INSA LYON

UNIVERSITÄT PASSAU

Distributed and Multimedia InformationSystems
Prof. Dr. Harald Kosch

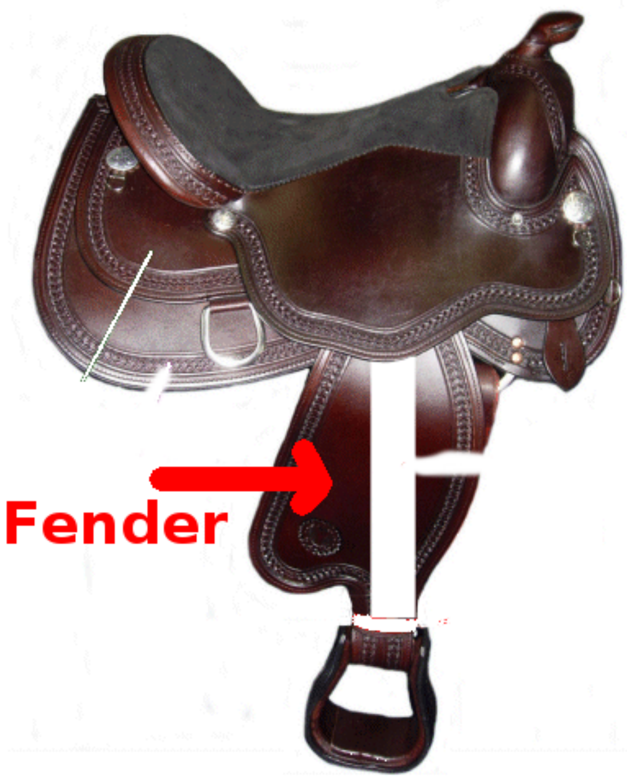
Collaborative(ly) Personalized PageRank

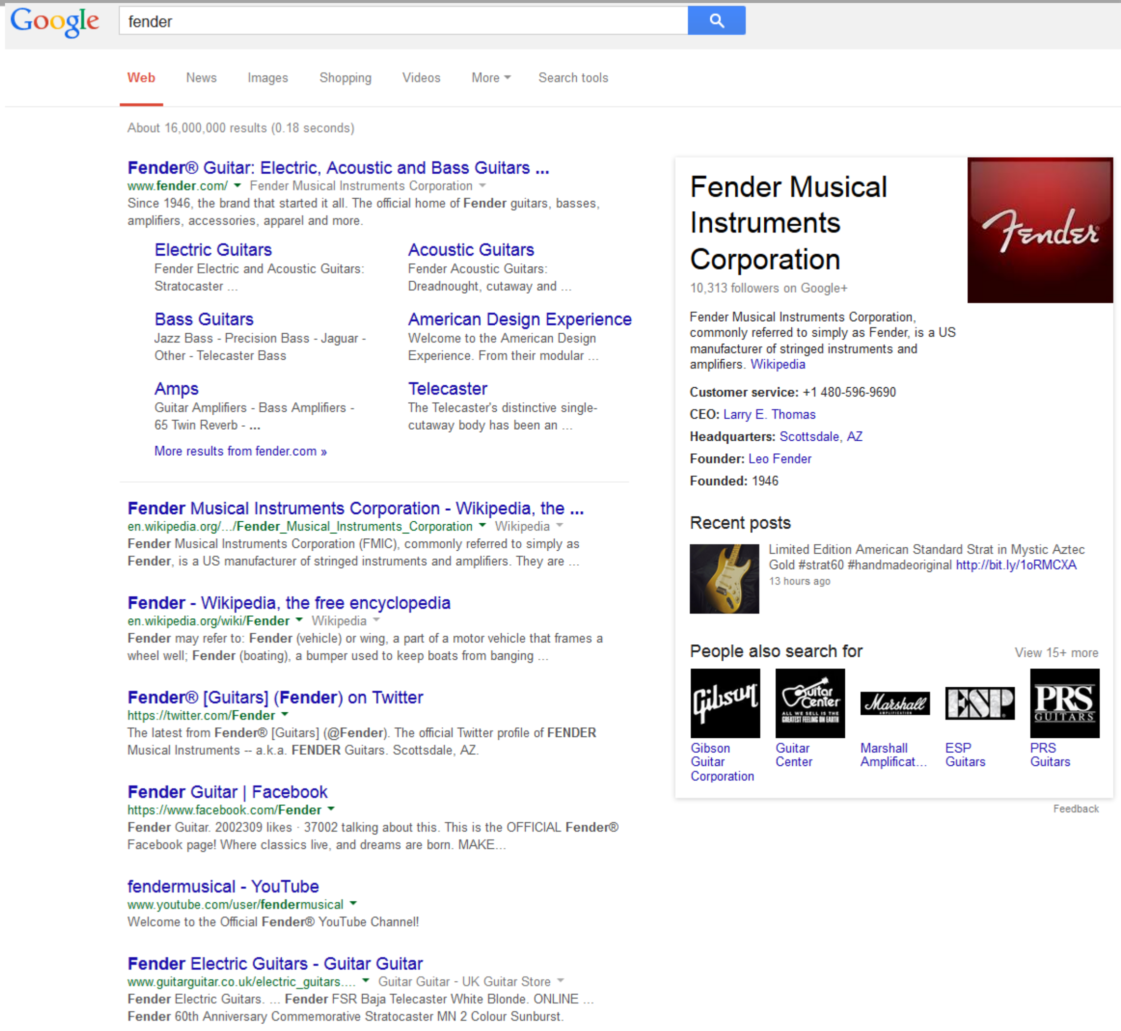
September 2014

Théo Dubourg - theo.dubourg@insa-lyon.fr

This presentation must not be redistributed without express consent of the author.

Now’s when I am convincing you (“3’ to convince”(INSA))





About 16,000,000 results (0.18 seconds)

Fender® Guitar: Electric, Acoustic and Bass Guitars ...
www.**fender**.com/ Fender Musical Instruments Corporation
Since 1946, the brand that started it all. The official home of **Fender** guitars, basses, amplifiers, accessories, apparel and more.

Electric Guitars
Fender Electric and Acoustic Guitars: Stratocaster ...

Acoustic Guitars
Fender Acoustic Guitars: Dreadnought, cutaway and ...

Bass Guitars
Jazz Bass - Precision Bass - Jaguar - Other - Telecaster Bass

American Design Experience
Welcome to the American Design Experience. From their modular ...

Amps
Guitar Amplifiers - Bass Amplifiers - 65 Twin Reverb - ...

Telecaster
The Telecaster's distinctive single-cutaway body has been an ...

More results from fender.com »

Fender Musical Instruments Corporation - Wikipedia, the ...
en.wikipedia.org/.../**Fender**_Musical_Instruments_Corporation Wikipedia
Fender Musical Instruments Corporation (FMIC), commonly referred to simply as **Fender**, is a US manufacturer of stringed instruments and amplifiers. They are ...

Fender - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Fender** Wikipedia
Fender may refer to: **Fender** (vehicle) or wing, a part of a motor vehicle that frames a wheel well; **Fender** (boating), a bumper used to keep boats from banging ...

Fender® [Guitars] (**Fender**) on Twitter
https://twitter.com/**Fender**
The latest from **Fender**® [Guitars] (@**Fender**). The official Twitter profile of **FENDER** Musical Instruments -- a.k.a. **FENDER** Guitars. Scottsdale, AZ.

Fender Guitar | Facebook
https://www.facebook.com/**Fender**
Fender Guitar. 2002309 likes · 37002 talking about this. This is the OFFICIAL **Fender**® Facebook page! Where classics live, and dreams are born. MAKE...

fendermusical - YouTube
www.youtube.com/user/**fender**musical
Welcome to the Official **Fender**® YouTube Channel!

Fender Electric Guitars - Guitar Guitar
www.guitarguitar.co.uk/electric_guitars.... Guitar Guitar - UK Guitar Store
Fender Electric Guitars. ... **Fender** FSR Baja Telecaster White Blonde. ONLINE ... **Fender** 60th Anniversary Commemorative Stratocaster MN 2 Colour Sunburst.

Fender Musical Instruments Corporation
10,313 followers on Google+

Fender Musical Instruments Corporation, commonly referred to simply as Fender, is a US manufacturer of stringed instruments and amplifiers. Wikipedia

Customer service: +1 480-596-9690
CEO: Larry E. Thomas
Headquarters: Scottsdale, AZ
Founder: Leo Fender
Founded: 1946

Recent posts

Limited Edition American Standard Strat in Mystic Aztec Gold #strat60 #handmadeoriginal http://bit.ly/1oRMCXA
13 hours ago

People also search for — View 15+ more

Gibson Guitar Corporation
Guitar Center
Marshall Amplificat...
ESP Guitars
PRS Guitars

Feedback

UNIVERSITÄT PASSAU

Slides Aren’t Mandatory

Hey ! I am a slide.

I am projected, shiny, I’m quite big, but it is not mandatory to read me.

Listening is enough, if needed to look at me, the guy over there will tell you.

Context: Web Search, IR Limits

- Only use the text/document
- Not aware of the “environment”
- Human beings are not really compatible with IR: They do not express themselves the right way.
  “Human beings do not know how to search” - Someone

Context: Web Search, Some Solutions

Use the environment/specificities of the web:

- Linking information: PageRank
- Techniques that Google commonly calls “antispam”: filters, rules, to prune the set of results and filter out the majority that is of low quality and keep the minority of higher quality
- Anchor analysis
- Other graph-based approaches: TrustRank, etc. ...

Context: Web Search Personalization

1. Ideas coming from **Recommender Systems**:
   - **Tailoring** the system's output to the current user.
   - Making "**recommendations**" of certain items vs. others to the user.
2. **Set of items =** all items returned by the IR engine
3. **Recommended items =** items that should be ranked higher / rank-merging with the IR score
4. Can be seen as another type of "**filter**"
5. **Content-based filtering:** based on user profile + item profile
6. **Collaborative filtering:** based on collaboratively collected info

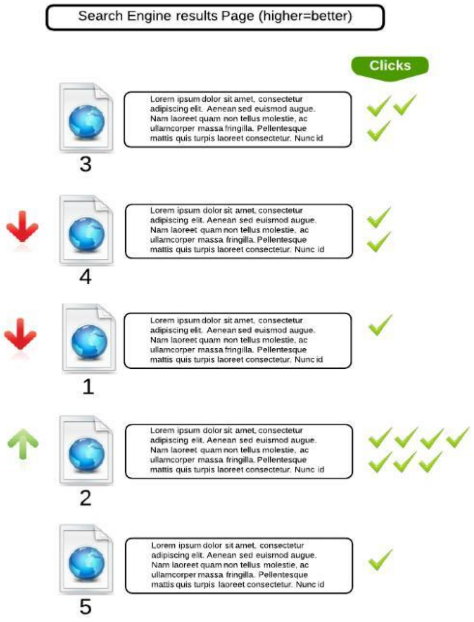
My Proposal: Original Idea

Link/Graph Analysis vs Recommender Systems

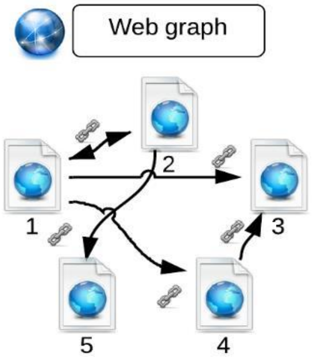
My Proposal: Previous Works Basis

- “A Large Scale Evaluation of Personalized Search Strategies”, Dou, 2007
- “Topic-Sensitive PageRank”, Haveliwala, 2002

Clickthrough-based re-ranking / collaborative filtering


Search Engine results Page (higher=better)
Clicks
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean sed euismod augue. Nam laoreet quam non tellus molestie, ac ullamcorper massa fringilla. Pellentesque mattis quis turpis laoreet consectetur. Nunc id
3
4
1
2
5

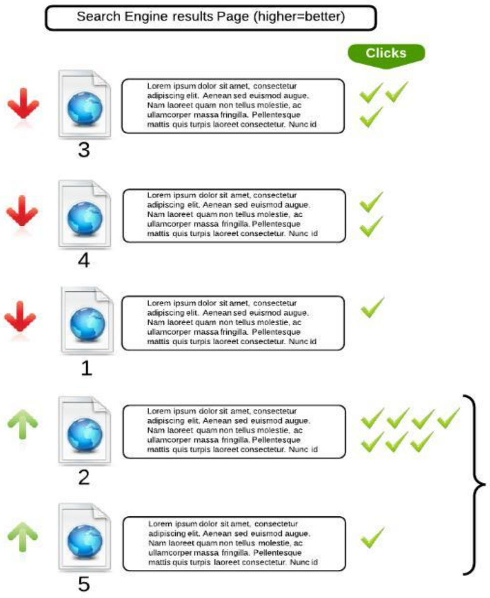
PageRank & Personalized PR



PageRank assumption:
if X has a link to Y then X and Y deal with similar information and X “asserts” for quality of Y’s content.

Personalized PageRank:
Gives a bias to the PageRank computation towards a given set of pages

Collaborative Personalized PageRank


Search Engine results Page (higher=better)
Clicks
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean sed euismod augue. Nam laoreet quam non tellus molestie, ac ullamcorper massa fringilla. Pellentesque mattis quis turpis laoreet consectetur. Nunc id
3
4
1
2
5

2 has a link to 5, so 5 benefits from clickthrough of 2 too

Project Definition

Features we want to achieve:

- **Web search personalization**: Results depending on the user
- **Collaborative approach:** results depending on users similar to current user
- **Usage-based approach:** the system adapts to what you do
- **Implicit and automated:** As a user, you do not need to do anything for the personalization to take place

Data we have: AOL Search Logs (2006), Internet-connection-reachable data

Project Schedule(s)

Original

<table>
<tr><th colspan="2">March</th><th colspan="2">April</th><th colspan="2">May</th><th colspan="2">June</th><th colspan="2">July</th><th colspan="2">August</th><th colspan="2">September</th></tr>
<tr><td colspan="2">Subject definition</td><td colspan="2">Literature review (2)</td><td colspan="6">Thesis writing</td><td colspan="2">Thesis written finalizatio</td><td colspan="2">Finalization</td></tr>
<tr><td colspan="2">Literature review (1)</td><td colspan="6">System design</td><td colspan="4">Evaluation</td><td colspan="2">Defense</td></tr>
<tr><td colspan="2"></td><td colspan="2">Minor development</td><td colspan="4">Implementation</td><td colspan="6"></td></tr>
</table>

Effective

<table>
<tr><th colspan="2">March</th><th colspan="2">April</th><th colspan="2">May</th><th colspan="2">June</th><th colspan="2">July</th><th colspan="2">August</th><th colspan="2">September</th></tr>
<tr><td colspan="2">Subject definition</td><td colspan="2">Literature review (2)</td><td colspan="2"></td><td colspan="5">Thesis writing</td><td colspan="2">Thesis written finalization</td><td colspan="1"></td></tr>
<tr><td colspan="2">Literature review (1)</td><td colspan="8">System design</td><td colspan="2"></td><td colspan="2">Finalization</td></tr>
<tr><td colspan="2"></td><td colspan="2">Minor development</td><td colspan="7">Implementation: System &amp; Evaluation system</td><td colspan="1">Evaluation (User Study)</td><td colspan="1">Defense</td><td colspan="1"></td></tr>
</table>

Project Main Tasks

- User Model Definition
- Usage Extraction
- Collaboration
- Web Graph Personalized Scoring
- PageRank Personalization
- SERP Personalization

Usage Extraction

- Queries clustering $c_i(q) = \frac{|kw(q, cluster(i))|}{|q|}$

$$c(q) = \begin{pmatrix} c_0(q) \\ c_1(q) \\ \vdots \\ c_{130}(q) \end{pmatrix}$$

- User profile $c_l(u) = \sum_{p \in \mathcal{Q}(u)} P(q|u)w(q)c(q)$ $P(q|u) = \frac{clicks(q, u)}{clicks(\bullet, u)}$

- Implementation: Runs fast enough. Using Numpy vectors.

Personalization

- User-to-user similarity $sim(u_1, u_2) = \frac{c_l(u_1) \cdot c_l(u_2)}{||c_l(u_1)||\ ||c_l(u_2)||}$
- Top 100 similar users (excluding sim = 1.0 ones)
- Scoring using similar users (collaboration) : $score(u, q, p) = \frac{\sum_{u_s \in \mathcal{S}_u} (sim(u_s, u) \cdot |clicks(q, p, u_s)|)}{\beta + \sum_{u_s \in \mathcal{S}_u} |clicks(q, \bullet, u_s)|}$
- 3 implementations of the scoring:
  - **Straightforward**: store in DB, retrieve from DB when needed, with caching ⟶ Hugely² slow
  - **DB accesses grouped**: download & DB accesses by batches + caching ⟶ Hugely slow
  - **Store in RAM**, process in-RAM (no cache needed) ⟶ Quite OK
    - Multiprocessed in-RAM computation ⟶ Viable solution (~ 1 day) (10e-7s/sim)
    - Could be scaled using more CPUs / servers

PageRank Personalization

- Recall: Standard PageRank → $R = c(M + E)R$
- Our *personalization vector*:

$$E(q,u) = \begin{pmatrix} p_0 \\ \vdots \\ p_i \\ \vdots \\ p_n \end{pmatrix}, p_i = \begin{cases} \frac{1}{N} & i \notin clicks(\mathcal{S}_u, \bullet, \bullet) \\ score(u,q,i) & i \in clicks(\mathcal{S}_u, \bullet, \bullet) \end{cases}$$

- CPPR formula → $CPPR(q,u) = c(M + E(q,u))CPPR(q,u)$

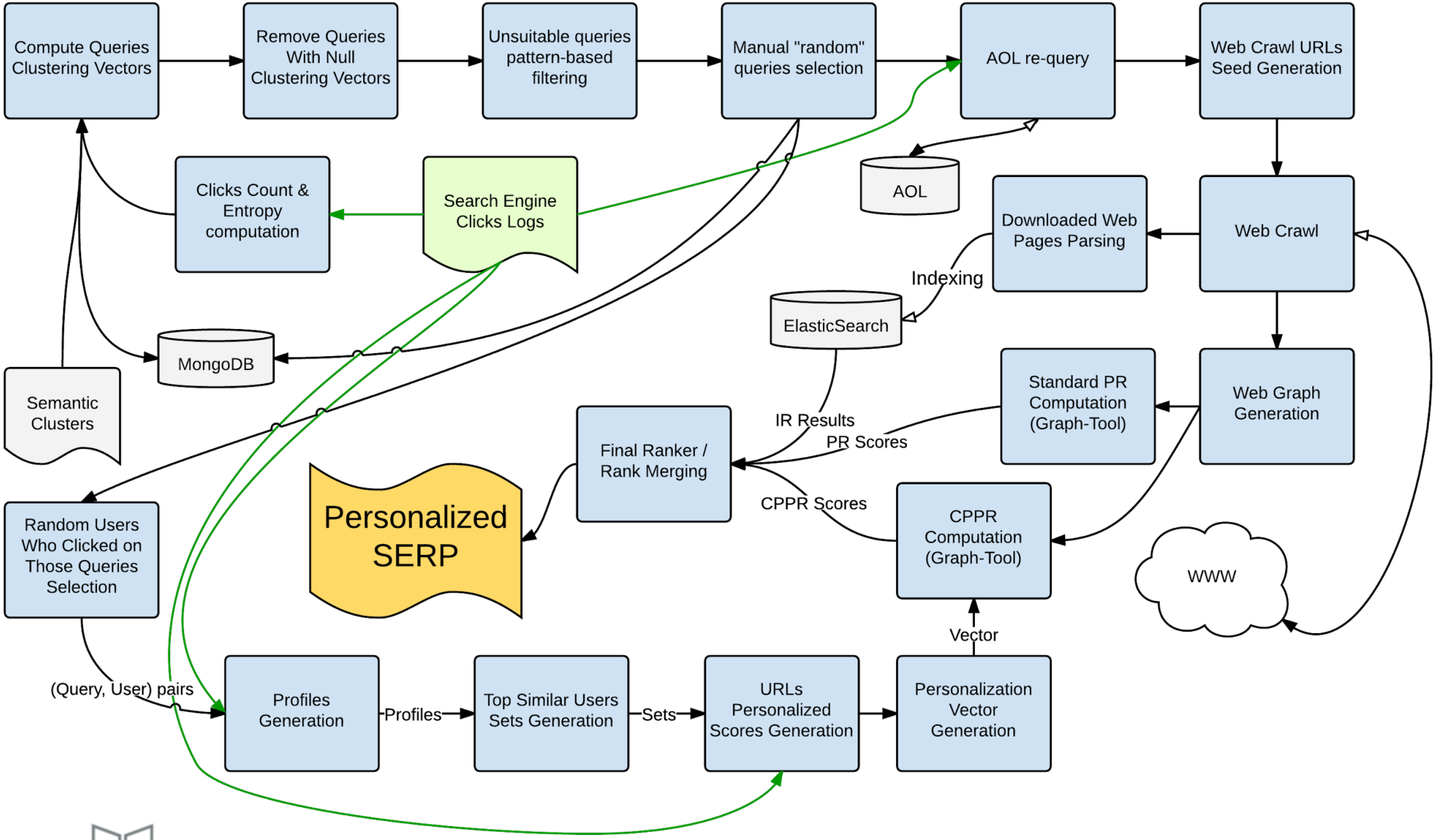
AOL Re-Query / Web Crawl / Indexation

- AOL Re-Querying System
  - Loads keywords & related logs entries
  - Loads the SERP
  - Analyses SERP vs. logs to decide if we keep this SERP
  - Anti-bots protections workarounds: proxies, delays, tor, etc. ...

- Web Crawl
  - 7/3 other domains/same domain links following strategy

- Web Crawl Indexation
  - ElasticSearch with BM25
  - Several processing servers committing → central ElasticSearch Server

All Together



Compute Queries Clustering Vectors
Remove Queries With Null Clustering Vectors
Unsuitable queries pattern-based filtering
Manual "random" queries selection
AOL re-query
Web Crawl URLs Seed Generation
Clicks Count & Entropy computation
Search Engine Clicks Logs
AOL
Downloaded Web Pages Parsing
Web Crawl
Indexing
ElasticSearch
MongoDB
Semantic Clusters
Standard PR Computation (Graph-Tool)
Web Graph Generation
IR Results
PR Scores
Final Ranker / Rank Merging
CPPR Scores
Personalized SERP
CPPR Computation (Graph-Tool)
WWW
Random Users Who Clicked on Those Queries Selection
Vector
(Query, User) pairs
Profiles Generation
Profiles
Top Similar Users Sets Generation
Sets
URLs Personalized Scores Generation
Personalization Vector Generation

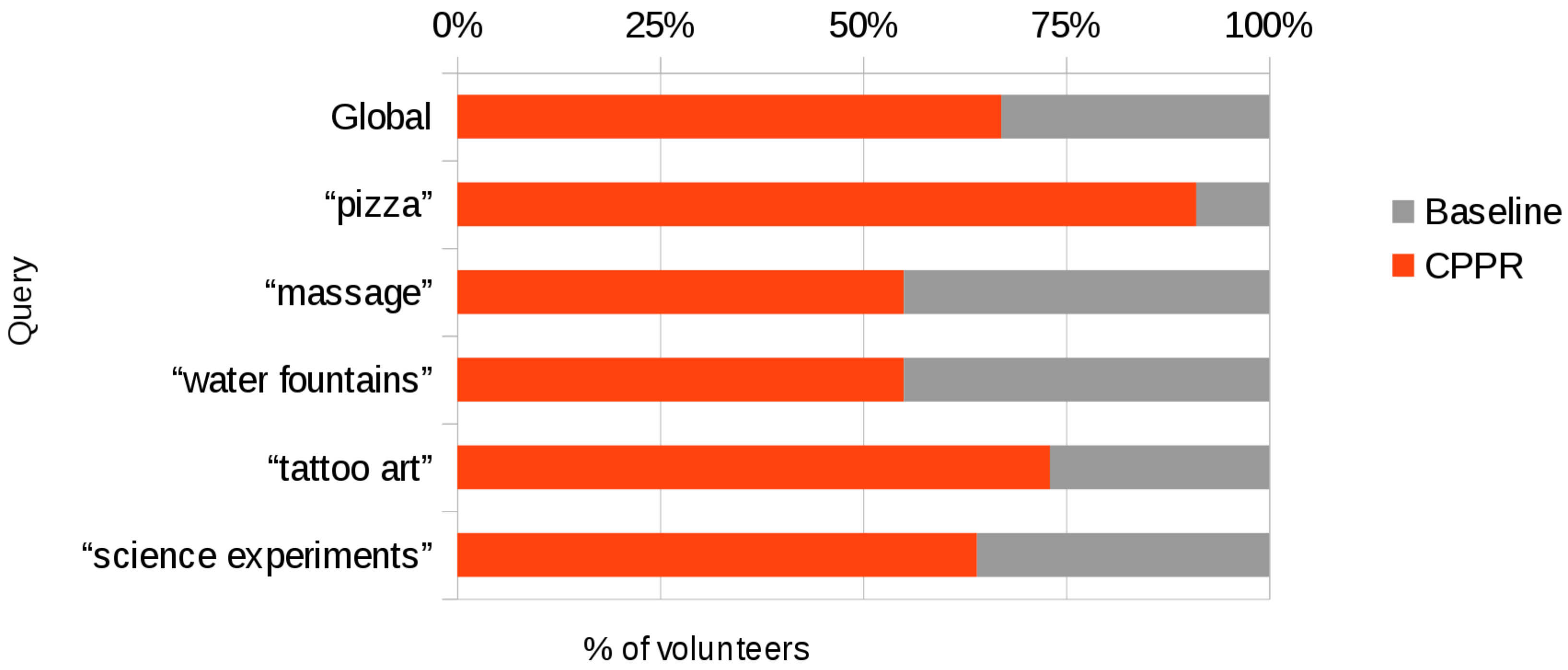
User Study

- 5 queries
- 5 contexts (user + history)
- 11 volunteers
- Asked which preferred ranking
- Asked to select “at most 5 relevant links” for every ranking
- *Precision* metric: $$p_a(q,u) = \frac{\sum_{r \in \mathcal{R}(q,u)} (11 - rank(r))}{\sum_{i \in [[1,5]]} (11 - i)}$$

Results (1)

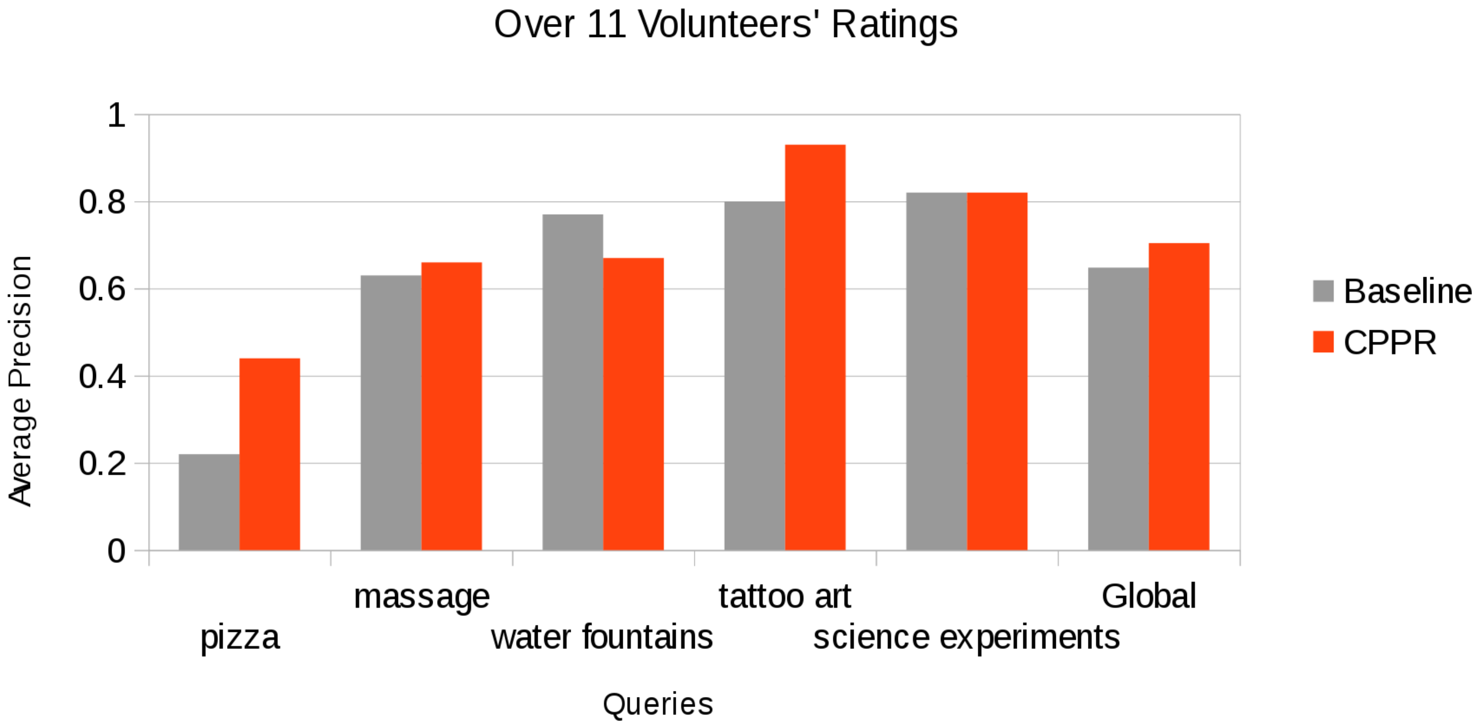
- Preferred ranking: CPPR

"Best ranking overall" selection



Results (2)

- Average precision gain: 21.6%



Retrospective/Figures/Taking a look behind…

>760
commits

1
student

200GB
inter-servers exchanged data

95
days github streak!

45k
written lines of codes

400GB
cloud SSD peak usage

1
computer uptime record

100GB
laptop-uploaded data

700
hours of work

22
CPUs peak usage

2M
indexed documents

>300GB
crawled websites

11
volunteers

some figures might have been ceiled :)

2
cats pictures

Technologies Used (that I already knew)

- Docker! (LXC) (now part of the “developer survival kit”)
- CentOS, Fedora, Ubuntu, Debian
- Data Crunching: Python (http://python.org)
- Web Crawl: Scrapy (http://scrapy.org)

Technologies Used (never used before) & learnt

- **Written work:** Multimarkdown
- **User Study Online Platform:** Express.js (http://expressjs.com)
- **PageRank computation:** Graph-Tool (http://graph-tool.skewed.de/)
- **HTML Parsing / Web Crawl Post-Processing:**
  - Chardet (https://github.com/chardet/chardet)
  - BeautifoulSoup4 (http://www.crummy.com/software/BeautifulSoup/)
- **IR/BM25 indexation & search:** ElasticSearch (http://elasticsearch.org)
- **Database:** MongoDB (http://mongodb.org)
- **Heavy computation:** Google Compute Engine (http://cloud.google.com/products/compute-engine/)
- **Python modules:** MultiProcessing, GZip, Pickle, JSON

References / Bibliography

Please see the bibliography of the written thesis.

The End

THE

END

Thanks for your attention.
Any questions?

Web Crawl & Indexation (optional slide, for questions)

- Web Crawler based on Scrapy framework
- Follows links:
  - 3 links to same domain
  - 7 links to different domains
  - < 255chars
  - Some patterns excluded
  - Pictures, css, js, etc., excluded
- Trials with several different settings
- ~1M docs in <1d with “nice” settings (not hitting server too heavily)
- Indexation is another story: ~1.6 page/sec
  - Need for a several servers to get it to one day
  - Indexation process not built to be ran in parallel: most processes doing work that has already been done…
  - Separating data chunks by hand in the end...

AOL Re-Query (optional slide, for questions)

- Web scraper
  - Loads keywords & related logs entries
  - Loads the SERP
  - Analyses SERP vs. logs to decide if we keep this SERP

- Anti-bots protections workarounds:
  - slow down ⟶ too slow, or banned
  - proxies ⟶ all banned
  - tor ⟶ banned
  - tor + slow down ⟶ too slow, or banned
  - proxies + slow down ⟶ tricky to add to the framework, but works OK