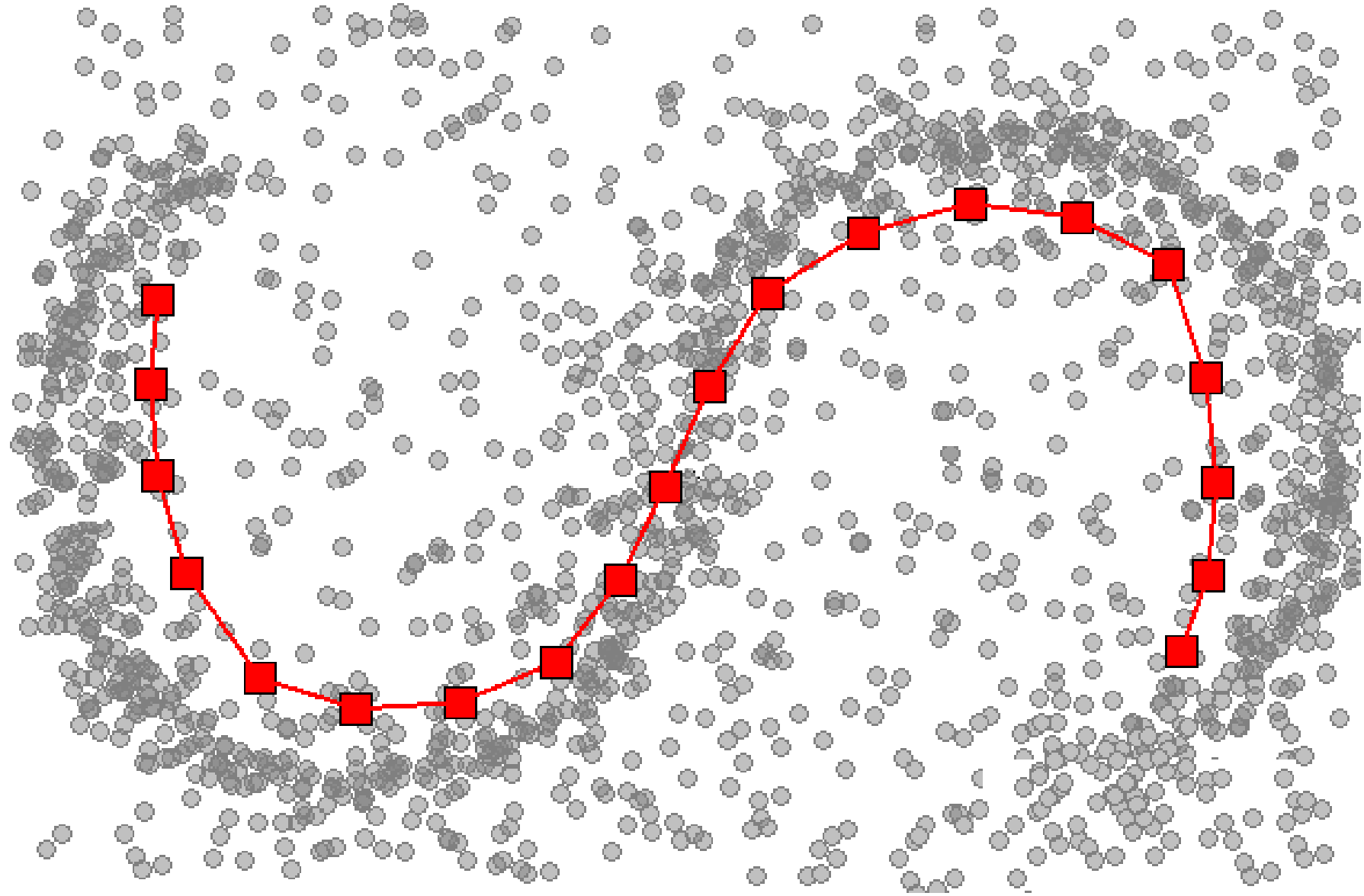
…

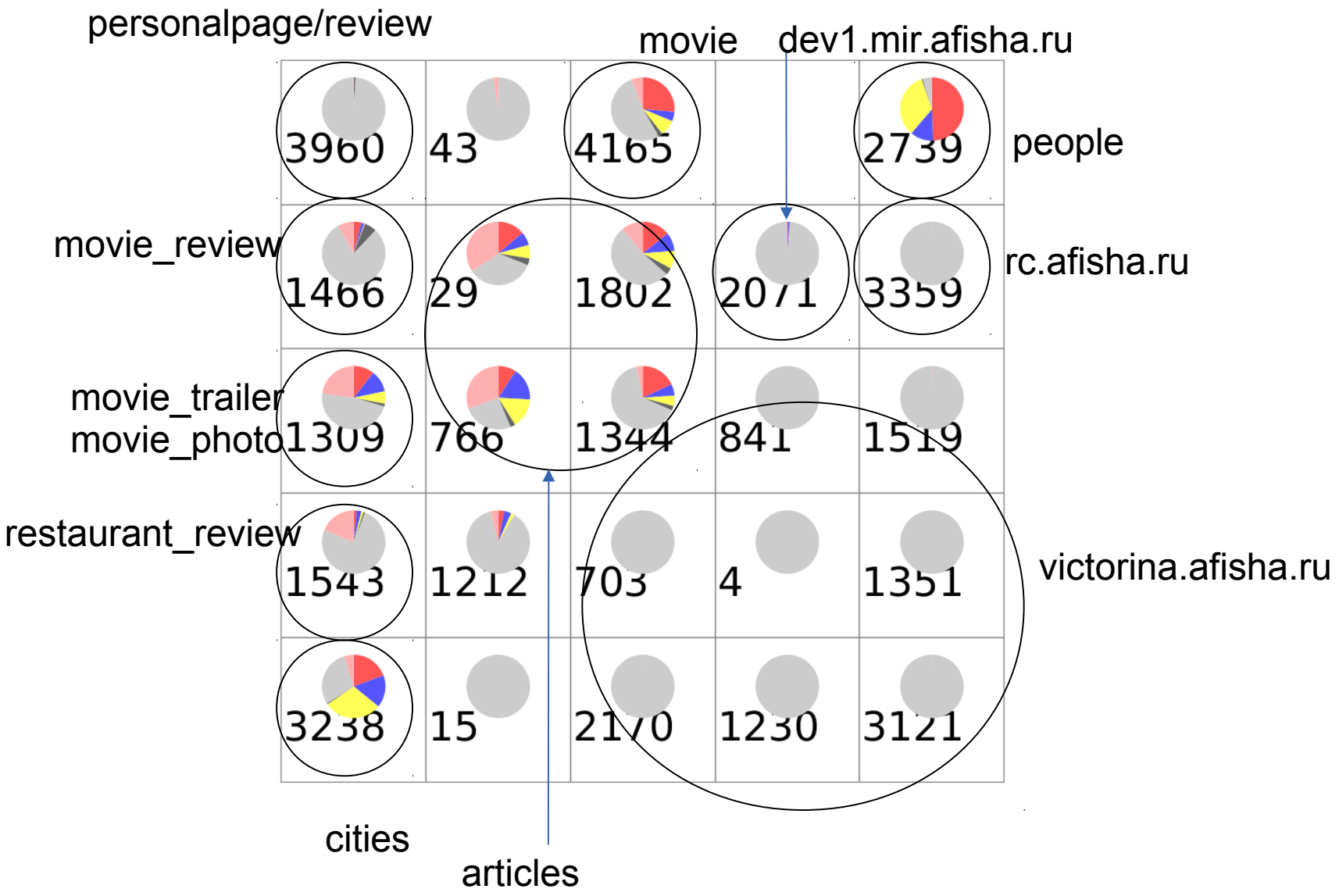
Sekitei: application

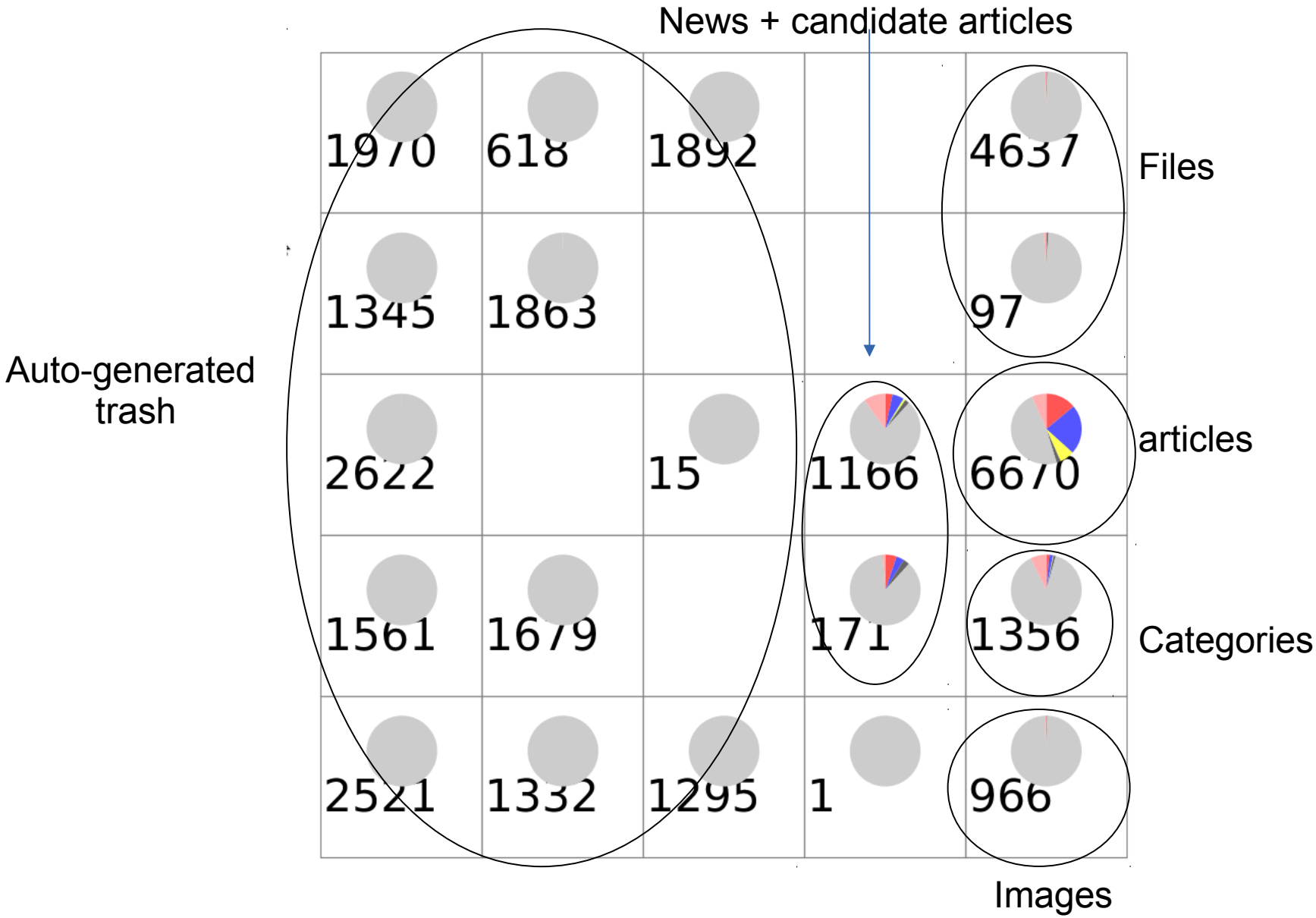
- 404, 500, timeout etc.
- spam, porno etc.
- duplicate detection
- news
- navigation removal
- etc.



SOM / Kohonen map




personalpage/review
movie
dev1.mir.afisha.ru
3960
43
4165
2739
people
movie_review
1466
29
1802
2071
3359
rc.afisha.ru
movie_trailer
movie_photo
1309
766
1344
841
1519
restaurant_review
1543
1212
703
4
1351
victorina.afisha.ru
3238
15
2170
1230
3121
cities
articles


News + candidate articles
1970
618
1892
4637
Files
1345
1863
97
Auto-generated trash
2622
15
1166
6670
articles
1561
1679
171
1356
Categories
2521
1332
1295
1
966
Images

Quoting

go.mail.ru

Goal: to build index with fixed size

Site:

- for hosting (about 1000 sites in ru) domain level 3
- otherwise domain level 2

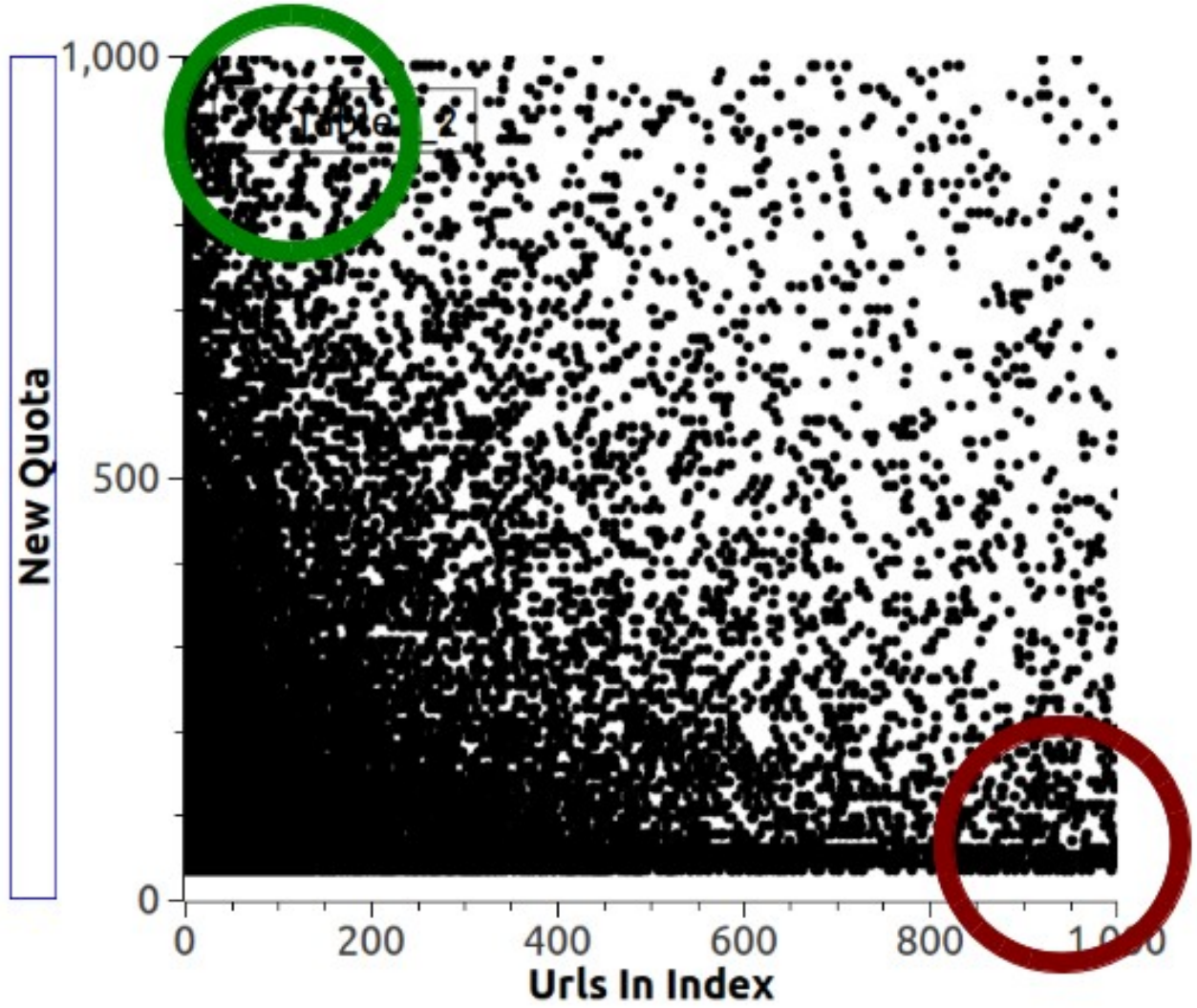
Quota:

- MIN_QUOTA ~ 100
- QUOTA = #Pages with Qlinks * MIN_QUOTA

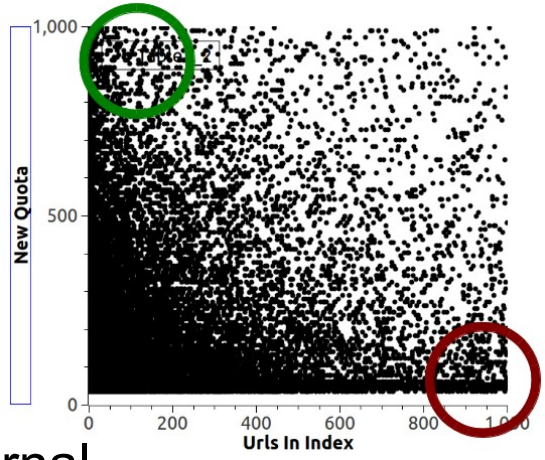
Diversity:

- We do quoting between stones inside of big sites

Quoting: compare with old algorithms



Sekitei: algorithm



Good

Blogs

blogspot/livejornal

Good sites

use4blog.com

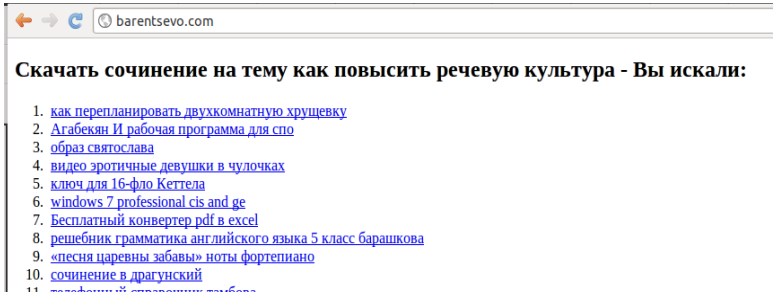
gagadget.com

Popular foreign sites

Robots, ban, etc.

Bad

Trash sites



Static rank

go.mail.ru

Features:

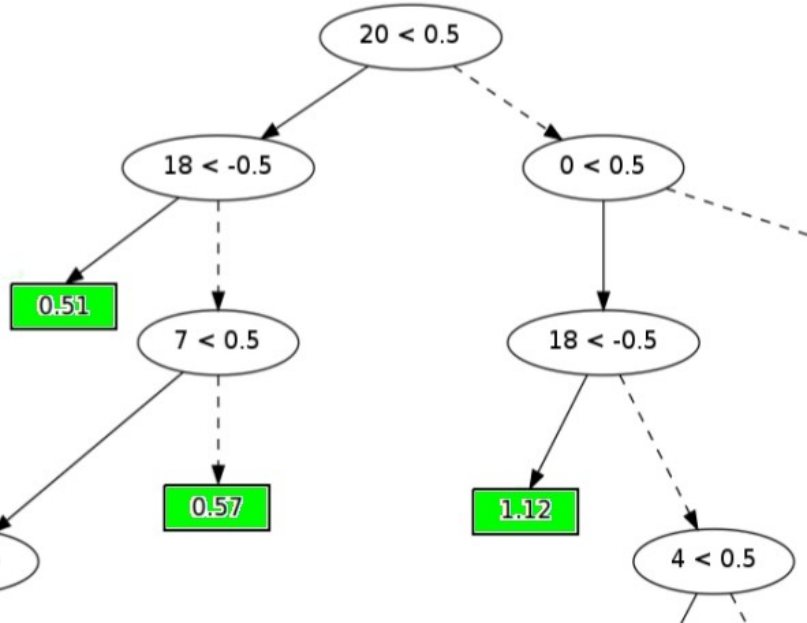
- Sekitei features
- Link features (Indegree, PR, etc.)
- Antispam features (e.g. #links from bad sites)

Model: gradient boosting decision trees

Target: predict #qlink for a page

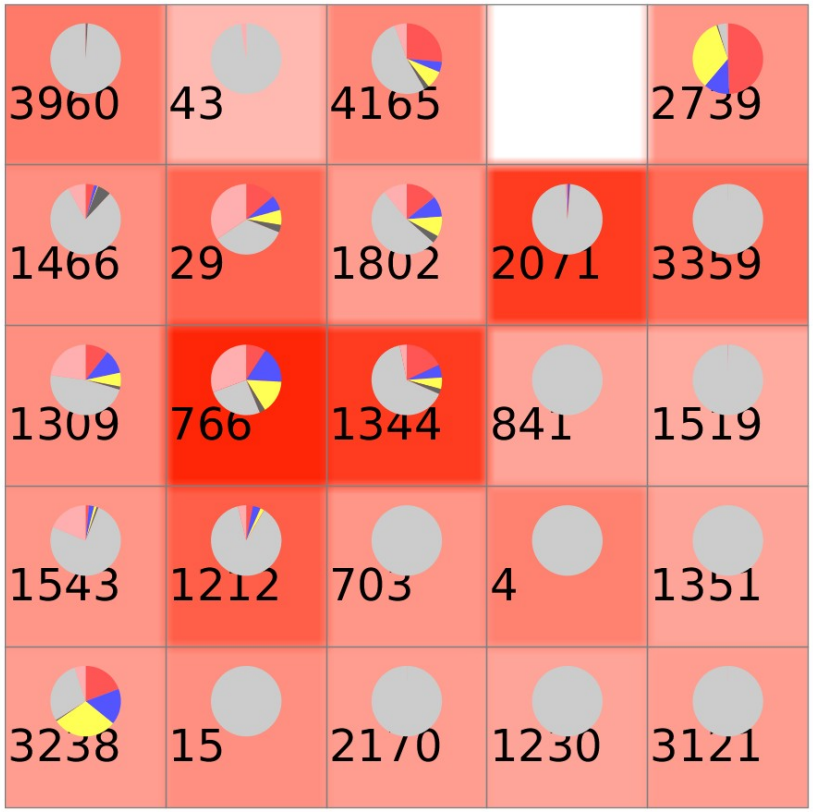
2 kind of models:

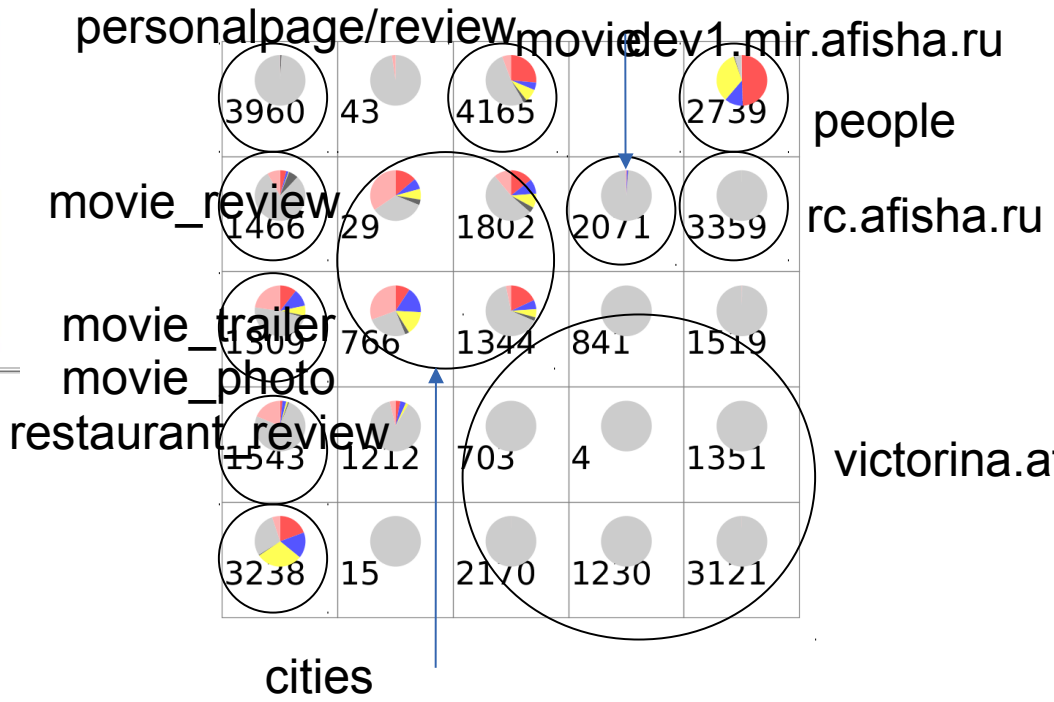
- individual (for big sites)
- one “wide” model for other

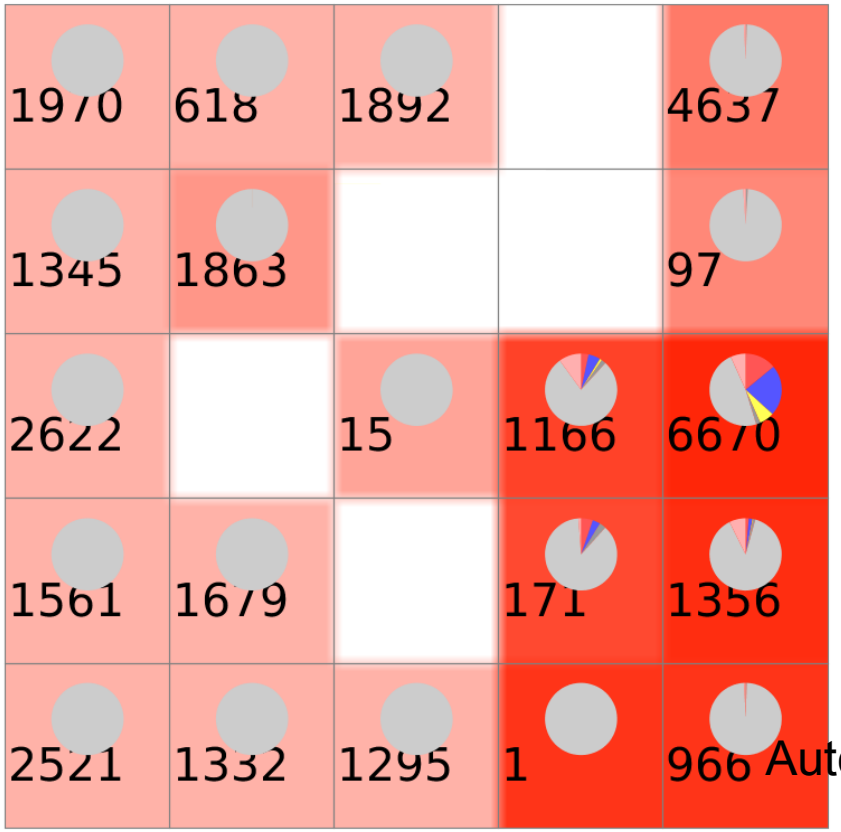


afisha.ru

go.mail.ru

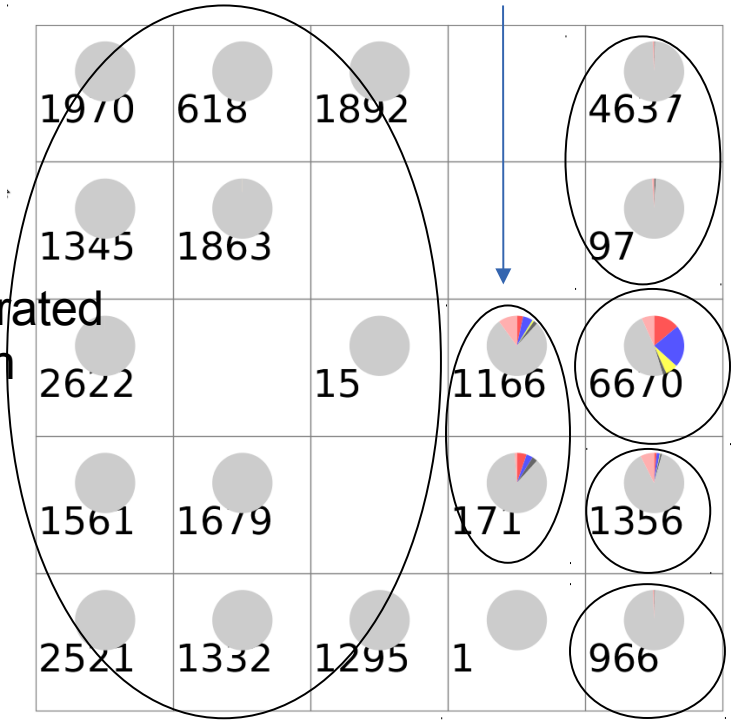






Auto-generated trash

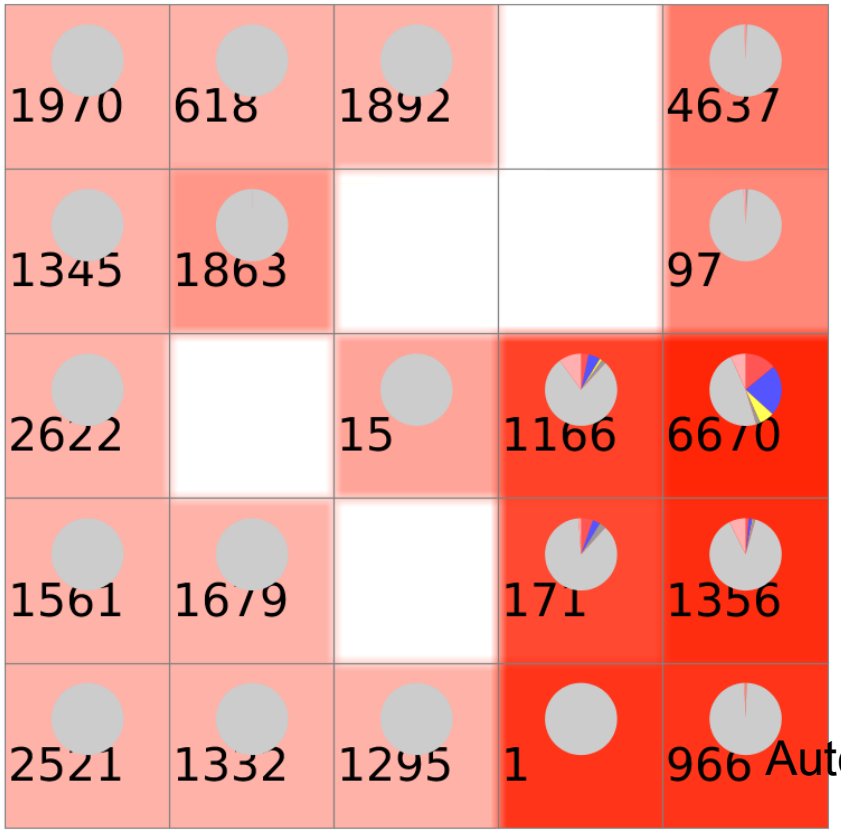
News + candidate articles



Files

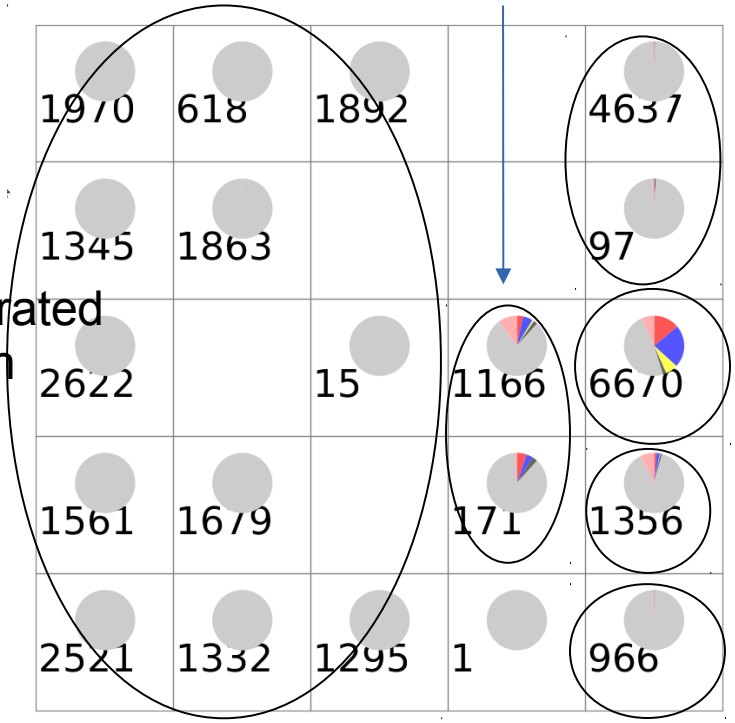
articles

Catego



Auto-generated trash

News + candidate articles

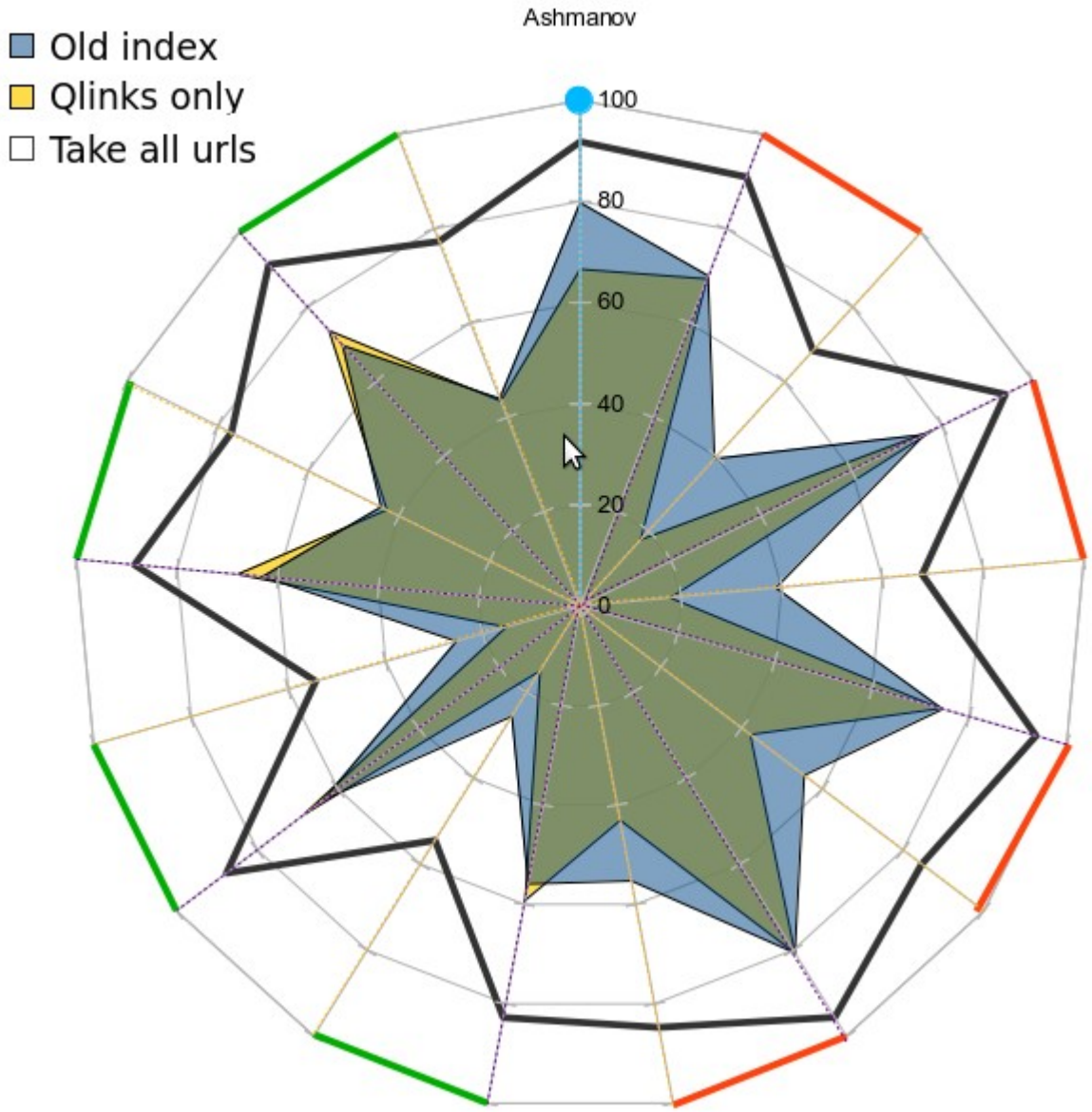


Files

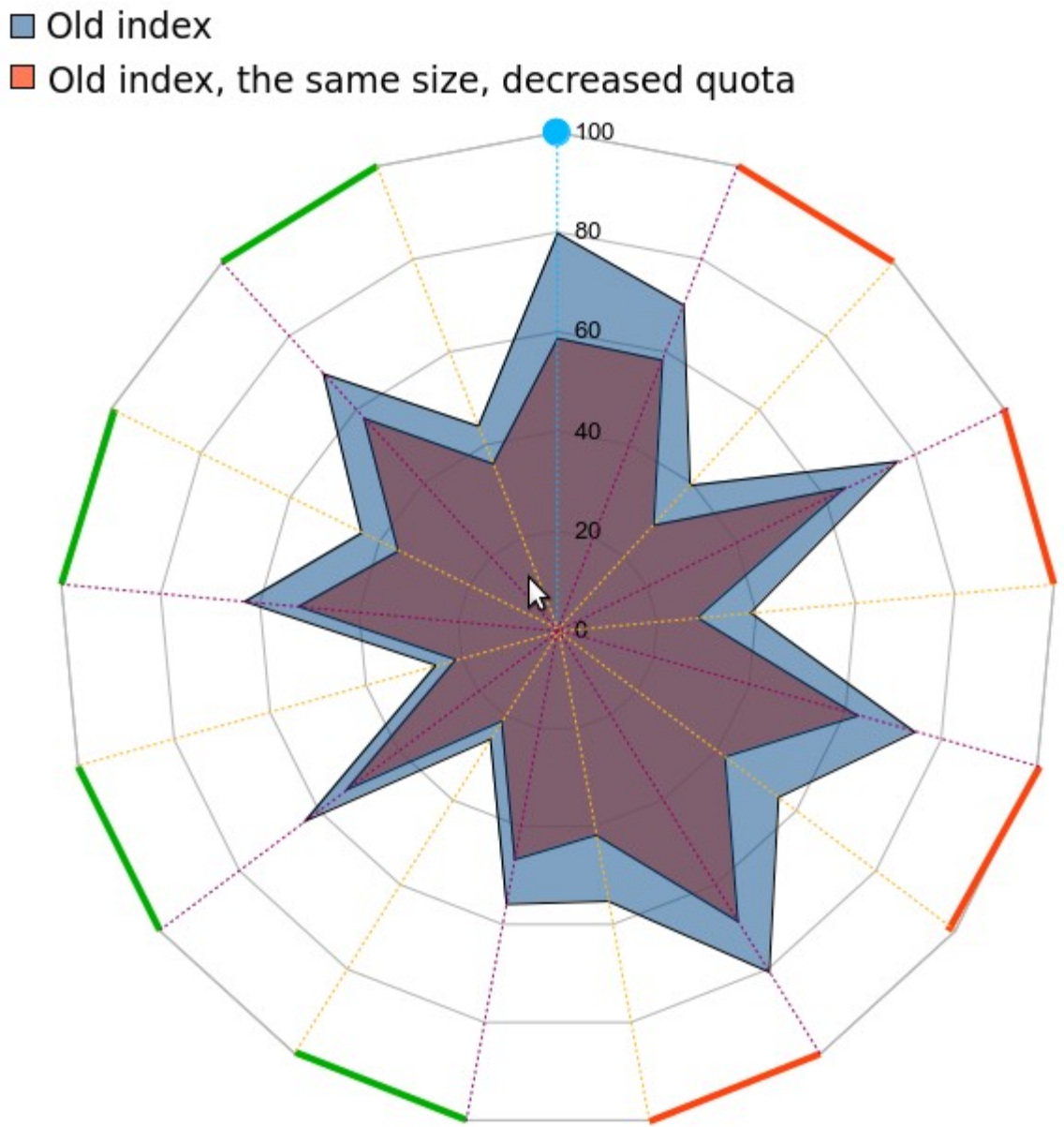
articles

Catego

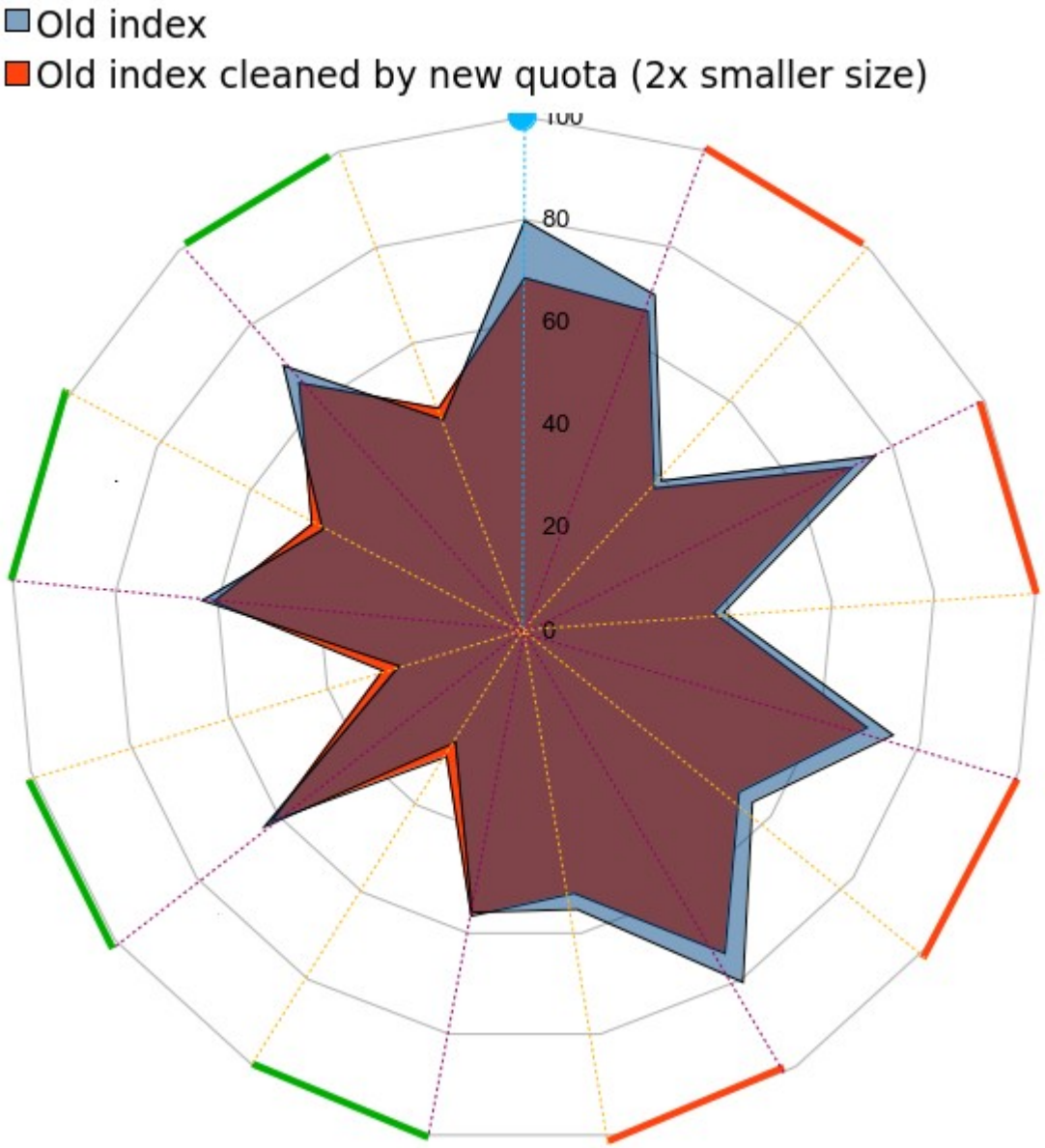
Index quality: baseline



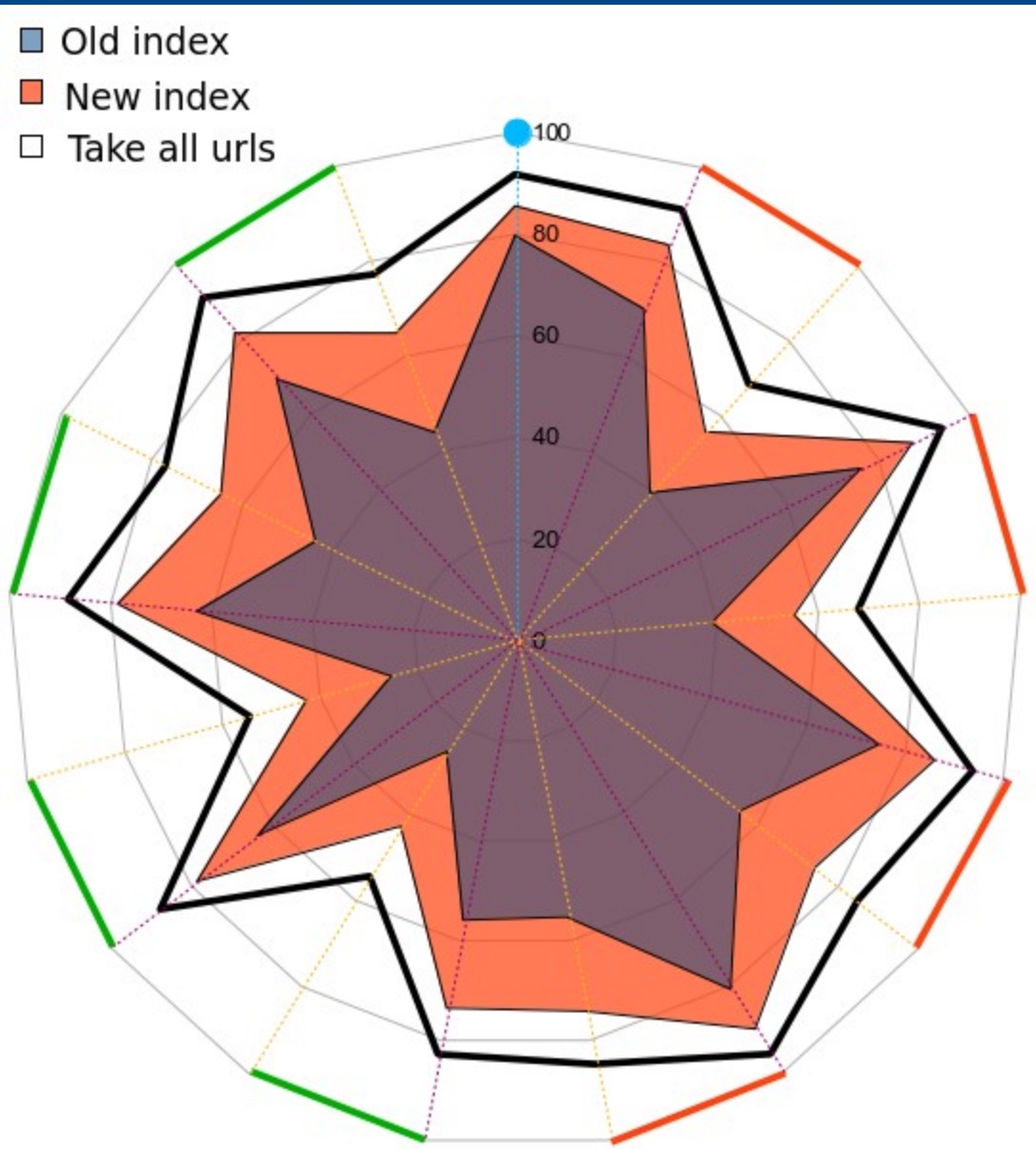
Index quality: old quota impact



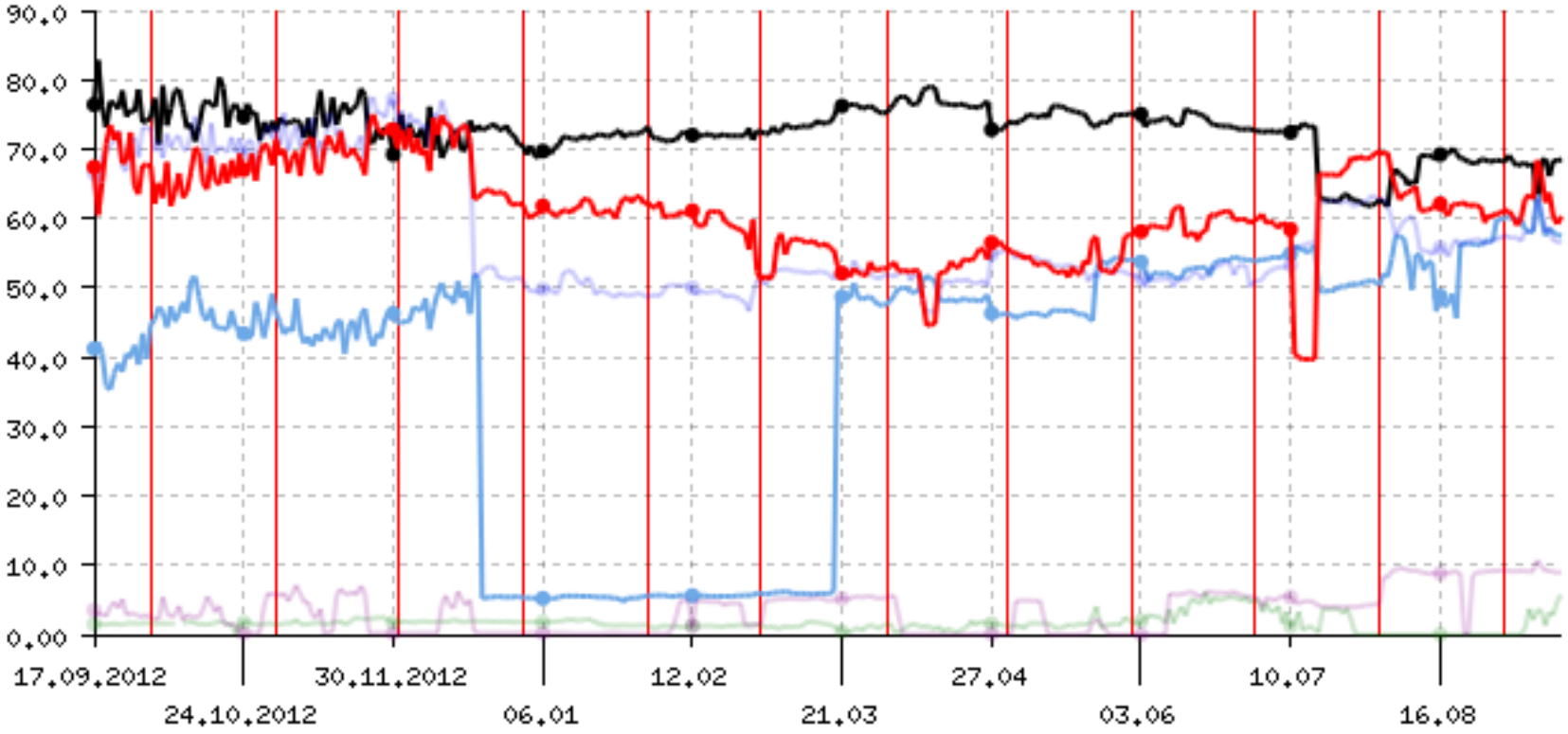
Index quality: new quota



Index quality: final



recall



Mail

Yandex

Google

Static rank

go.mail.ru

№	Запрос	Bing (5.7 +0.8)	Google (68.6 +0.4)	Mail (57.4 -0.3)	Rambler (56.4 -0.4)	Yahoo (9.0 +0.1)	Yandex (60.3 +1.1)
1	абузничество	0	1	4	1	0	1
2	авантюрствовать	4 +4	16	17	17	4	17
3	антибомонд	6	40	28	19	6	19
4	алтайминг	0 -1	3	28	25	1	26
5	артпоповый	0	36 -2	15	19	1	20

Pin-Up Went Down - 342 (2010) - MetalArea
metalarea.org › ... › Releases Area / Релизы › Archive ▾ Translate this page
Jul 9, 2010 - 30 posts - 16 authors
а альбом занятный. **артпоповый**, извините (с), такой. первый был концептуален, но я его переосмыслил и потер почти все, окромя:

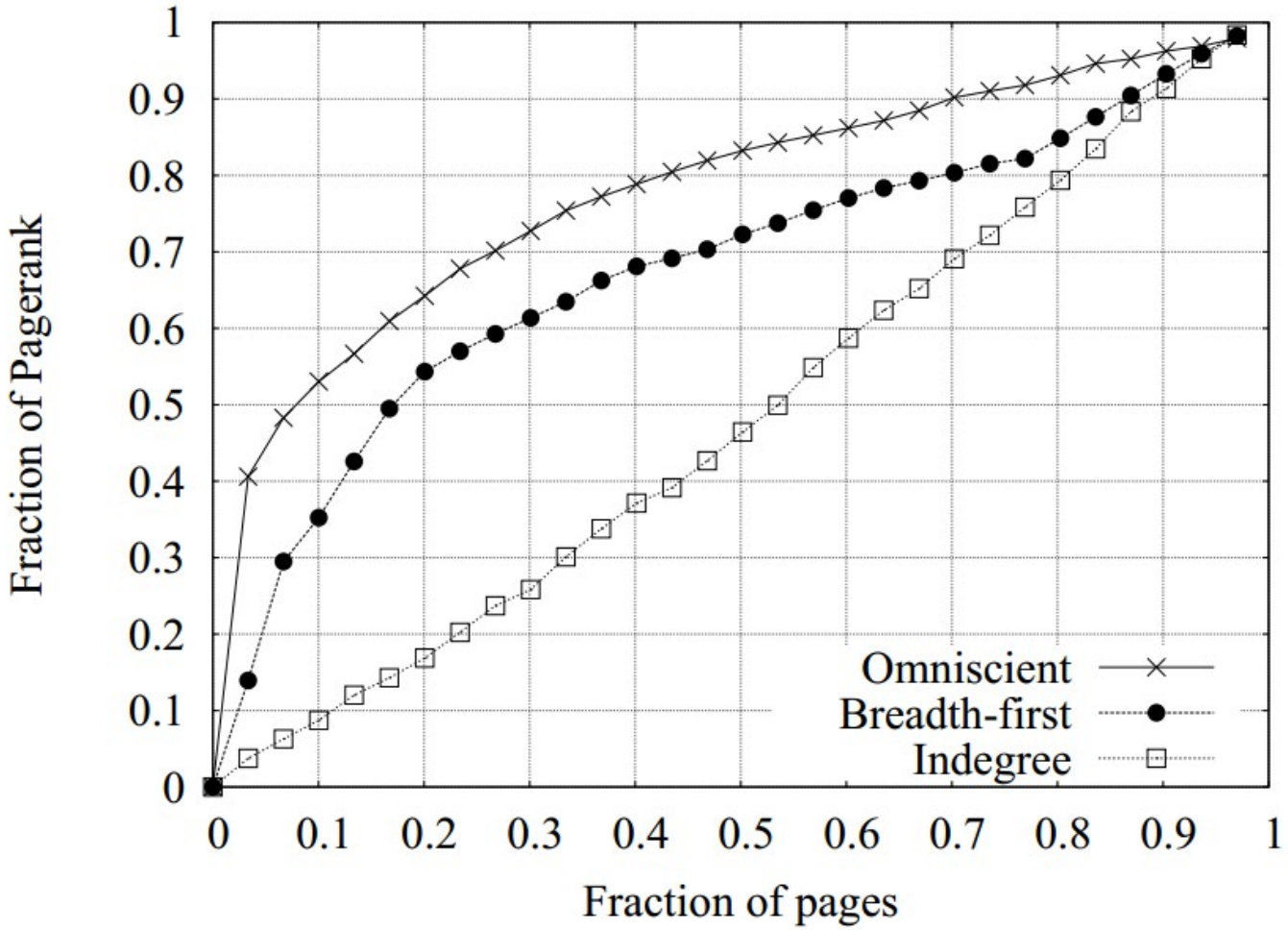
Pin-Up Went Down - 342 (2010) - MetalArea
metalarea.org › ... › Releases Area / Релизы › Archive ▾ Translate this page
Jul 9, 2010 - 30 posts - 16 authors
а альбом занятный. **артпоповый**, извините (с), такой. первый был концептуален, но я его переосмыслил и потер почти все, окромя:

Pin-Up Went Down - 342 (2010) - MetalArea
metalarea.org › ... › Releases Area / Релизы › Archive ▾ Translate this page
Jul 9, 2010 - 30 posts - 16 authors
а альбом занятный. **артпоповый**, извините (с), такой. первый был концептуален, но я его переосмыслил и потер почти все, окромя:

Google Russia

Static rank

go.mail.ru



Ricardo Baeza-Yates. Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition), 2011

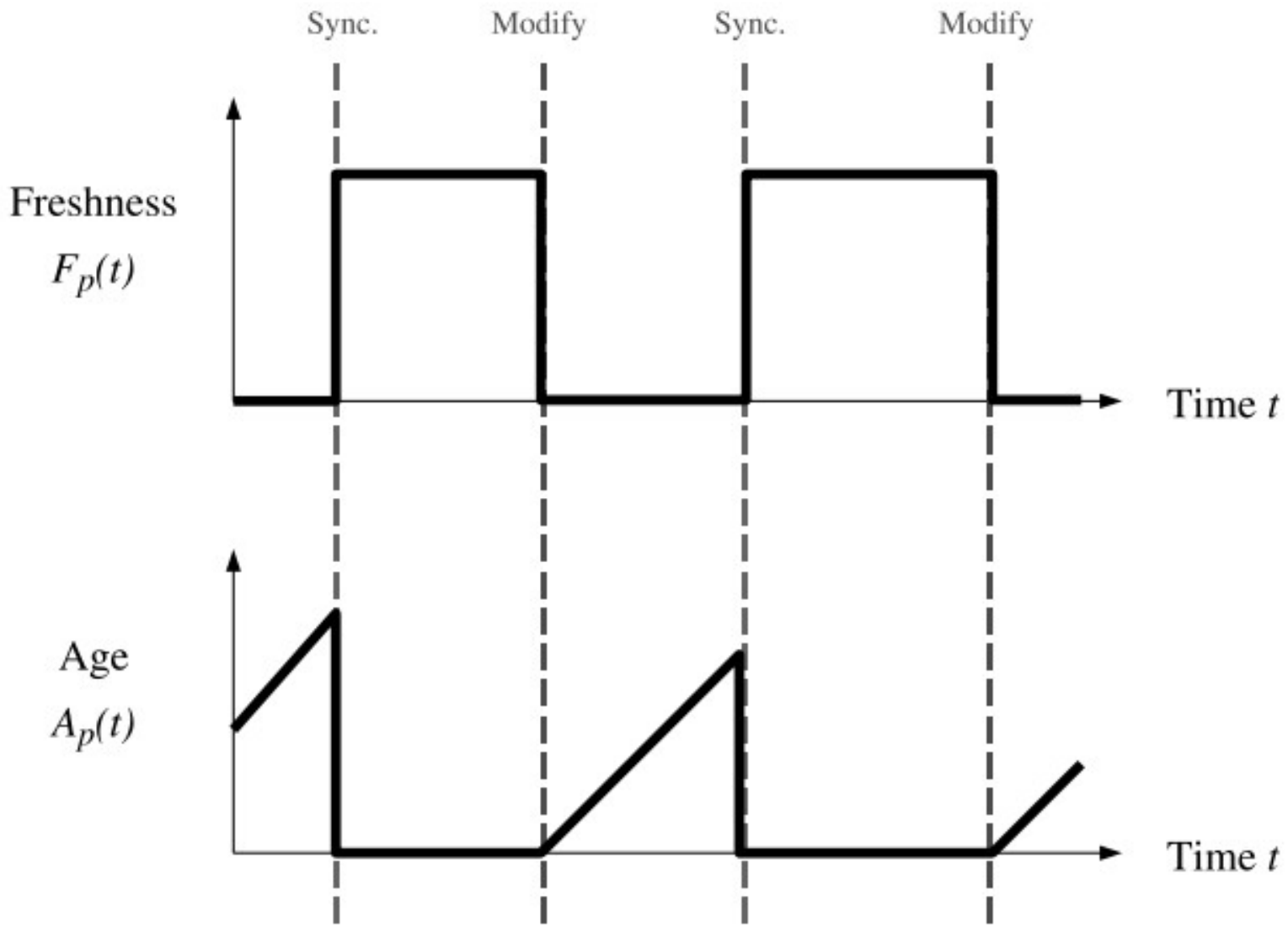
Revisit policy

go.mail.ru

Events

- **Creation**
  - We use special discovery scheduler
- **Deletion**
  - Permanent vs temporal
  - Special issue with ban/robots
- **Update**
  - How to understand, that page was changed?
  - How to predict update?

Update prediction



Ricardo Baeza-Yates. Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition), 2011

Central storage

go.mail.ru

- Distributed file system (HDFS)
- MapReduce framework
- Hbase: no-sql database

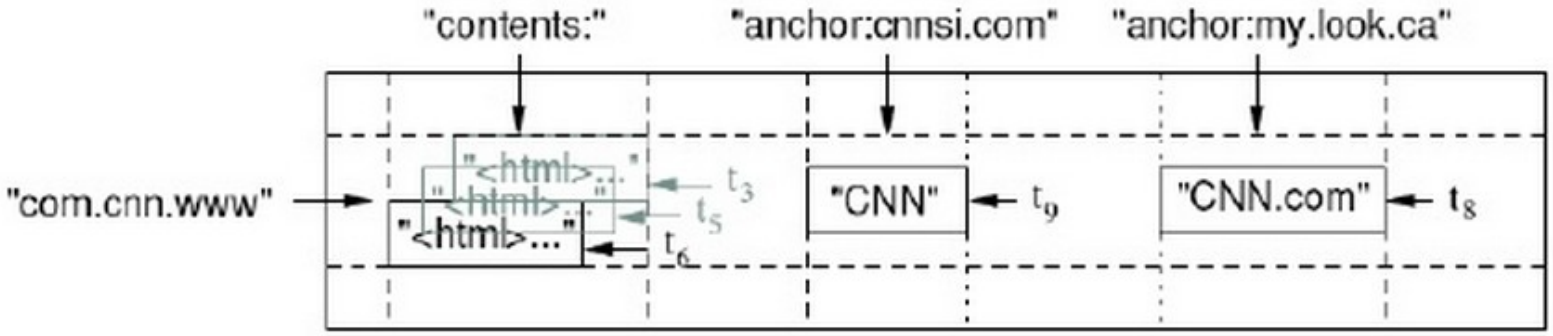
Hbase

go.mail.ru

APACHE HBASE

- Doesn't support a full relational data model
- Multi-dimensional sorted map
- Indexed by a row, column, timestamp

$$(row: string, column: string, time: \text{int } 64) \rightarrow string$$



- Column-oriented storage

- Most queries only involve a few columns out of many, so greatly reduces I/O.

Thank you!

Reference:

– **Ricardo Baeza-Yates**. Modern Information Retrieval:
The Concepts and Technology behind Search (2nd Edition), 2011