Web search

Serge Abiteboul

INRIA & ENS Cachan

Web indexing

Indexing: scaling

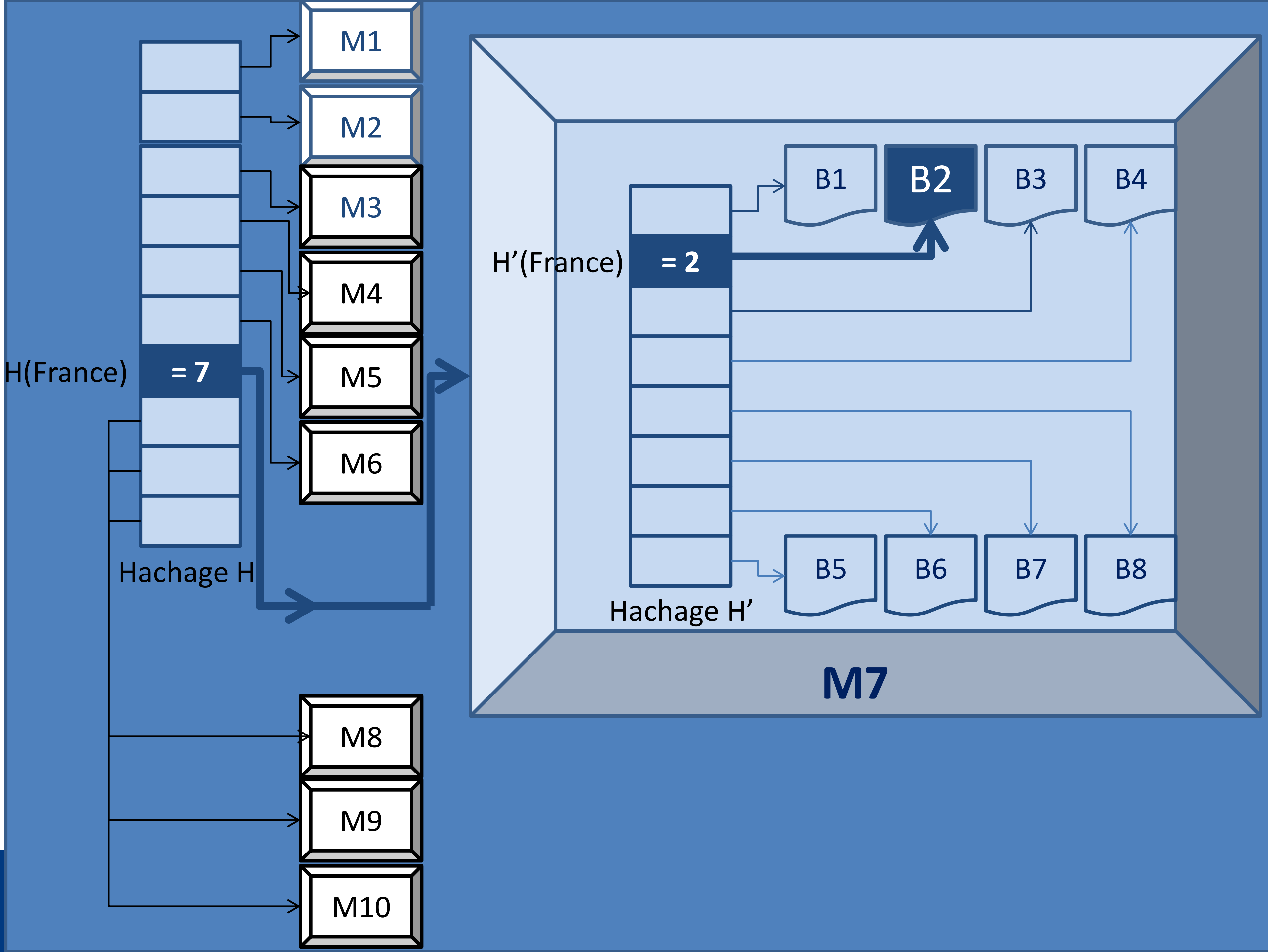
If the number of indexed pages grows, the server needs more storage to keep the index, and each query is becoming more expensive to evaluate

If the number of users grows, the server receives more requests.

In both cases, the server is quickly overwhelmed.

Solution

- the technique of hashing and
- parallelism.


M1
M2
M3
M4
M5
M6
H(France)
= 7
Hachage H
M8
M9
M10
B1
B2
B3
B4
H'(France)
= 2
B5
B6
B7
B8
Hachage H'
M7

Google

The scale

- Billions of pages
- The size of the index is roughly that of the indexed text
- Dozains of billions of queries per month

A query should require zero or very few disk accesses

Farms everywhere in the world with thousands of machines

Main issue: how to select the pages proposed in first page of answer?

Distributed PageRank [WWW03]

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY – ÎLE-DE-FRANCE

On-line vs. off-line computation

Off-line algorithm

- Crawls the Web and builds a Link-matrix
- Stores the link matrix and update it – very expensive
- Starts an off-line computation on a frozen link matrix

On-line Page Importance Computation

- Does not require storing the link matrix
- Works continuously together with crawling
- Keeps improving and updating the estimate

Static Graphs: OPIC

We assign to each page a small amount of cash

When a page is read, its cash is distributed between its children

The total of cash in all pages does not change

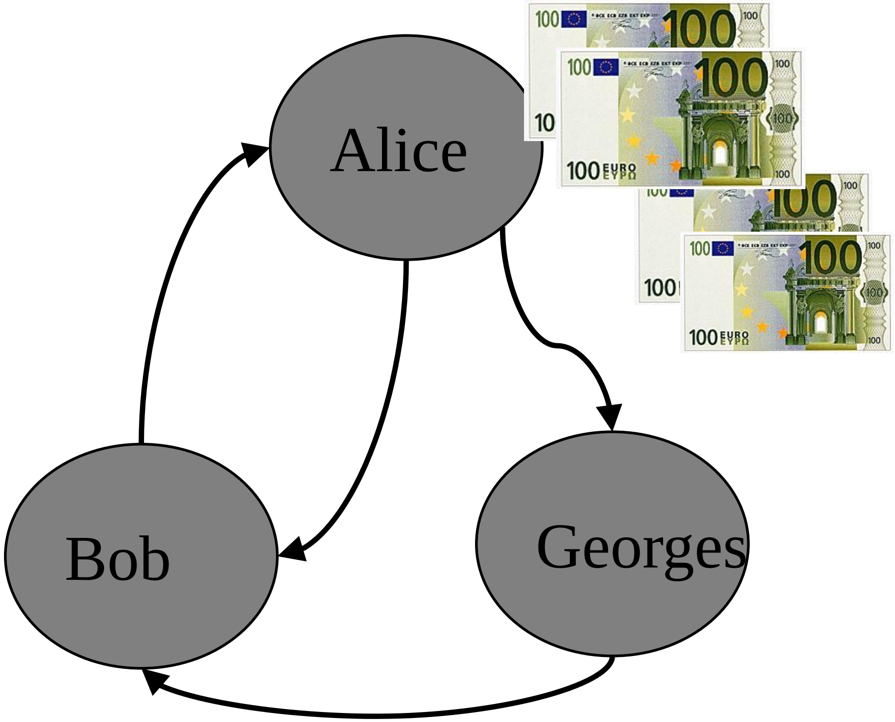
The page importance for a given page is computed using the history of cash of that page

Example

Small Web of 3 pages

Alice has all the cash to start

Importance independent of the original position



What happened?

Cash-Game History:

Alice received	600	(200+400)
Bob received	600	(200+100+300)
Georges received	300	(200+100)

Solution:

I(Alice)	= 40%
I(Bob)	= 40%
I(Georges)	= 20%

It is the fixpoint

I(page) = History(page)/ Sum(Histories)

Two issues

The Web is a changing graph: Adaptive Algorithm

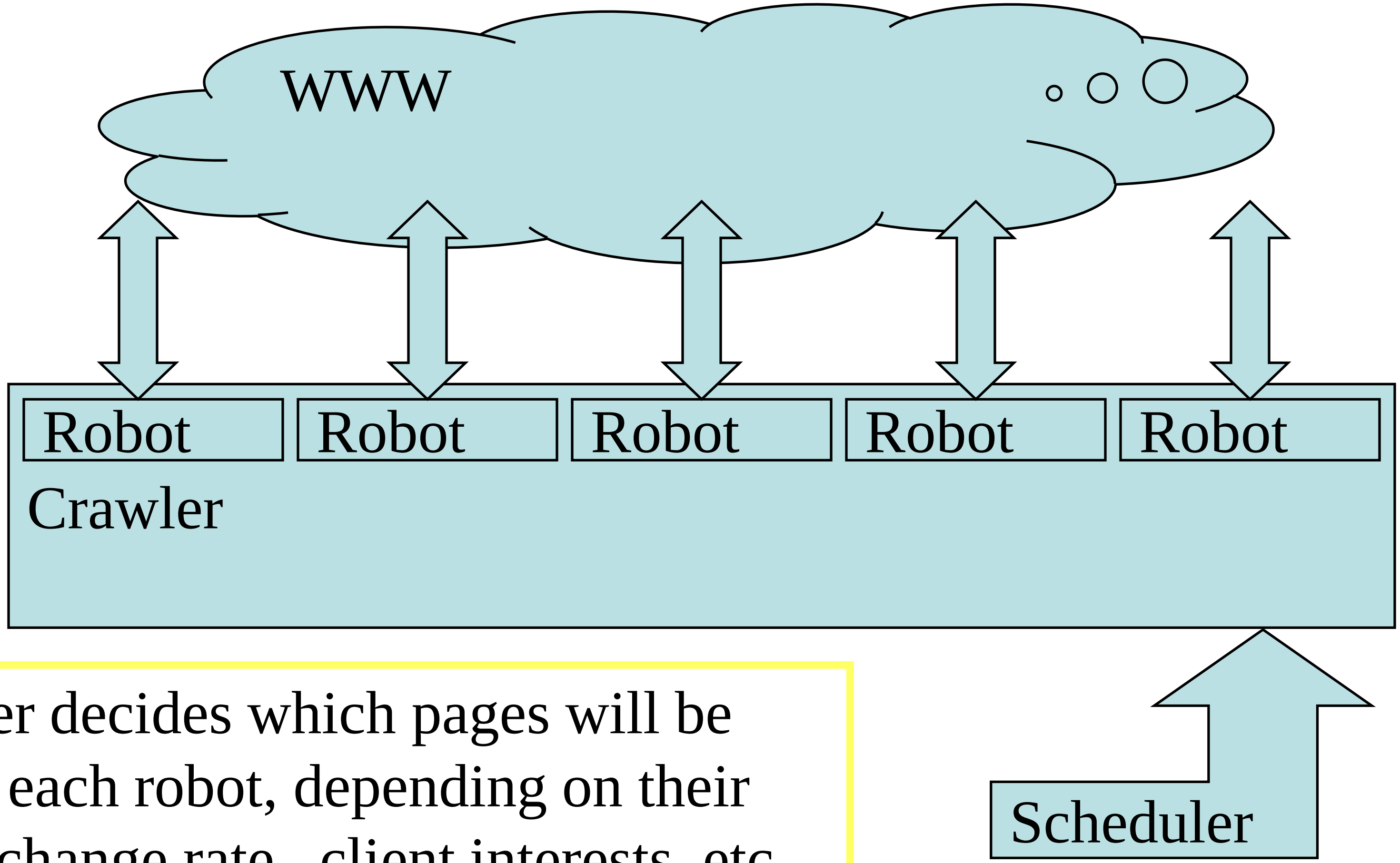
- The Web changes continuously, so does the importance of pages
- Consider only the recent part of the cash history for each page
- Consider what happened in a “time window”

Distributed OPIC

- Issue: a peer can cheat to increase the importance of a “friend”
- Give him all the cash from the pages he crawls
- “Coloring technique” to fix that – in charge of distributing cash to a small random number of pages

Experiments

Overview of Crawler



The scheduler decides which pages will be read next by each robot, depending on their importance, change rate, client interests, etc.

Experiments

Experiments were conducted using the crawlers of INRIA & Xyleme

- Cluster of 8 PCs/Linux with 1.5Gb of memory each
- Code in C++, communications use Corba
- Each crawler: about 4 million pages per day – about 100 millions total
- A difficulty: map between URL and integers

Experiments lasted for several months - one billion URLs

The crawling strategy: “greedy” read the pages with maximum cash – optimize the speed of convergence

After several trials, we set the window size to 3 months

60% pages were never read

Experiments (2)

Crawler

- Up to 100 robots running simultaneously on a single PC
- Average of 50 pages/seconds on an (old)PC (4 millions/day)
- Limiting factor is the number of random disk access

Performance and Politeness

- Pages are grouped by domain to minimize the cost of DNS (Domain Name Server) resolution
- To avoid rapid firing, we maintain a large number of accessible sites in memory (1 million domains).

Knowledge about visited pages: 100 million pages in main memory

- For each page, the exact disk location of the info structure (4 bytes) + a counter that we use for page rank and for the crawling strategy
- One disk access per page that is read

Negative information

Motivation

Negative references on the Web

- In recent news: A service company ranked first with mostly negative references

Opinions and recommendations

- Did Elvis actually died?
- What is the best pizza in La Plagne?

In social networks

- Where is Alice today?
- Who is her boyfriend?
- When is her birthday?
- What would she like as a gift?

Some work on corroboration [WSDM10]

Alban Galland, Amélie Marian et Pierre Senellart

One often finds contradictions/inconsistencies on the Web

Use voting: gives already good results

Alternative

- Assess the truth using voting
- Assess expertise of participants based on the truth and what they voted
- Access again the truth using voting biased by expertise of voters
- Access expertise based on the new truth

. . . until fixpoint

Get better results that voting in some cases

- When there are participants with different level of expertise

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE