Engineering at Scale
THE CHALLENGES OF PREDICTING QUERIES IN WEB SEARCH ENGINES
Paul Baecke
restaurants paddington
restaurants paddington station
restaurants paddington basin
restaurants paddington central
restaurants paddington area
restaurants paddington street london
restaurants paddington brisbane
restaurants paddington london trip
Pétrus - Gordon Ramsay Restaurants
Ad www.gordonramsayrestaurants.com/Pétrus
Style & Modern European Cuisine Incredibly Presented. Book Now!

Ad parkgrandlondon.com
Book Now Pay Later - Stay & Dine Package - Celebration Package

The 10 Best Restaurants Near Paddington Station, London ...
www.tripadvisor.co.uk - England - London - London Restaurants
Restaurants near Paddington Station. London on TripAdvisor: Find traveller reviews and candid photos of dining near Paddington Station in London, United Kingdom.

Restaurants near Paddington Tube Station | Squaremeal
www.squaremeal.co.uk/restaurants/station/paddington-tube-station
We've found 160 Restaurants near Paddington Tube Station. Click here to read Square Meal's independent reviews, check out restaurant menus and make reservation

Paddington Restaurants | OpenTable
https://www.opentable.co.uk/london/paddington-restaurants
Find Paddington restaurants in the West London area and other locations such as Kensington, Notting Hill, Hammersmith, and more. Make restaurant reservations and ...

Paddington Taxis
Ad www.allpaddingtonaxis.co.uk
Call Now - Prompt Local Tax Service. Est Local Company in Paddington
Learn More on About.com
Ad - About.com/Experts
Related Articles on Trending Topics 85+ Million Visitors - Search Now
Introduction

How is what we do ‘Extreme Computing’?
What is the product
Complexity online
Complexity offline
Complexity of systems
Some examples
Some numbers (online)

10^{12} \text{ requests served per year}

10^{16} \text{ bytes of data logged per year}

10^{14} \text{ ms of CPU time used per year}

Average QPS

> 100k

Data

> 1 GB

CPU

> 5 million ms

Thousands \textit{Years of CPU Time per year!}
More numbers (offline)

Some data is refreshed 12 times per day
All data is updated daily
Models updated weekly
Data scientists run 100s of experiments per week
Availability goal is > 99.995% uptime
Latency goal is < 50ms average
Why this matters

At this scale, every engineering decision matters

There is a deep focus on efficient data structures and algorithms

Every ms of CPU time saved, every byte of storage optimized:

Saves money & time

Allows for better experiences to be built

Makes users happier

Keeps our engineers on the cutting edge of research and best practices
Bing & Autosuggest Infrastructure

WHAT IT TAKES TO SERVE BING & KEEP LIVE SITE HEALTHY
Bing Usage

>500 Million Bing Users
In 240 Countries/territories
>260 Million queries/day
>450 Million Windows Users
Serving 500M users requires massive scale

- Five datacenters
- 300,000 Servers
- ~100 Edge Nodes
- >$1 Billion/Year infrastructure cost
Bing AutoSuggest in Numbers

150k+ Keystrokes/ Second
500+ # AS servers
~50ms Server latency
1.1B+ Suggestions (USA)
~30 Supported countries
Outages are Newsworthy

How long will big-name customers like Netflix put up with Amazon cloud outages?

Amazon’s cloud went down, again, this time on Christmas Eve, for 12 hours, blacking out 7% of AWS customers. Will continued cloud outages erode confidence in the public cloud?

Microsoft Bing Outage Brings Down Yahoo Search, Cortana, and Siri

Microsoft’s search engine went down on Friday morning due to what seems to be a bad update rolled out by the freeload-based software giant and which also affected a number of other services, including Yahoo Search, Windows Phone’s personal assistant Cortana, and Apple’s Siri.

Twitter crashes hard, Internet freaks out

NEW YORK (CNNMoney) -- Cue the collective internet freakout! Twitter went down for several hours on Thursday afternoon, deprivation users of a place to complain that Twitter was down.

Bing is down: ‘The page you want isn’t available’ (update: fixed)

Google, YouTube Outages Whip Twitter Into Frenzy: ‘Is This a Global Emergency?’

Gmail Went Down And Everyone Panicked

It’s not just you. Gmail went down, and everyone flipped. At around 2 p.m. EST, Google’s email service became largely inaccessible, as confirmed by the website Down For Everyone Or Just Me. But you probably already knew that if you’ve been checking Twitter.

Twitter goes down, chaos and productivity ensue (UPDATE: It’s back)

Facebook, Instagram briefly go down; Twitter freaks out

Microsoft Searches for Cause of Bing Outage

Failure related to internal testing after major outage

Amazon EC2 Outage Shows Risks of Cloud

Amazon’s EC2 cloud went dark last week, knocking sites like Foursquare, SoundCloud, and Quora offline, and affecting hundreds of Amazon cloud customers. The outage is a black eye for the young cloud services industry and gives businesses a reason to think twice.

Is Facebook down? A history of outages

When Facebook goes down it’s a serious issue: bored office workers are bereft of distractions, children are obliged to talk to their families, media executives cry over lost traffic and nobody gets pissed ... (no seriously, that still exists).

In addition, Facebook engineers frantically rush to get everything back online – but it wasn’t always like that. In 2010 when Facebook’s site went down it could be fixed just by turning it off and on again, literally.

Google Outage: Internet Traffic Plunges 40%

The web giant is refusing to discuss why all its services from Google Search to Gmail to YouTube stopped working across the world.

Yahoo’s search goes dark following Bing outage

Twitter apologizes for worldwide outage

Why is Google silent on its outage?
AutoSuggest
Predicting your query before you type it

What should we suggest?

Alice
User previous queries:
- movie streaming
- imdb ranking

Bob
User location:

Charlie
Day of week: Sunday
Time of day: 12h30
Device: mobile

netflix
new york times
nearby restaurants
Why is it useful?

1. Reduce query formulation effort
2. Prevent misspellings
3. Provide more relevant search results
   • Search Result Pages (SERPs) tend to be optimized for popular queries
4. Provide direct answers
AUTOSUGGEST - BEHIND THE SCENES
Overall architecture
(very simplified)

OFFLINE

<table>
<thead>
<tr>
<th>Data</th>
<th>Process</th>
<th>Pre-rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map reduce</td>
<td>Distributed ML</td>
<td></td>
</tr>
</tbody>
</table>

ONLINE

User Context

Candidates

Suggestion Database

Typed Prefix

“r”

Diversifier

Previous User Queries

1. race cars
2. reddit
3. restaurants
4. recipes

1. race cars
2. reddit
3. restaurants
4. recipes

Ranker
Pre-ranking

• Each query $q$ is associated with (an estimate of) the probability $P(q)$ that a user will use it.

• The estimate is based on:
  • how many times $q$ was typed in the past
  • how recently
  • ... and other factors
Suggestion Database - Candidate Generation

(Scored) compacted tries

- Compressed data structure
  - 1.1B suggestions and their metadata fit in < 30 GB
- Very efficient retrieval of top-k completions
  [Hsu and Ottoviano, WWW 2013]
- Inverted Index over queries for non prefix match suggestions
AUTOSUGGEST – CHALLENGES
**Context Matters**

What should we suggest?

<table>
<thead>
<tr>
<th>Alice</th>
<th>Bob</th>
<th>Charlie</th>
</tr>
</thead>
<tbody>
<tr>
<td>User previous queries:</td>
<td>User location:</td>
<td>Day of week: Sunday</td>
</tr>
<tr>
<td>- movie streaming</td>
<td>new york times</td>
<td>Time of day: 12h30</td>
</tr>
<tr>
<td>- imdb ranking</td>
<td></td>
<td>Device: mobile</td>
</tr>
<tr>
<td><strong>netflix</strong></td>
<td><strong>new york times</strong></td>
<td><strong>nearby restaurants</strong></td>
</tr>
</tbody>
</table>

Model P (Query | UserContext, Time)
User Location

Some queries are much more popular in some places than in others.

Modify Query Prior Prob by Affinity \((\text{Query, Location})\) learnt from data.
Localized Suggestions in Action

User Location: Redmond, WA

User Location: New York, NY
Previous Query

Modify Query Prior by Affinity \((Query, Previous Query)\)

User Previous Query:

Querying again from Search Result Page:
5-15% of submitted queries contain spelling errors

- Offline spell corrections
  - Popular misspellings are stored in the trie with a pointer to their correction

- Online spell corrections
  - Exploration of the trie using an error model learned from the data (frequent typos have low penalty) [Duan and Hsu, WWW 2011]
Diversity

Avoid showing “duplicate” suggestions e.g.:
- “aol”, “aol homepage”, “aol.com official site”

Represent a diverse set of intents when input is underspecified
More Challenges

- Filtering queries with explicit sexual intent, offensive queries and queries inciting to commit crimes
- Filtering Spam queries
- Ensuring high availability and low latency
- ...

Trade-offs – what to build and how to build it

This is were understanding the fundamentals of computing at scale kicks in

There are no easy answers:

Each potential solution has a cost in terms of complexity, storage, compute usage, serving cost

These costs need to be weighed up against product impact (which we won’t cover here)

Everything needs to be measured:

- Instrument production systems
- Profile during coding, don’t assume
- Gate on performance during builds and deployments
- Continuously evaluate and re-evaluate as systems change
Systems vs fundamentals: Lesson 1

Core trie data structure optimized to do micro second lookups

Custom data structure with per processor optimization to reduce cost online

Extreme computing win, yes?

Partly:

Code so complex it’s almost impossible to maintain

Extreme computing requires extreme engineering: turns out the serialization code was 100x slower than the data structure
Let’s go back to basics. We have a middle tier workflow engine written in C#. You’ve been asked to check at runtime whether a string is a duplicate in a list of strings you’ve already. How do you do this?

A hash table/map? Right?

Depends.

It turns out for sufficiently small collections, iterating every time is not much slower.

Also, because every machine is serving multiple requests, it also reduces the impact on other requests by reducing memory overhead and garbage collection pressure.

Sometimes, simple data structures and algorithms are more efficient overall.
Systems vs fundamentals: Lesson 3

Map-reduce is awesome, it allows us to process petabytes of data.

However, at some point even map-reduce doesn’t scale.

Issues:

Data skew

Logs, even split across partitions, larger than can be read or processed.

Sampling as a solution.
Systems vs fundamentals: Lesson 4

The real complexity is not in the code or the individual components though

The system overall is larger than most engineers or scientists can reason over

A lot of extreme engineering is in place to allow extreme computing flourish

This abstraction needs to be balanced against needing to understand the complexity to produce elegant and efficient solutions

There is no easy answer: it’s part science, part art, part experience
Conclusion

The techniques you are learning in this course provide a strong foundation to work at scale.

We focus on these fundamentals:

- When we design systems
- When we interview candidates

However extreme the computing, it fails if the systems are in place to facilitate it.

Personal learning: the optimisations that mattered a generation or so ago matter again.
Be one who empowers billions. BUILD ON

New Grad & Internship opportunities

We’re hiring in the US and Europe!

Paul Baecke

Add me on LinkedIn if you want to stay in touch:

https://uk.linkedin.com/in/paulbaecke