Query Log Mining

Prof. Ricardo Baeza-Yates, Yahoo! Research, Barcelona, Spain
Dr. Fabrizio Silvestri, ISTI - CNR, Pisa, Italy

History in Search Engines

Alphonse de Lamartine

Source: Wikipedia

History Teaches Everything... Even the Future!

What is History?

- Past Queries
- Query Sessions
- Clickthrough Data

Web Mining

- Content:
  - text & multimedia mining
- Structure:
  - link analysis, graph mining
- Usage:
  - log analysis, query mining
- Relate all of the above
  - Web characterization
  - Particular applications

Dynamic

What's in Query Logs?

The 250 most frequent queried terms in the "famous" AOL query log!

Thanks to http://www.wordle.net for the tagcloud generator

Some Examples!

Some Examples

- AOL User 23187425 typed the following queries within a 10 minutes time-span:
- **you come forward** 2006-05-07 03:05:19
- **start to stay off** 2006-05-07 03:06:04
- **i have had trouble** 2006-05-07 03:06:41
- **time to move on** 2006-05-07 03:07:16
- **all over with** 2006-05-07 03:07:59
- **joe stop that** 2006-05-07 03:08:36
- **i can move on** 2006-05-07 03:09:32
- **give you my time in person** 2006-05-07 03:10:07
- **never find a gain** 2006-05-07 03:10:47
- **i want change** 2006-05-07 03:11:15
- **know who iam** 2006-05-07 03:11:55
- **curse have been broken** 2006-05-07 03:12:30
- **told shawn lawn mow burn up** 2006-05-07 03:13:50
- **burn up** 2006-05-07 03:14:14
- **was his i deal** 2006-05-07 03:15:13
- **i would have told him** 2006-05-07 03:15:46
- **to kill him too** 2006-05-07 03:16:18

I Love Alaska!

- http://www.minimovies.org/documentaires/view/ilovealaska
- “I love Alaska tells the story of one of those AOL users. We get to know a religious middle-aged woman from Houston, Texas, who spends her days at home behind her TV and computer. Her unique style of phrasing combined with her putting her ideas, convictions and obsessions into AOL's search engine, turn her personal story into a disconcerting novel of sorts.

  Over a period of three months, a portrait of a woman emerges who is diligently searching for likeminded souls. The list of her search queries read aloud by a voice-over reads like a revealing character study of a somewhat obese middle-aged lady in her menopause, who is looking for a way to rejuvenate her sex life. In the end, when she cheats on her husband with a man she met online, her life seems to crumble around her. She regrets her deceit, admits to her Internet addiction and dreams of a new life in Alaska.”

Query Logs Analyzed in the Literature

Query log name	Public	Period	# Queries	# Sessions	# Users
Excite '97	Y	Sep '97	1,025,908	211,063	$\sim 410,360$
Excite '97 (small)	Y	Sep '97	51,473	N.D.	$\sim 18,113$
Altavista	N	Aug 2^{nd} - Sep 13^{th} '98	993,208,159	285,474,117	N.D.
Excite '99	Y	Dec '99	1,025,910	325,711	$\sim 540,000$
Excite '01	Y	May '01	1,025,910	262,025	$\sim 446,000$
Altavista (public)	Y	Sep '01	7,175,648	N.D.	N.D.
Tiscali	N	Apr '02	3,278,211	N.D.	N.D.
TodoBR	Y	Jan - Oct '03	22,589,568	N.D.	N.D.
TodoCL	N	May – Nov '03	N.D.	N.D.	N.D.
AOL (big)	N	Dec 26^{th} '03 – Jan 1^{st} '04	$\sim 100,000,000$	N.D.	$\sim 50,000,000$
Yahoo!	N	Nov '05 – Nov '06	N.D.	N.D.	N.D.
AOL (small)	Y	Mar 1^{st} - May 31^{st} '06	36,389,567	N.D.	N.D.

Some Popular Terms: Excite and Altavista

query	freq.
Empty Query	2,586
sex	229
chat	58
lucky number generator	56
p****	55
porno	55
b****y	55
nude beaches	52
playboy	46
bondage	46
porn	45
rain forest restaurant	40
f****ing	40
crossdressing	39
crystal methamphetamine	36
consumer reports	35
xxx	34
nude tanya harding	33
music	33
sneaker stories	32

(a) Excite.

query	freq.
christmas photos	31,554
lyrics	15,818
cracks	12,670
google	12,210
gay	10,945
harry potter	7,933
wallpapers	7,848
pornografia	6,893
"yahoo com"	6,753
juegos	6,559
lingerie	6,078
symbios logic 53c400a	5,701
letras de canciones	5,518
humor	5,400
pictures	5,293
preteen	5,137
hypnosis	4,556
cpc view registration key	4,553
sex stories	4,521
cd cover	4,267

(b) Altavista.

Fabrizio Silvestri: **Mining Query Logs: Turning Search Usage Data into Knowledge**. *Foundations and Trends in Information Retrieval.* (To Appear).

Topic Distribution: Excite and AOL

Topic	Percentage
Entertainment or recreation	19.9%
Sex and pornography	16.8%
Commerce, travel, employment, or economy	13.3%
Computers or Internet	12.5%
Health or sciences	9.5%
People, places, or things	6.7%
Society, culture, ethnicity, or religion	5.7%
Education or humanities	5.6%
Performing or fine arts	5.4%
Non-English or unknown	4.1%
Government	3.4%

Excite

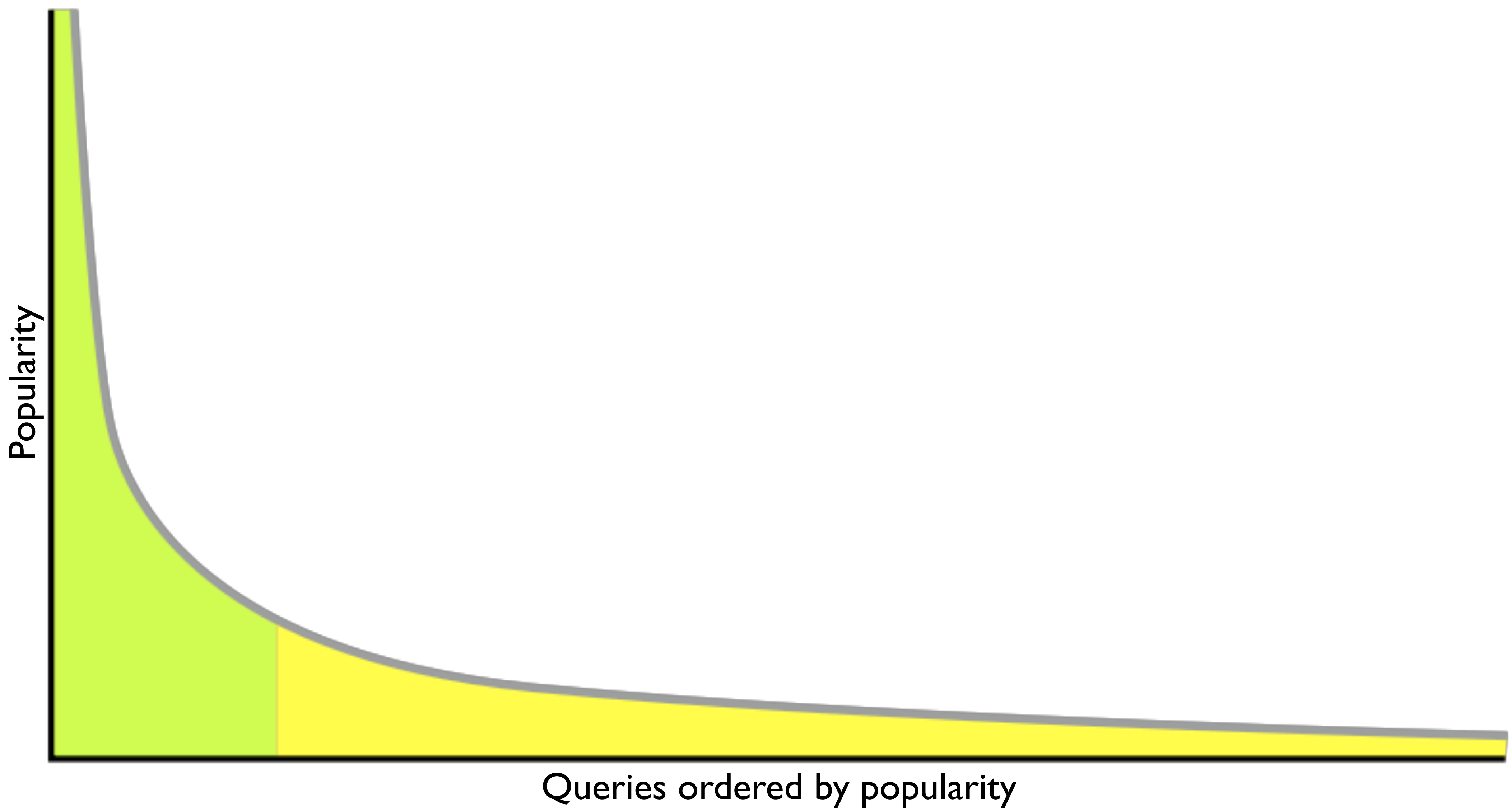
Topic	Percentage
Entertainment	13%
Shopping	13%
Porn	10%
Research & learn	9%
Computing	9%
Health	5%
Home	5%
Travel	5%
Games	5%
Personal & Finance	3%
Sports	3%
US Sites	3%
Holidays	1%
Other	16%

AOL

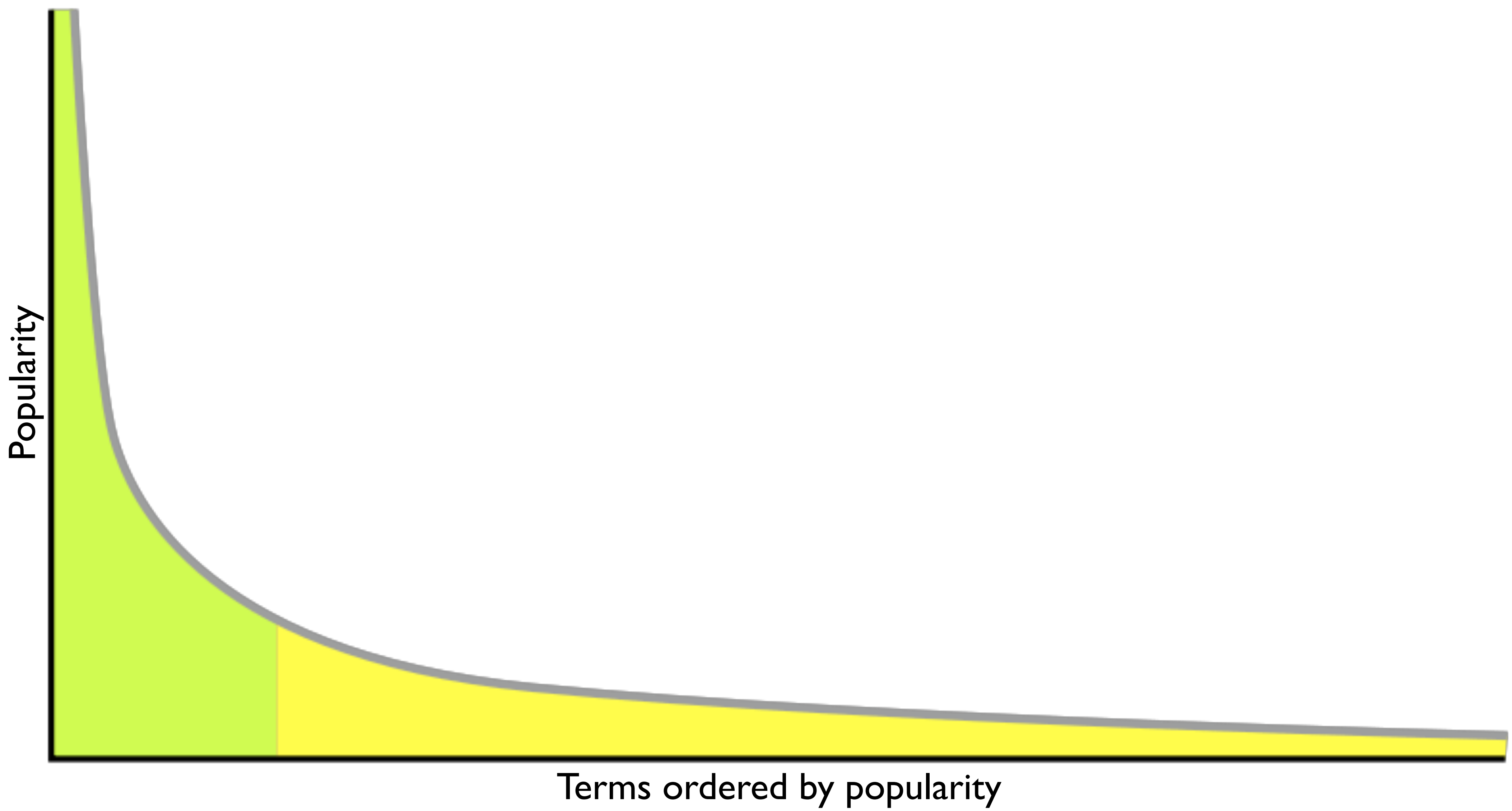
A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic, "**From e-sex to e-commerce: Web search changes**," Computer, vol. 35, no. 3, pp. 107–109, 2002.

S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman, "T**emporal analysis of a very large topically categorized web query log**," J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 2, pp. 166–178, 2007.

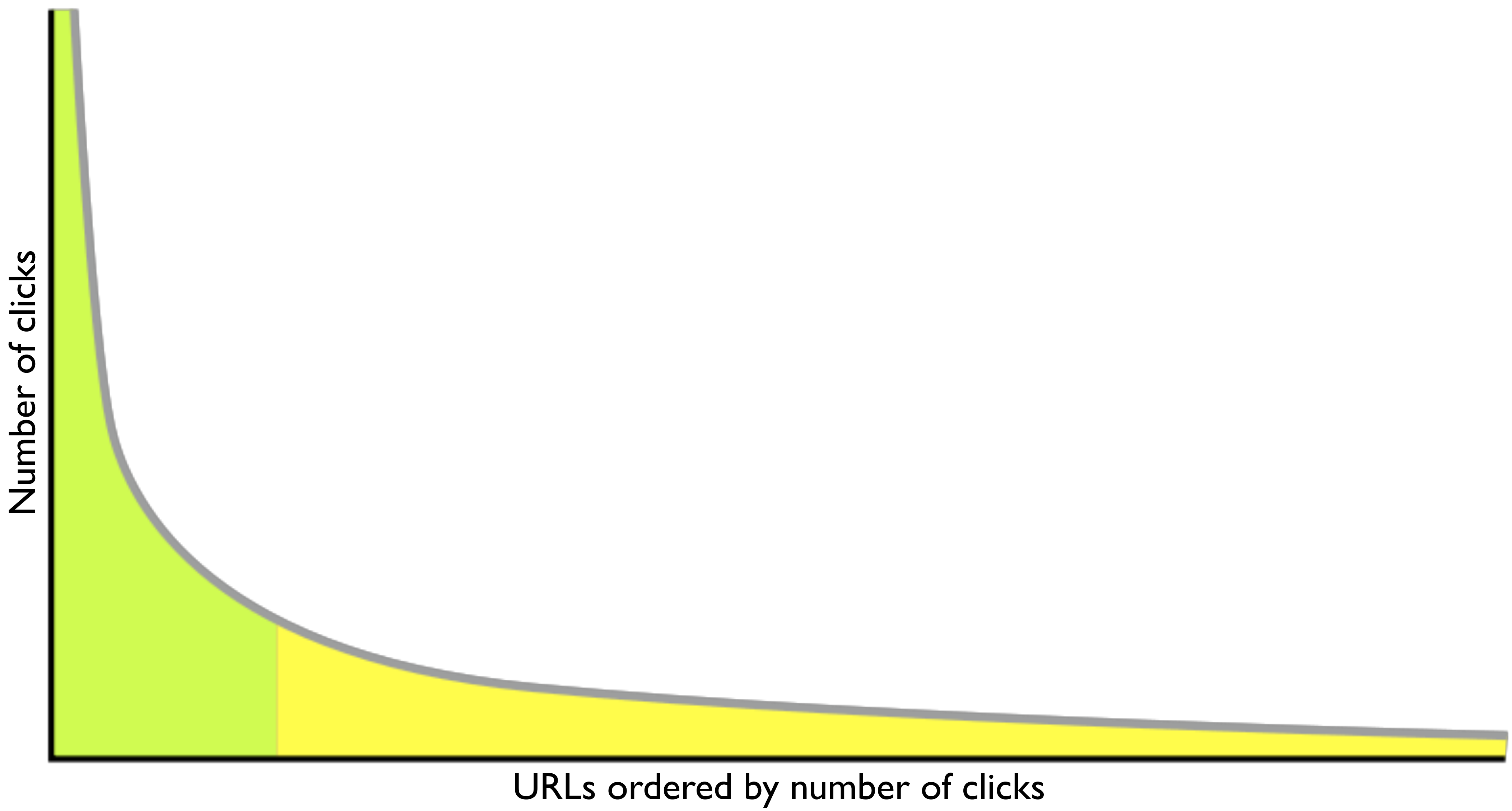
Long Tail Distribution



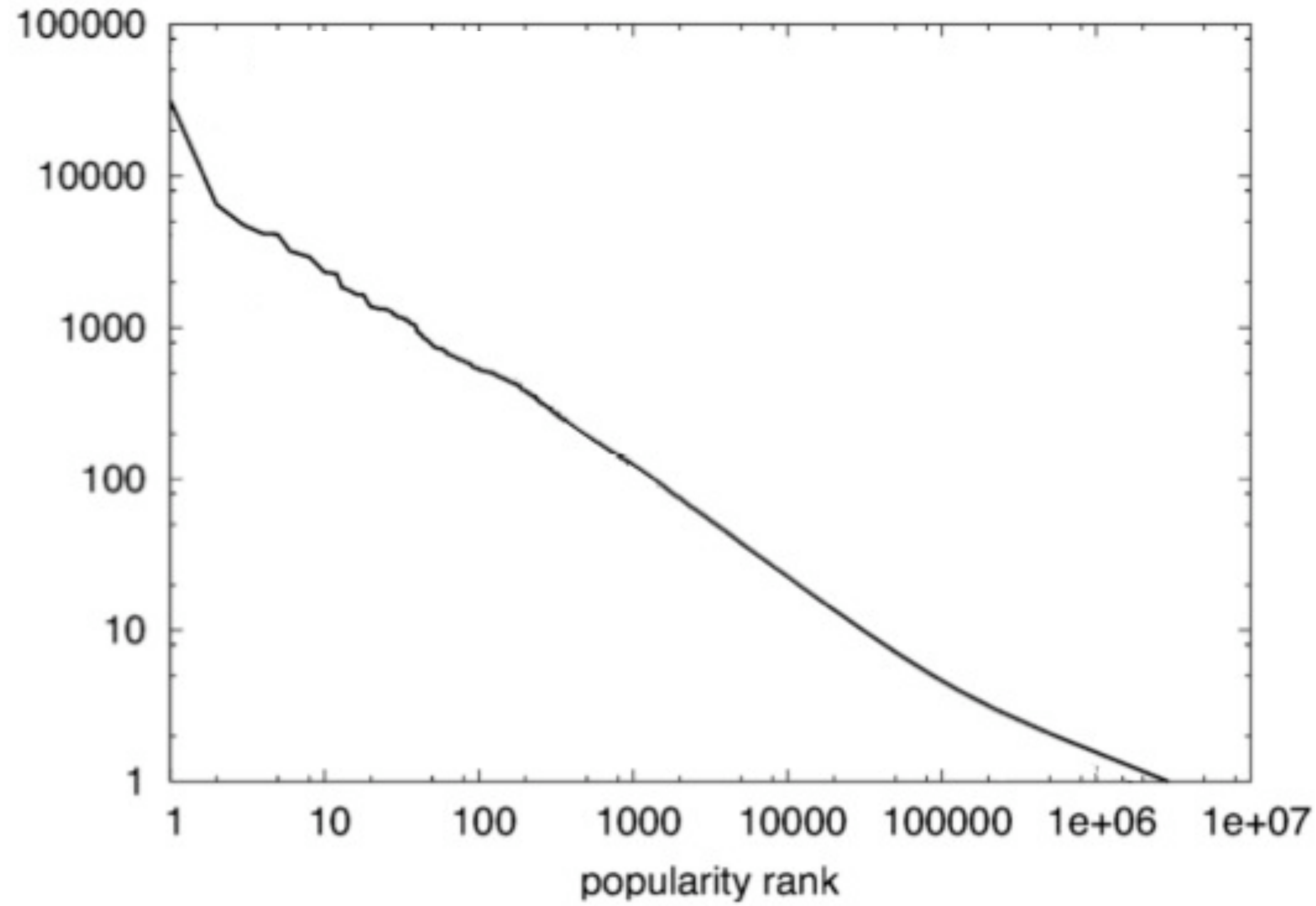
Long Tail Distribution



Long Tail Distribution

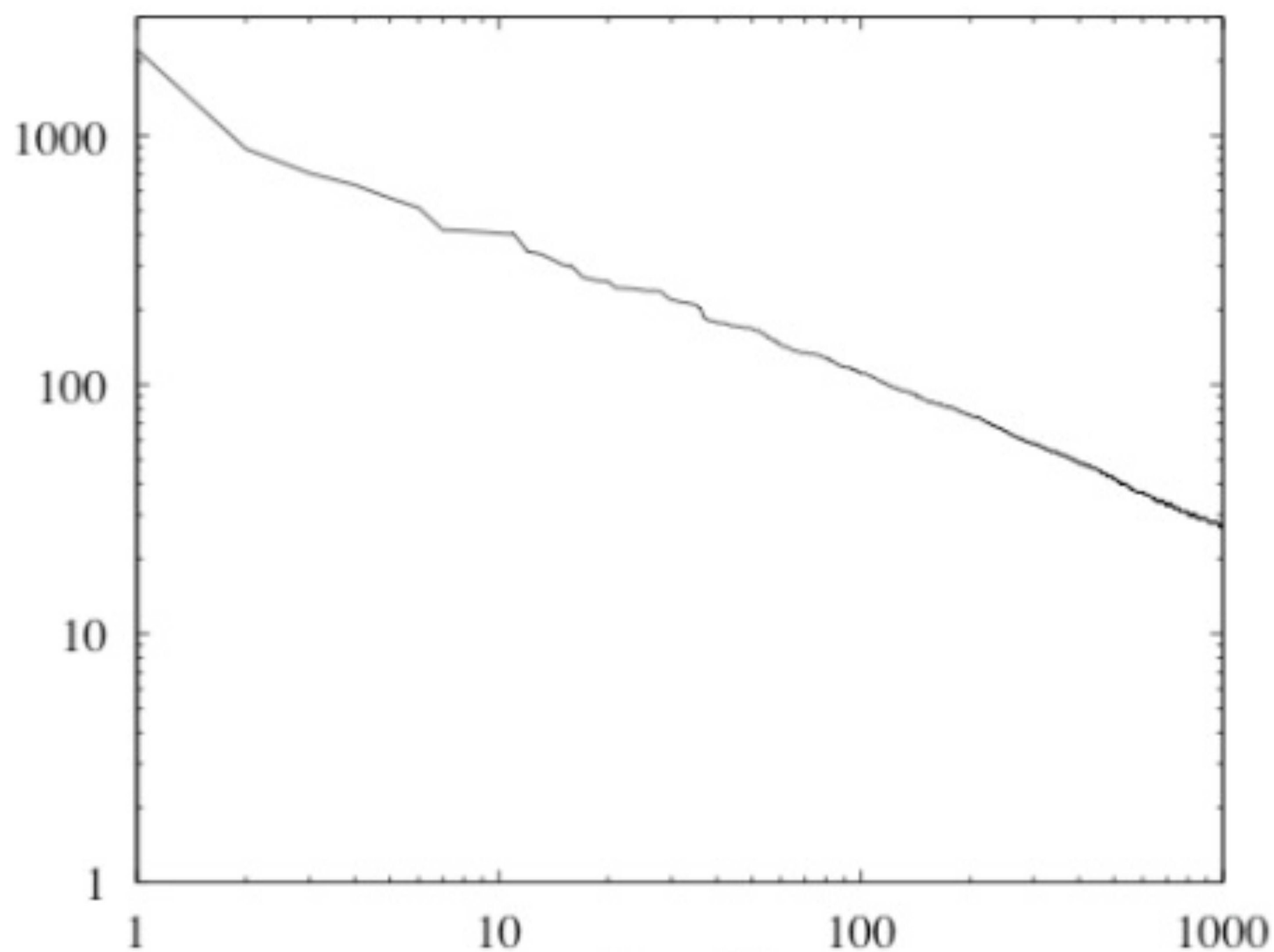


Power-Law In Query Popularity: Altavista



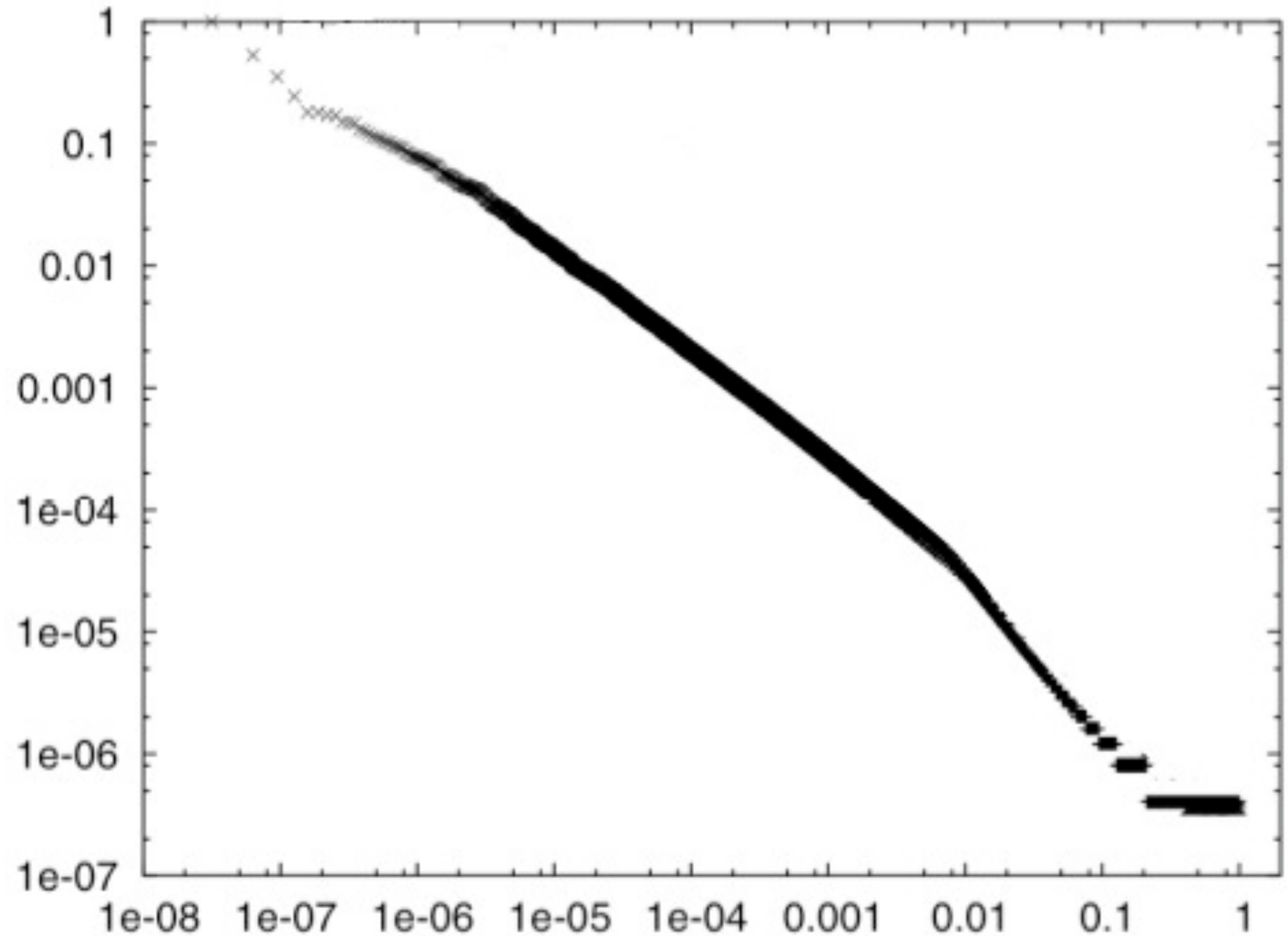
T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "**Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data**," ACM Trans. Inf. Syst., vol. 24, no. 1, pp. 51–78, 2006.

Power-Law In Query Popularity: Excite



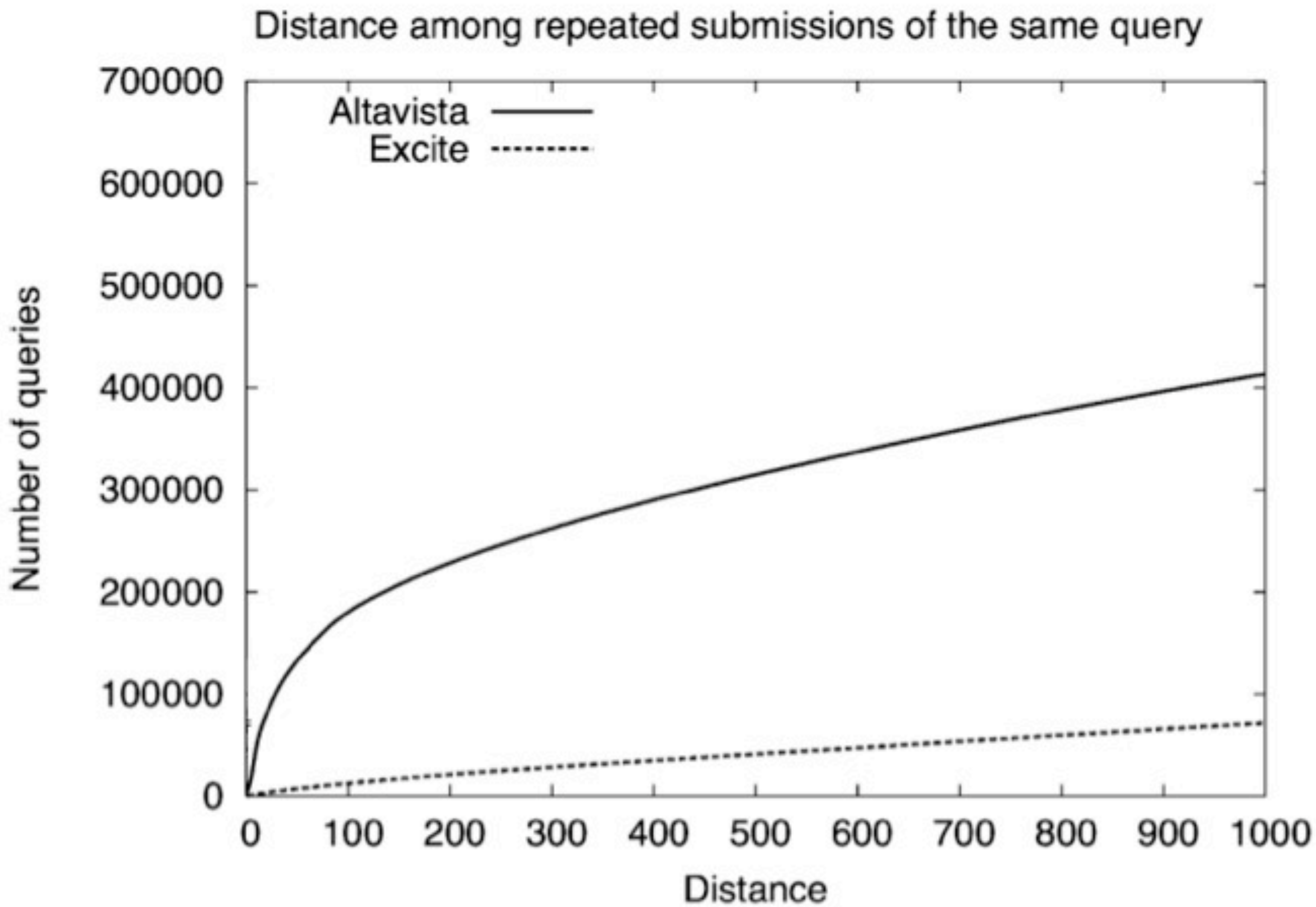
T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "**Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data**," ACM Trans. Inf. Syst., vol. 24, no. 1, pp. 51–78, 2006.

Power-Law In Query Popularity: Yahoo!



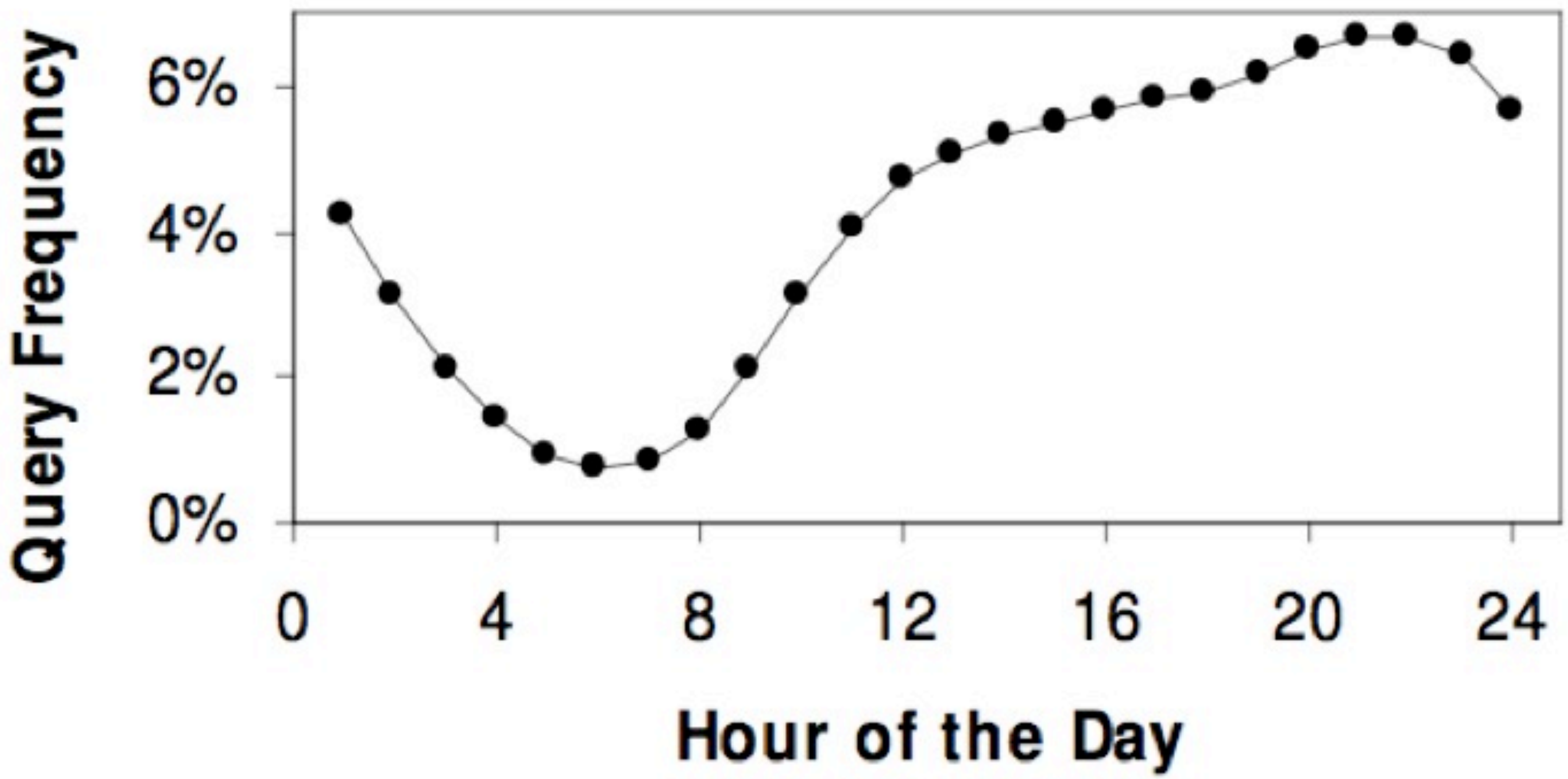
R. Baeza-Yates, A. Gionis, F. P. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri, "**Design trade-offs for search engine caching**," ACM Trans. Web, vol. 2, no. 4, pp. 1–28, 2008.

Query Resubmission



T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "**Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data**," ACM Trans. Inf. Syst., vol. 24, no. 1, pp. 51–78, 2006.

Frequency of Query Submission



S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman, "T**emporal analysis of a very large topically categorized web query log**," J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 2, pp. 166–178, 2007.

Query Statistics: Excite

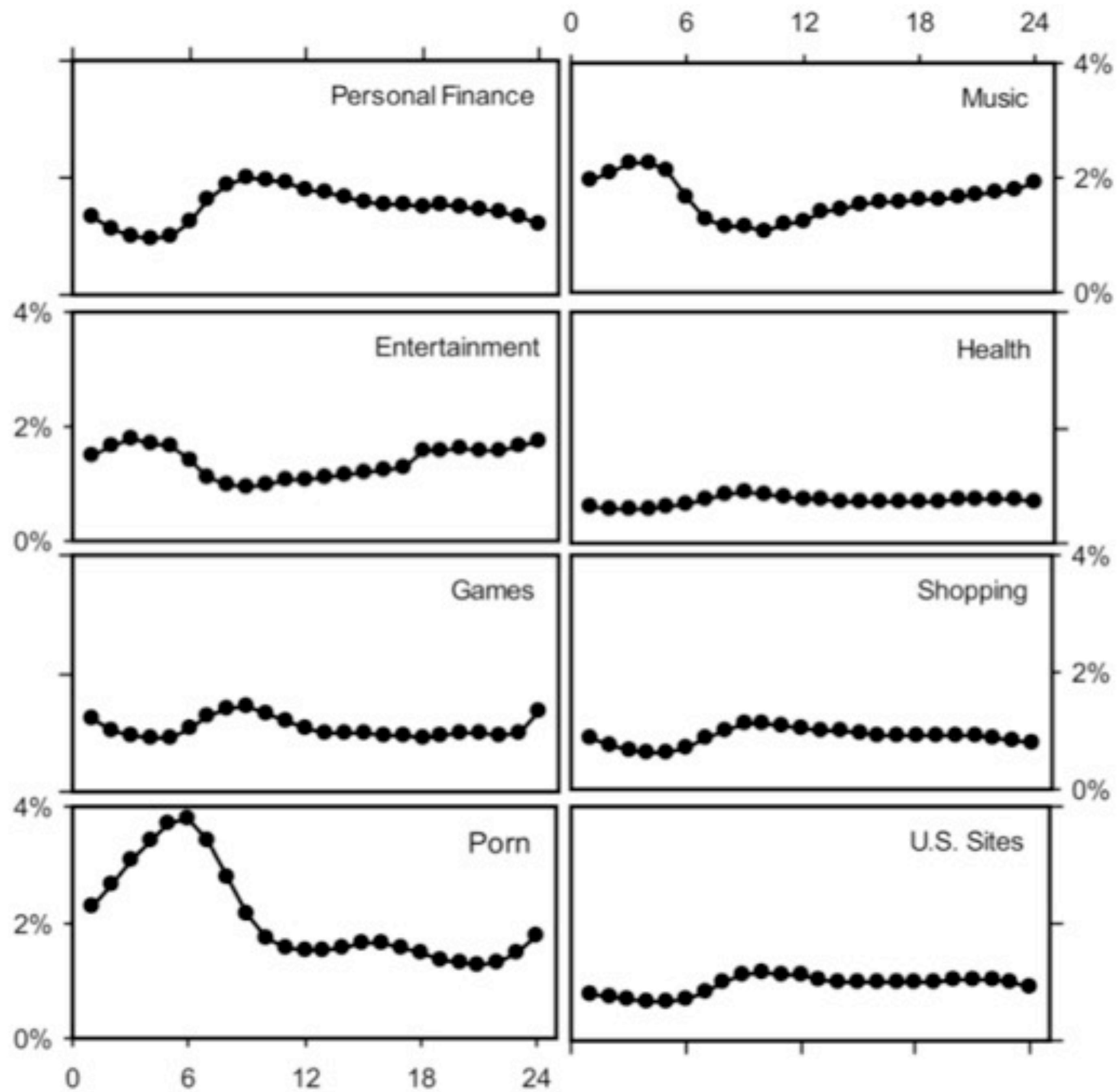
Characteristic	1997	1999	2001
Mean terms per query	2,4	2,4	2,6
Terms per query			
1 term			
2 terms			
3+ terms			
Mean queries per user	2,5	1,9	2,3

In 2008: 2.5 terms per query.

R. Baeza-Yates, A. Gionis, F. P. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri, "**Design trade-offs for search engine caching**," ACM Trans. Web, vol. 2, no. 4, pp. 1–28, 2008.

A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic, "**From e-sex to e-commerce: Web search changes**," Computer, vol. 35, no. 3, pp. 107–109, 2002.

Hourly Topic Distribution

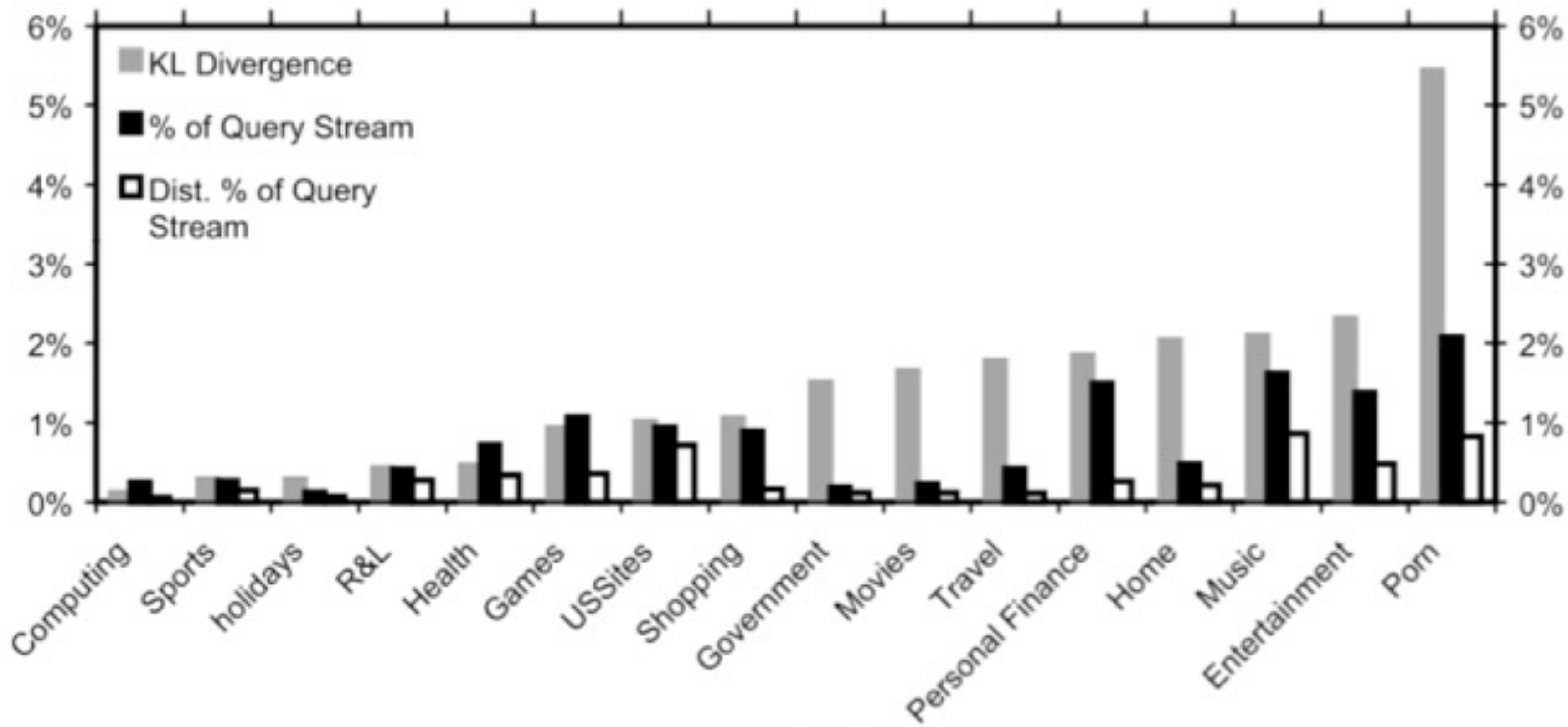


S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman, "T**emporal analysis of a very large topically categorized web query log**," J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 2, pp. 166–178, 2007.

Surprising Topics

$$D\left(p\left(q|t\right)\|p\left(q|c,t\right)\right)=\sum_{q}p\left(q|t\right)\log\frac{p\left(q|t\right)}{p\left(q|c,t\right)}$$

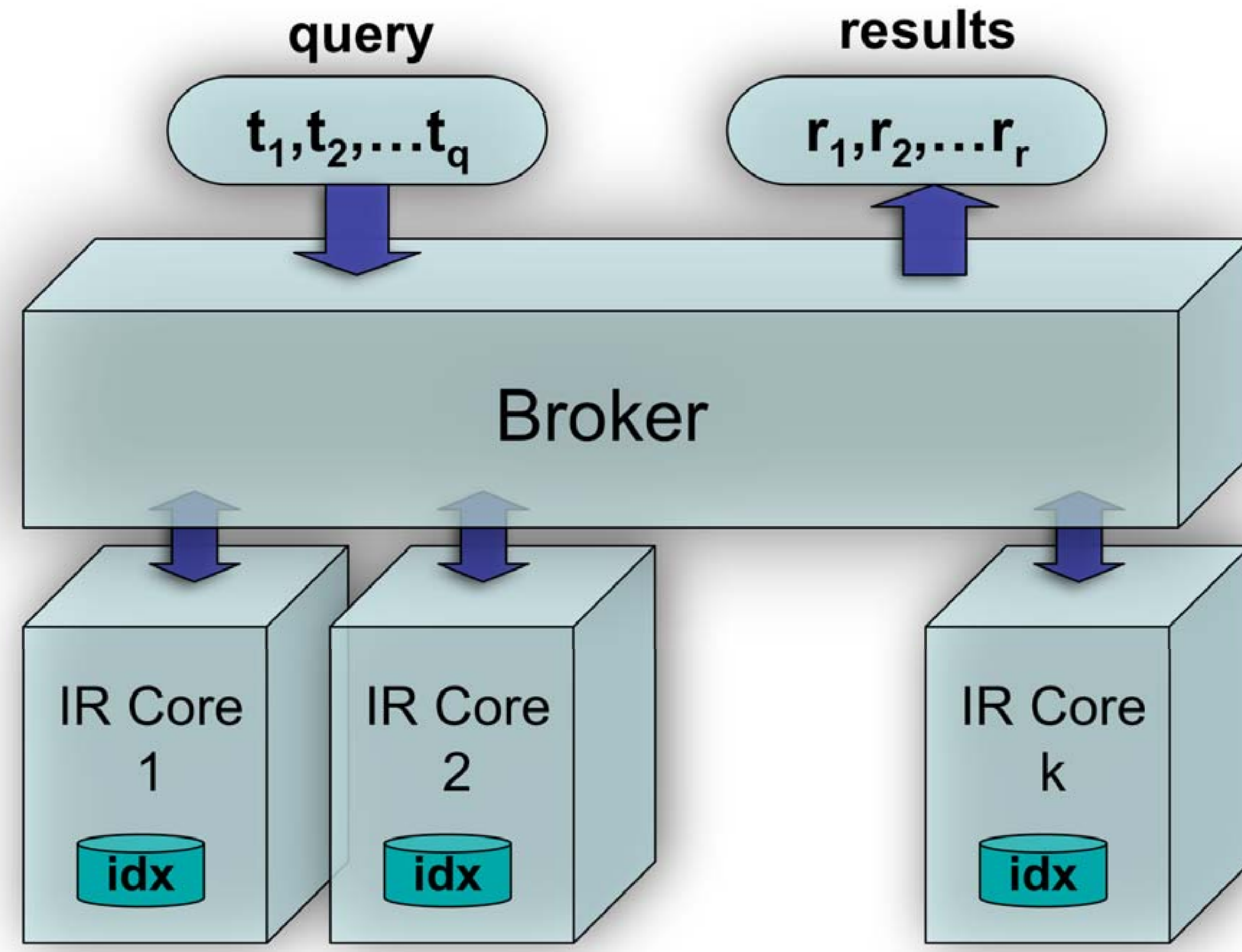
- KL-Divergence betwe
  observing a query topic u.a.r. and the actual topic observed.



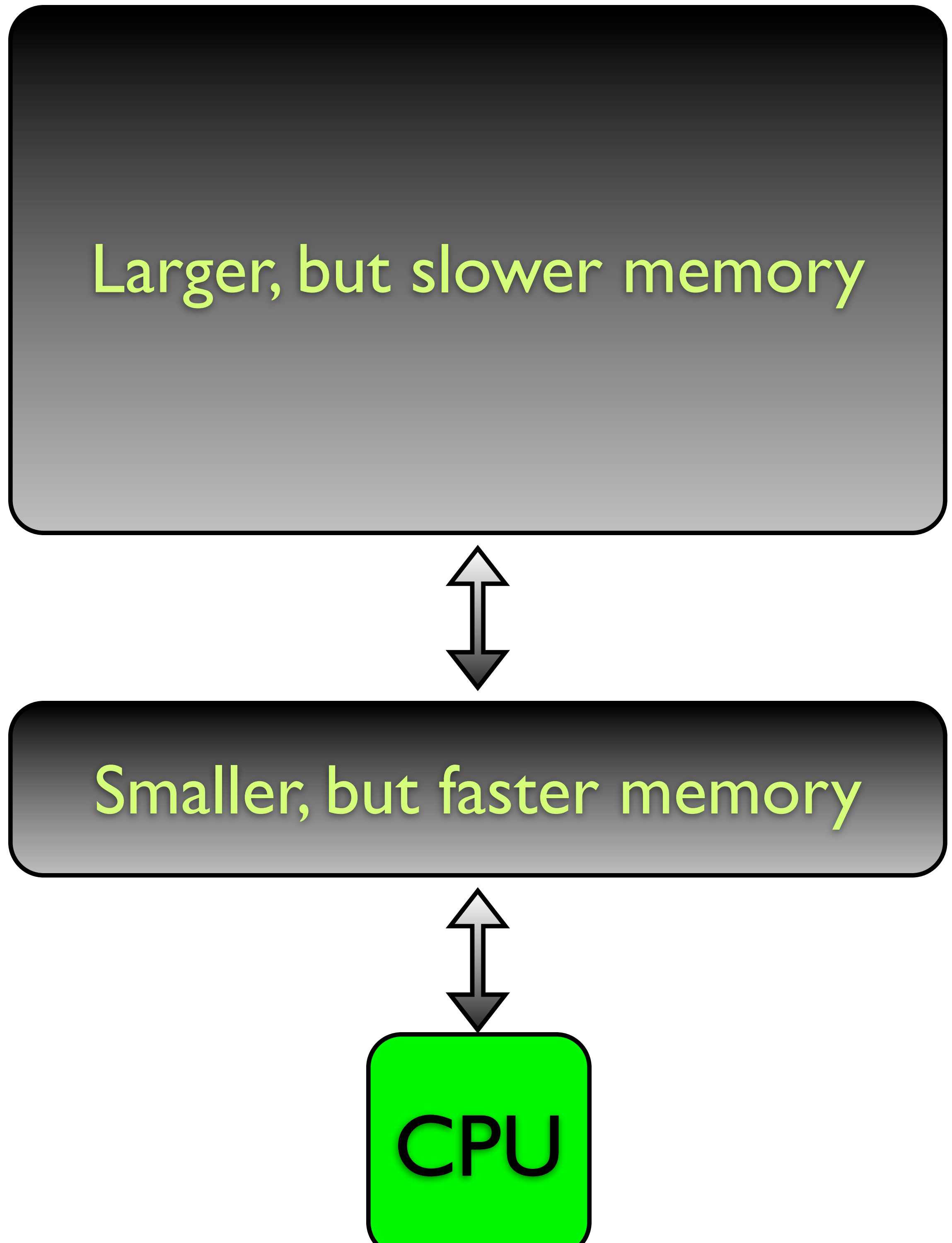
S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman, "T**emporal analysis of a very large topically categorized web query log**," J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 2, pp. 166–178, 2007.

Caching

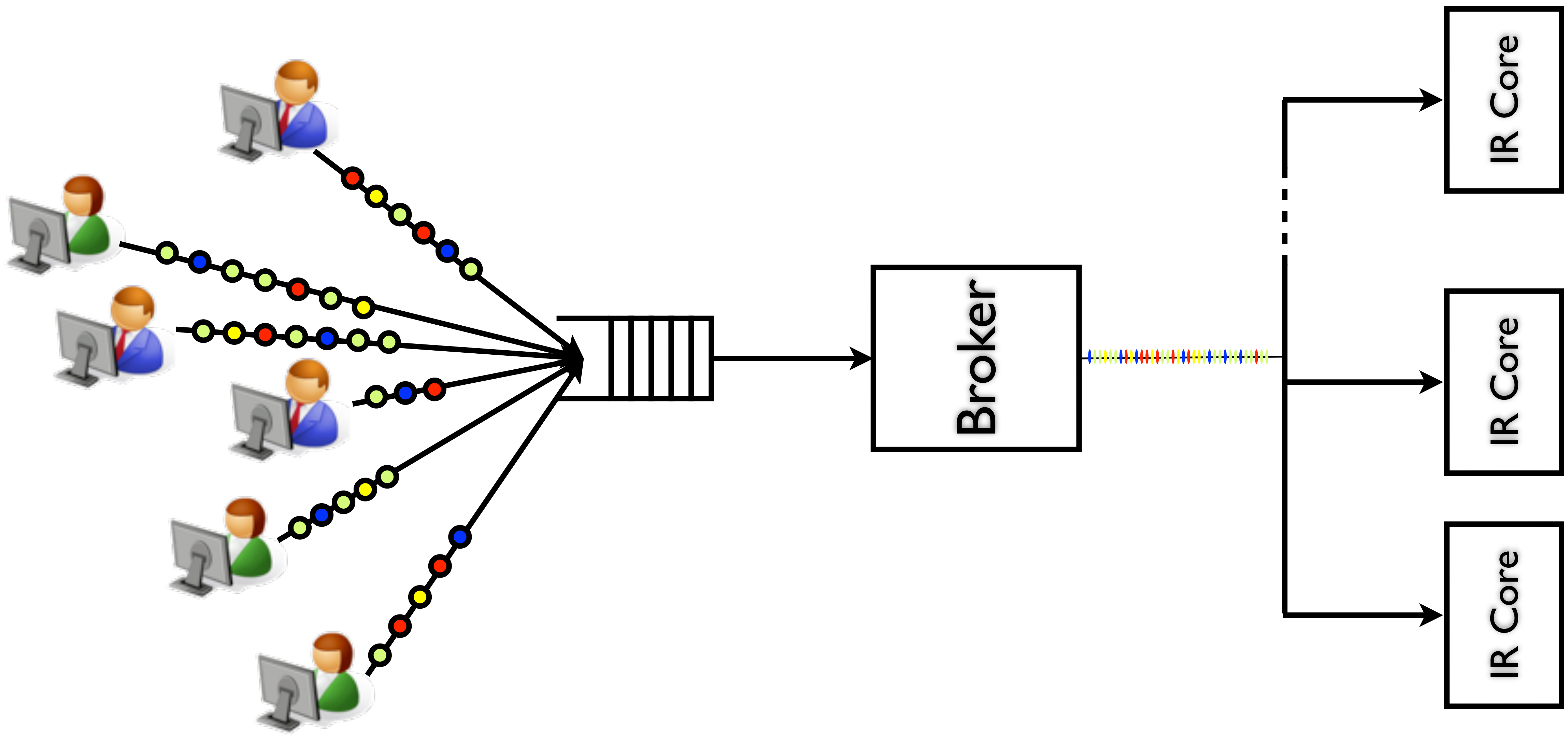
Sketching a Distributed Search Engine



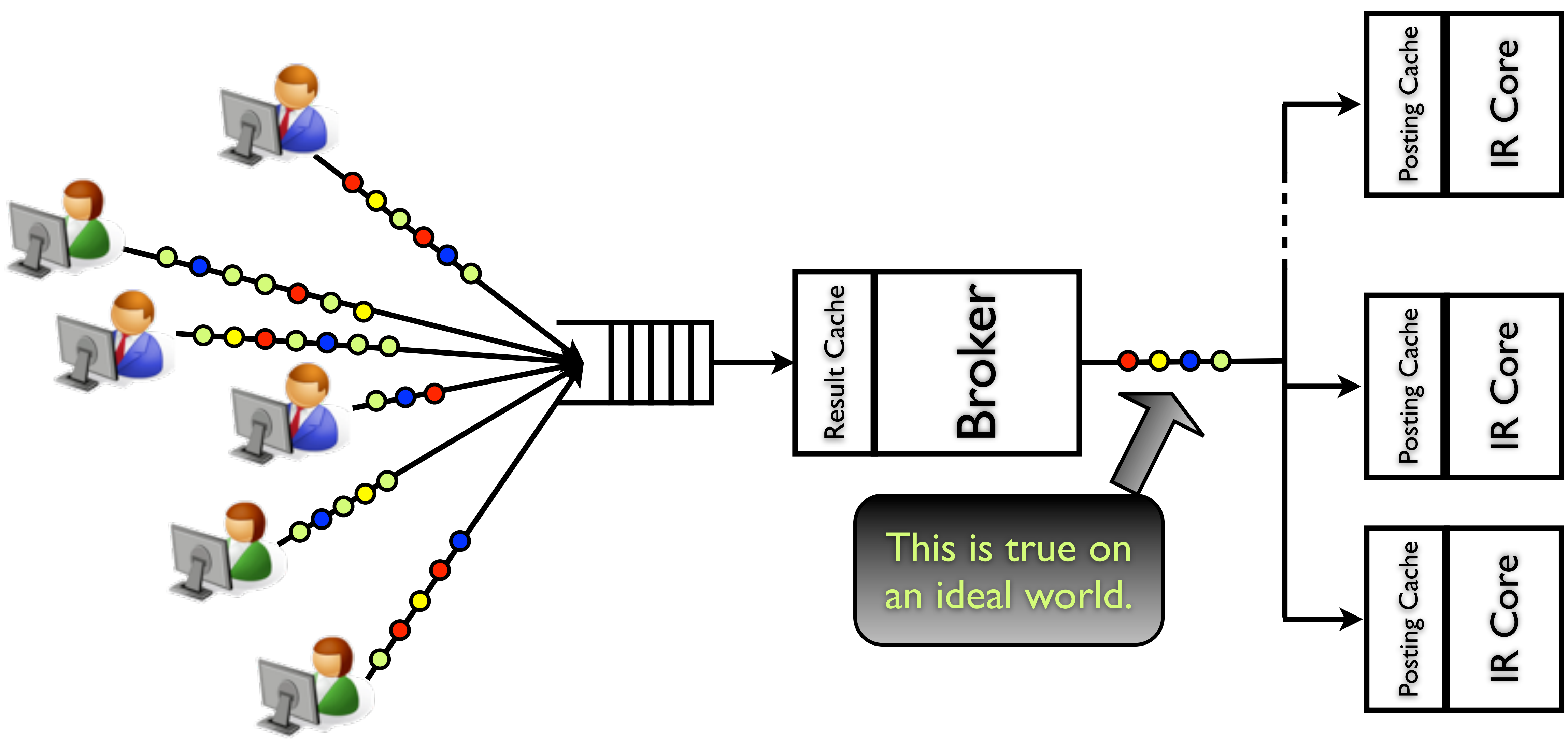
Caching in General



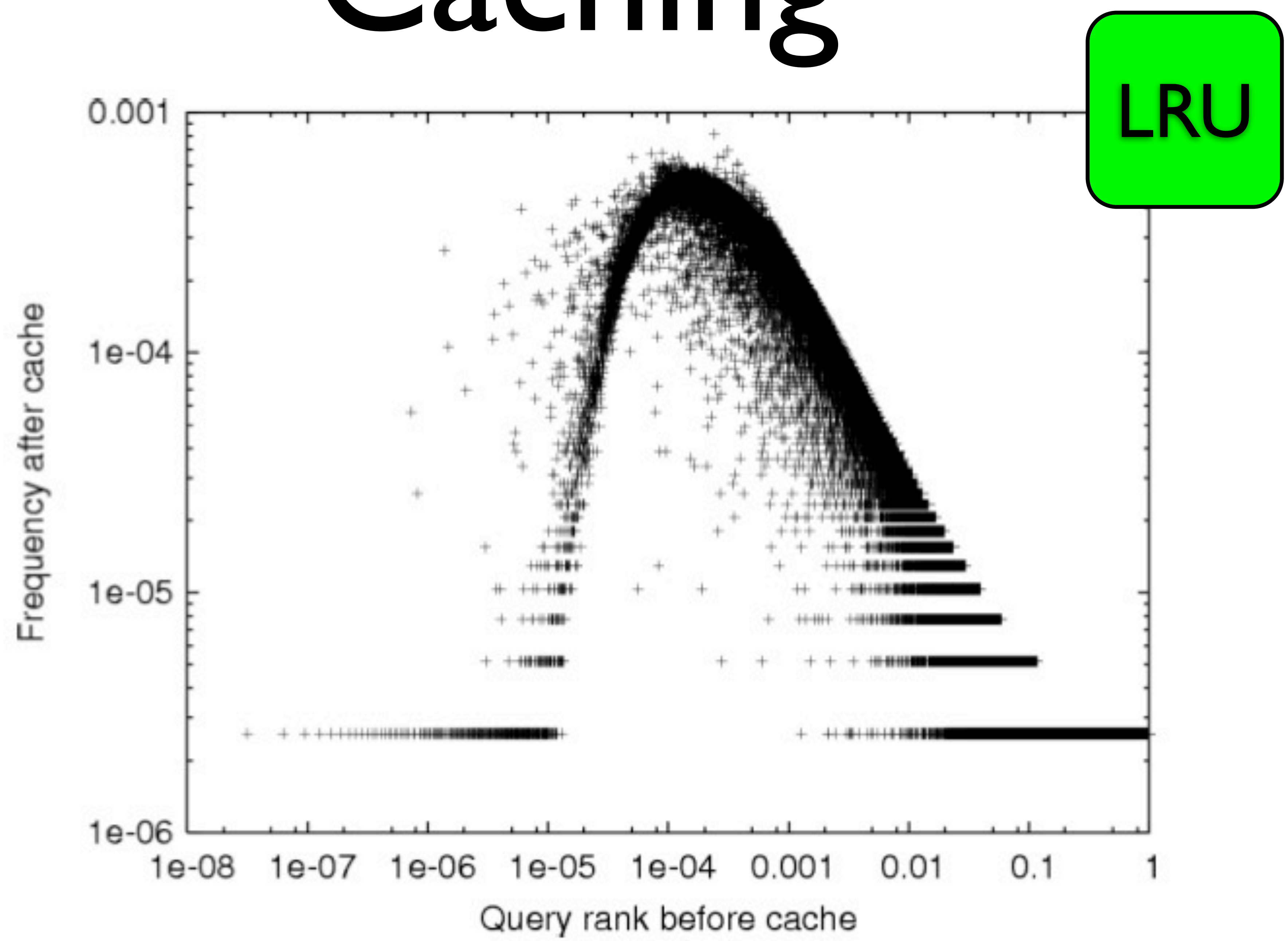
Caching



Caching


Result Cache
Broker
This is true on an ideal world.
Posting Cache
IR Core
Posting Cache
IR Core
Posting Cache
IR Core

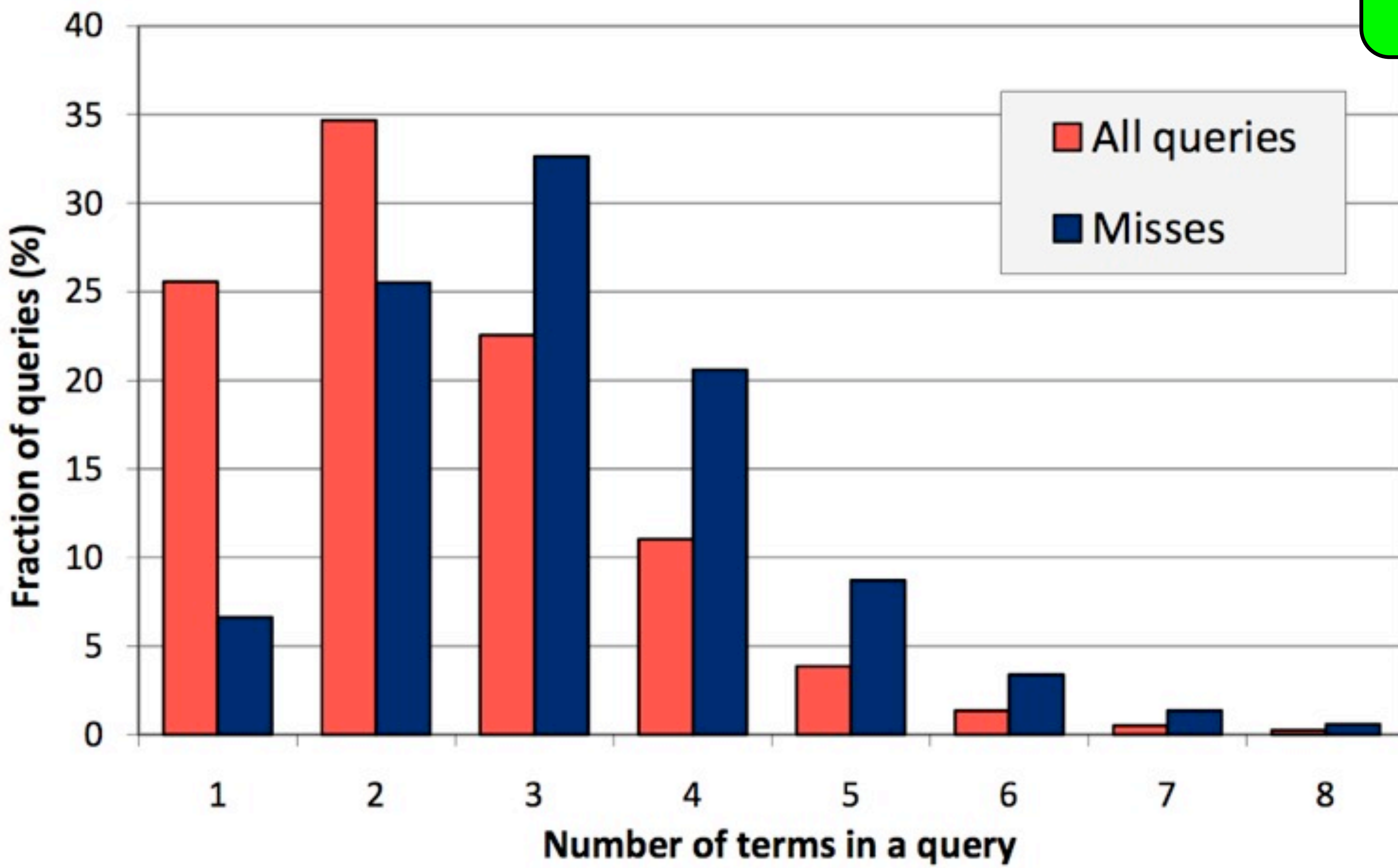
Filtering Effect of Caching



R. Baeza-Yates, A. Gionis, F. P. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri, "**Design trade-offs for search engine caching**," ACM Trans. Web, vol. 2, no. 4, pp. 1–28, 2008.

Filtering Effect of Caching

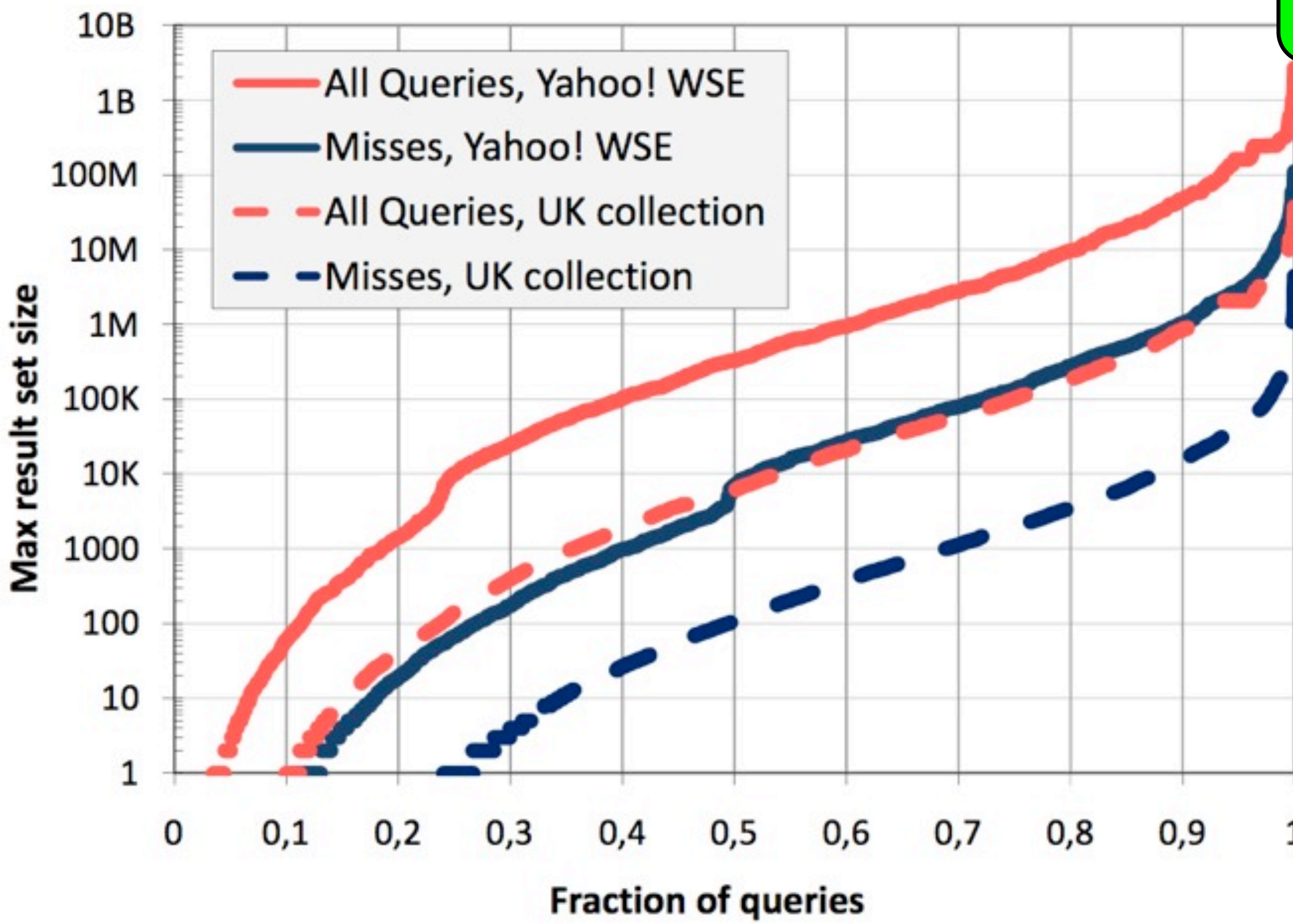
LRU



Skobeltsyn, G., Junqueira, F., Plachouras, V., and Baeza-Yates, R., "**ResIn: a combination of results caching and index pruning for high-performance web search engines**," SIGIR 2008, pp 131-138.

Filtering Effect of Caching

LRU



Skobeltsyn, G., Junqueira, F., Plachouras, V., and Baeza-Yates, R., "**ResIn: a combination of results caching and index pruning for high-performance web search engines**," SIGIR 2008, pp 131-138.

Caching Performance Evaluation

- **Hit-Ratio**: i.e. how many times the cache is useful
- **Query Throughput**: i.e. the number of queries the cache can serve in a second

- But... what really impacts on caching performance?

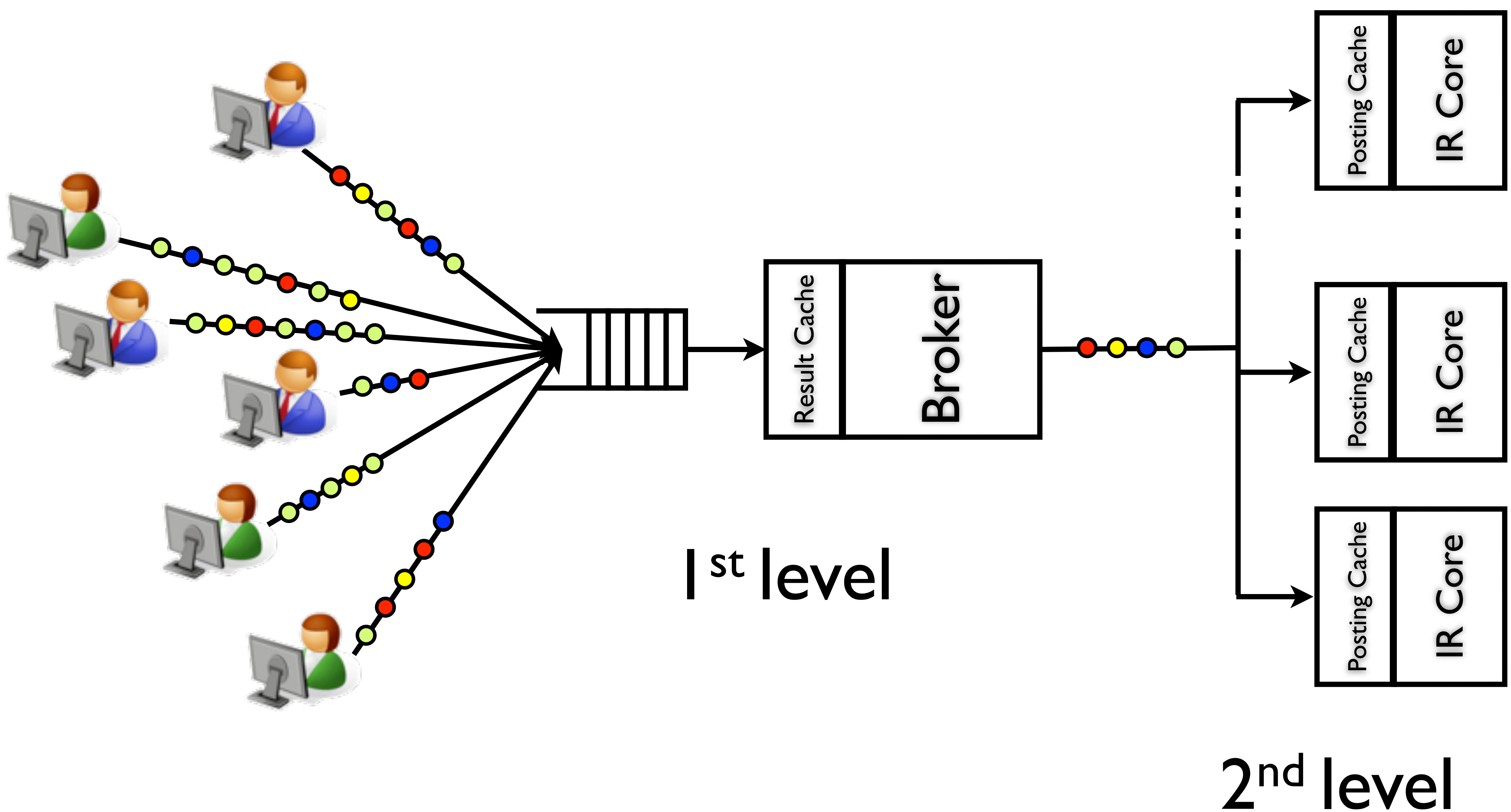
Caching for Search Engines Workloads

- Caching Architectures:
  - Two-Level Caching
  - Three-Level Caching
- Caching Policies
  - PDC
  - SDC
  - AC

Two-Level Caching

- Firstly studied in:
  - Saraiva, P. C., Silva de Moura, E., Ziviani, N., Meira, W., Fonseca, R., and Riberio-Neto, B. 2001. **Rank-preserving two-level caching for scalable search engines**. In Proceedings of ACM SIGIR '01. ACM, New York, NY, 51-58.
- Further analyzed in:
  - Baeza-Yates, R., Gionis, A., Junqueira, F. P., Murdock, V., Plachouras, V., and Silvestri, F. 2008. **Design trade-offs for search engine caching**. ACM Trans. Web 2, 4 (Oct. 2008), 1-28.

Two-Level Caching



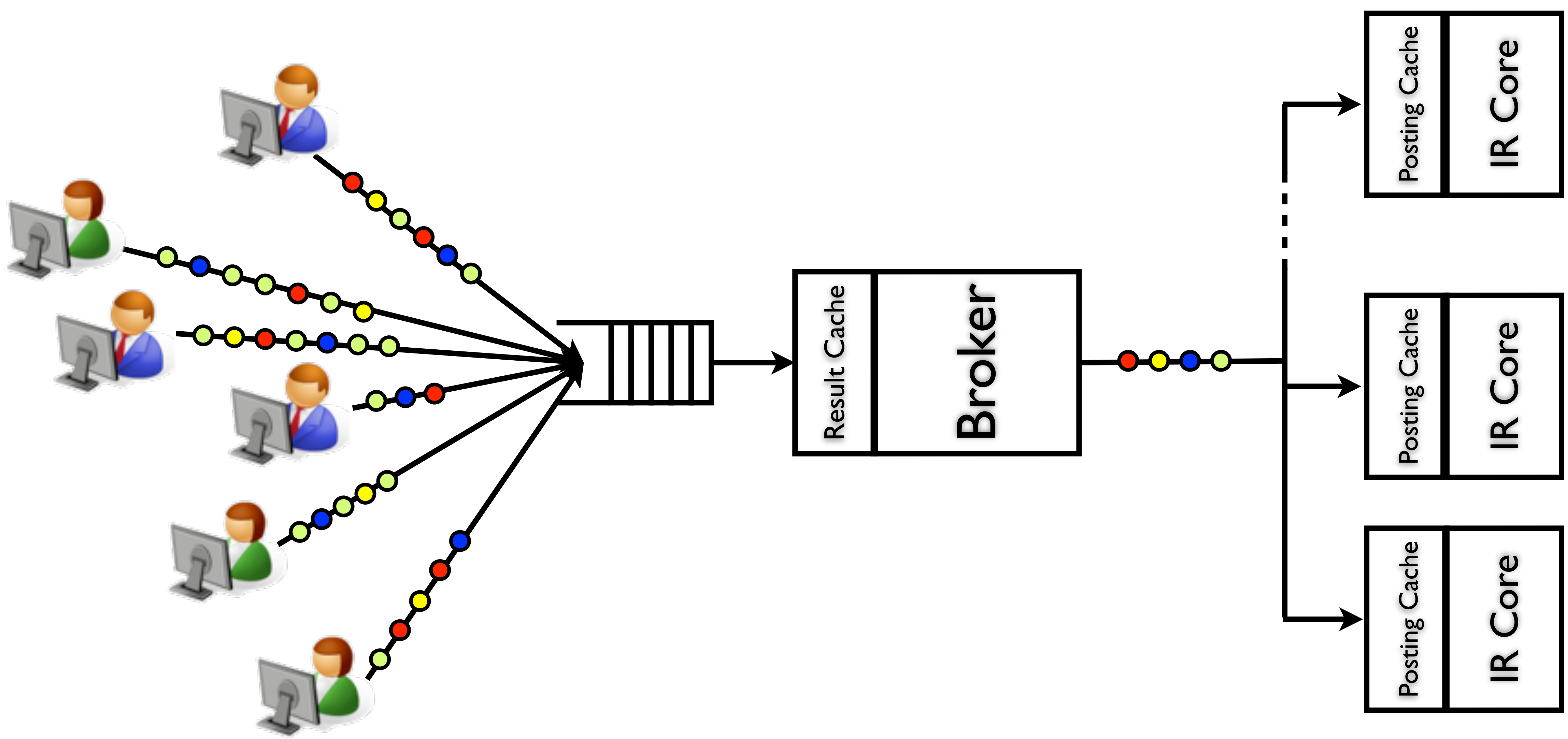
Three-level Caching

- Adds one level between results and posting lists cache.
- Usually stores frequently occurring pairs of terms.
  - Long, X. and Suel, T. 2005. **Three-level caching for efficient query processing in large Web search engines**. In Proceedings of the 14th international Conference WWW '05. 257-266.
  - Skobeltsyn, G., Junqueira, F., Plachouras, V., and Baeza-Yates, R. 2008. **ResIn: a combination of results caching and index pruning for high-performance web search engines**. In Proceedings of the 31st Annual international ACM SIGIR '08. 131-138.

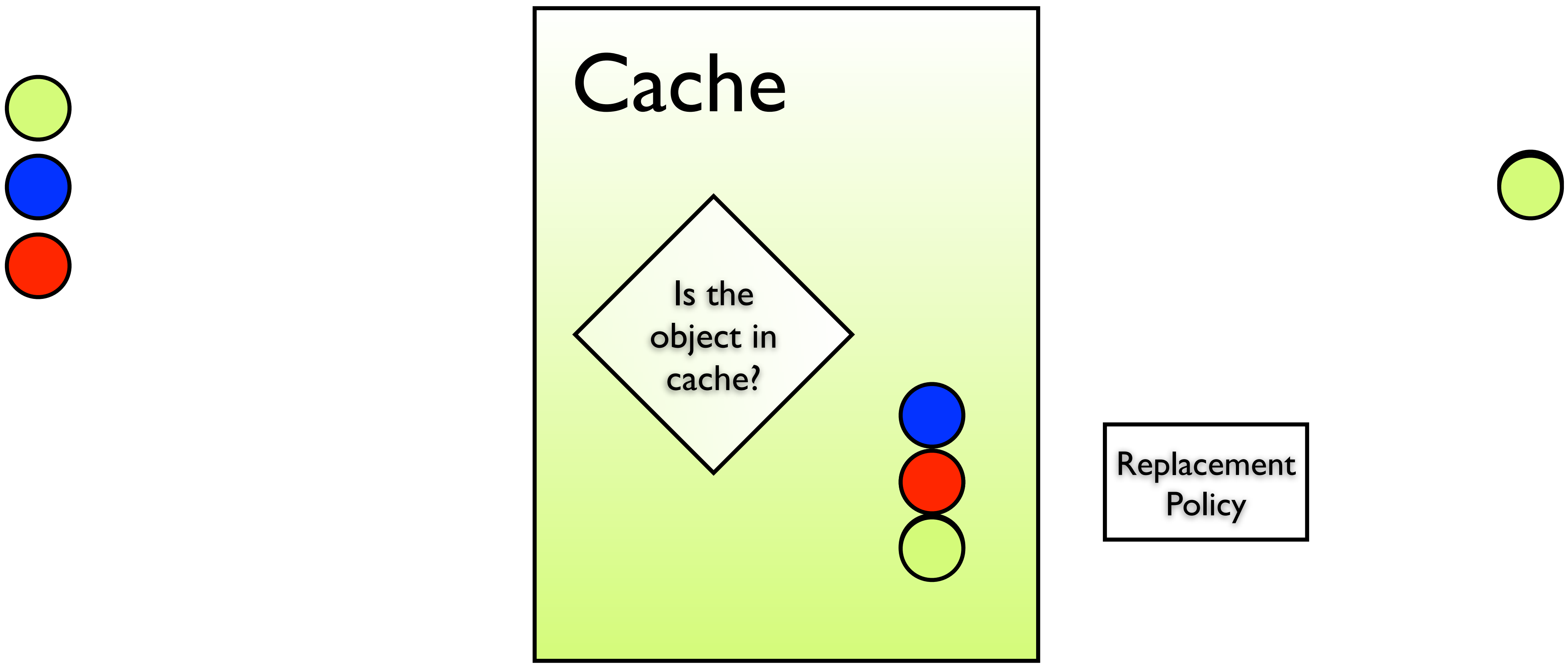
“Things” to Cache in Search Engines

- Results
  - in answer to a user query
- Posting Lists
  - e.g. for the query “new york” cache the posting lists for term new and for term york
- Partial queries
  - cache subqueries, e.g. for “new york times” cache only “new york”

Cache Replacement Policies



Cache Replacement Policies

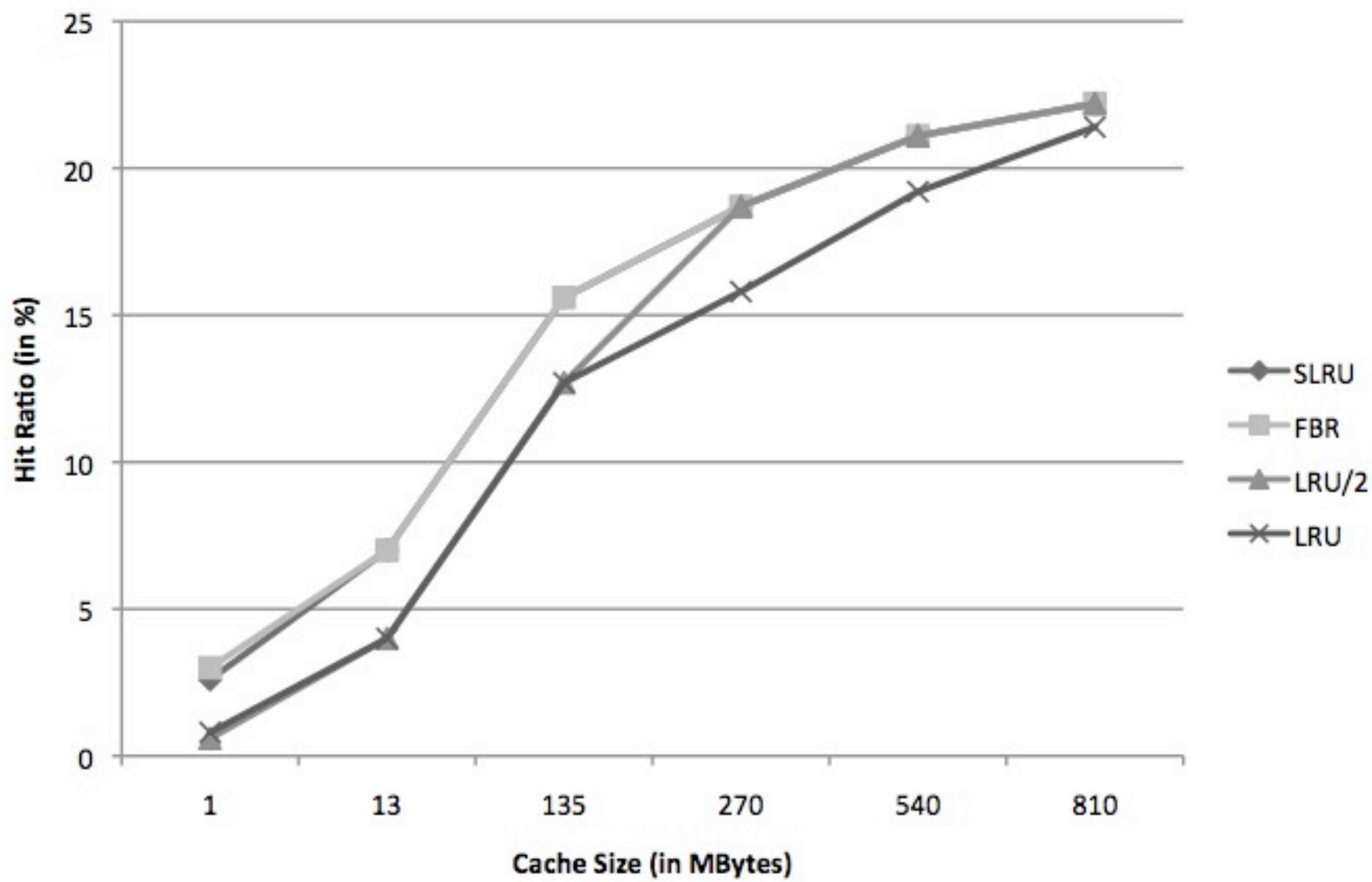


Traditional Replacement Policies

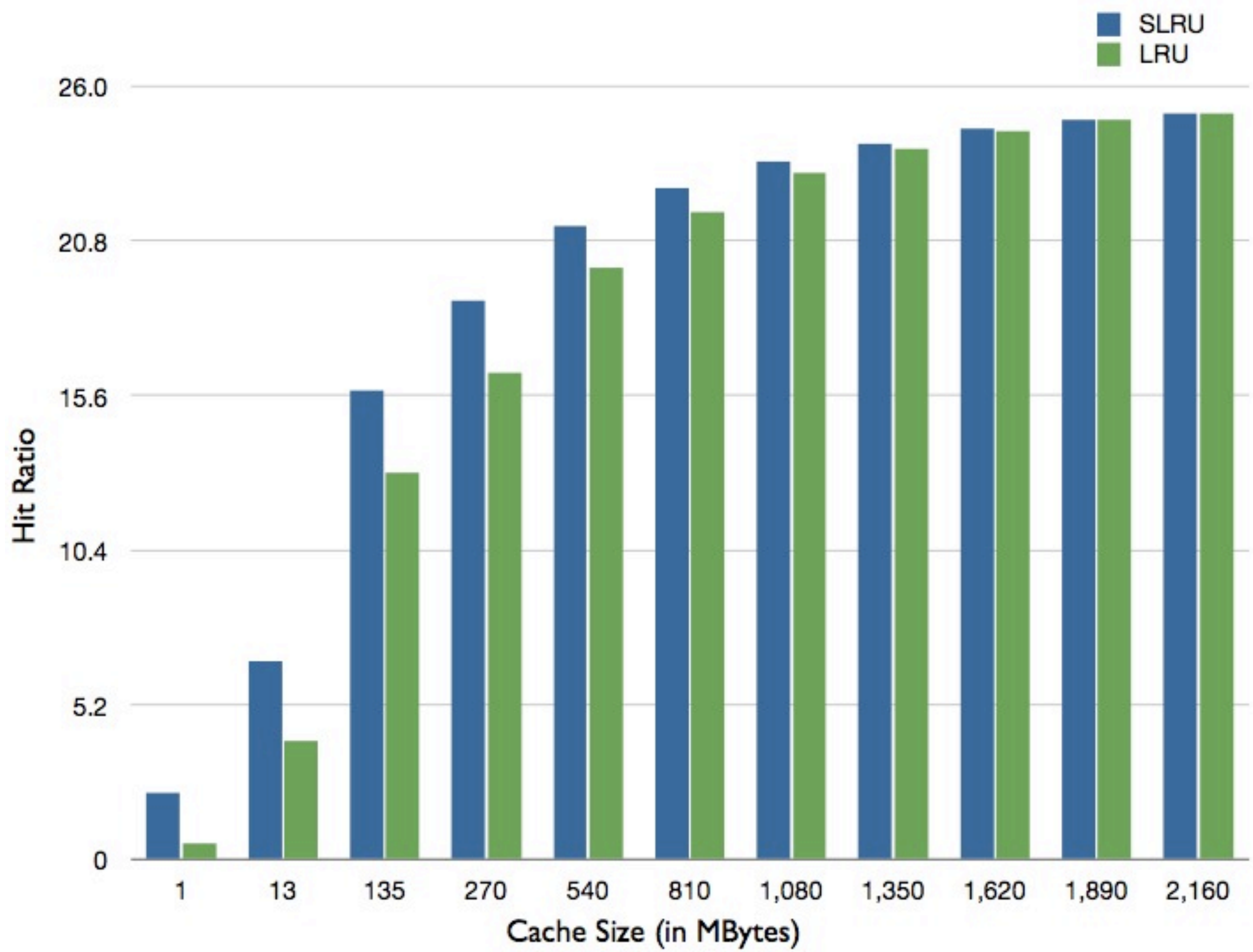
- LRU
- LFU
- SLRU
- ...

T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "**Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data**," ACM Trans. Inf. Syst., vol. 24, no. 1, pp. 51–78, 2006.

Hit Ratios on Excite



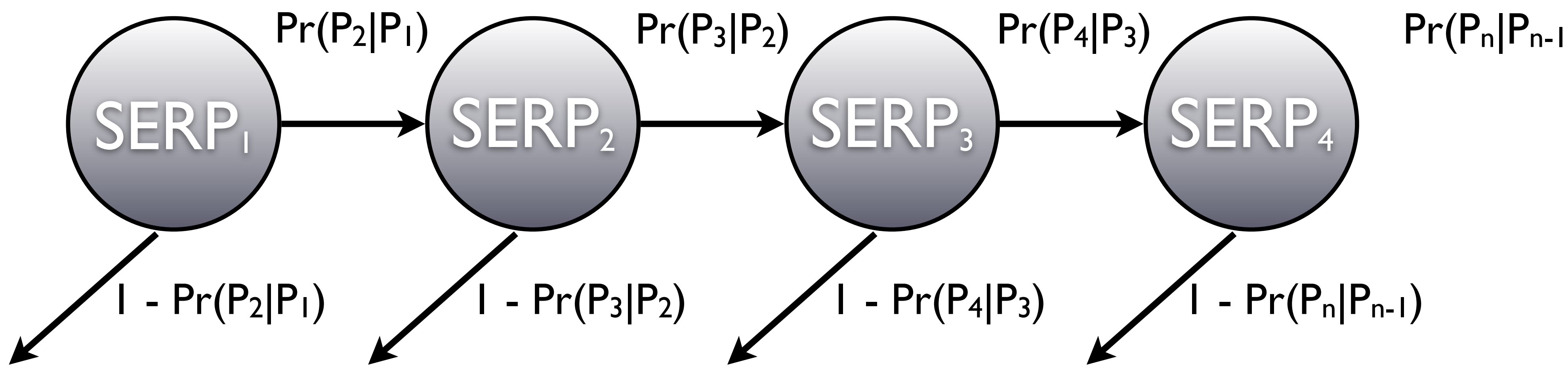
SLRU vs. LRU on Excite



Search Engine Tailored Policies

- PDC
  - Probability Driven Caching
- SDC
  - Static Dynamic Caching
- AC
  - Admission Control

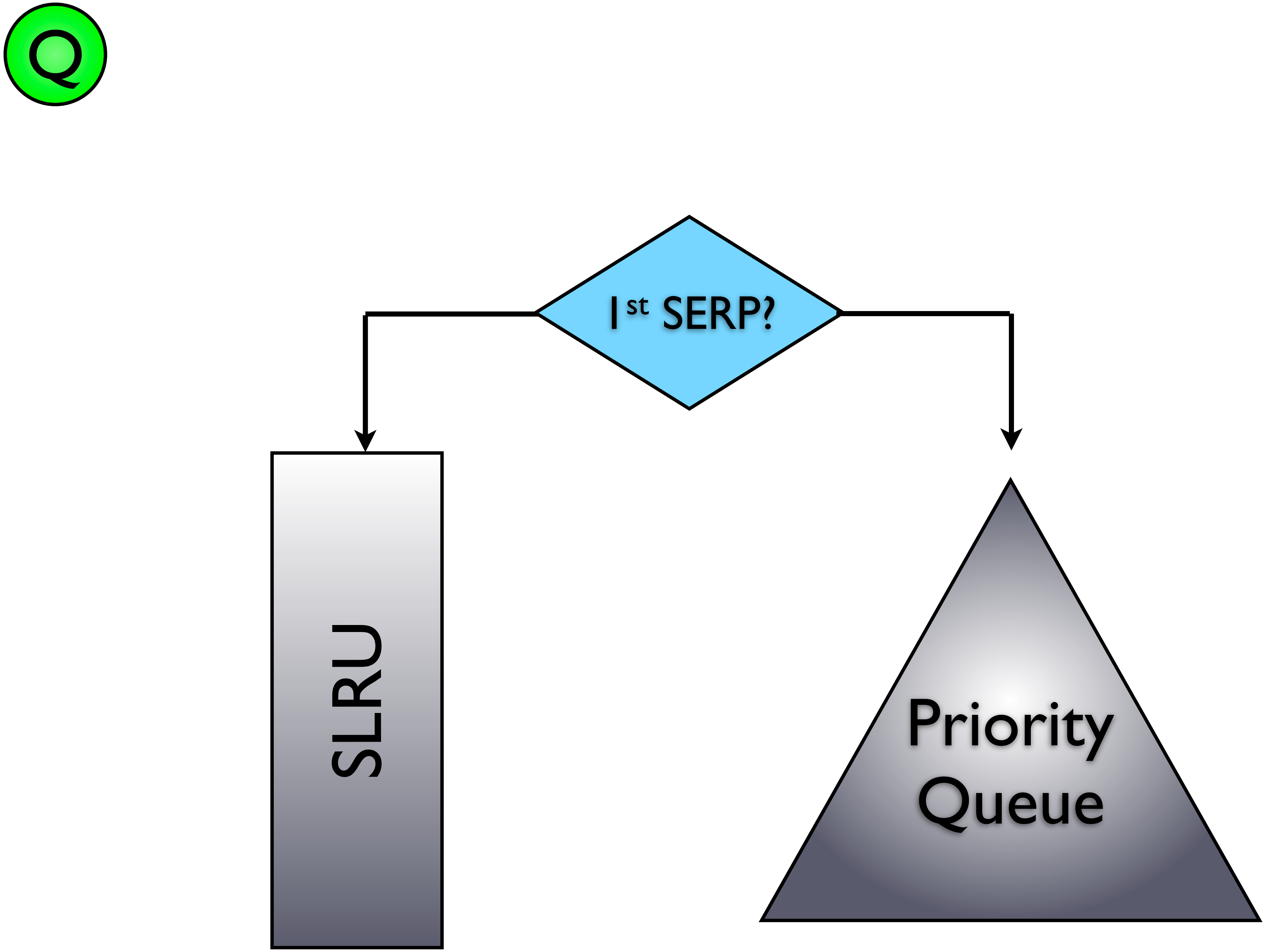
PDC



- IDEA: design a policy tailored over users' behavior on search pages
- With high probability users do not go beyond the first page of results
- For some query users browse many result pages.

Lempel, R. and Moran, S. 2003. **Predictive caching and prefetching of query results in search engines.** In Proceedings of WWW '03. 19-28.

PDC



PDC Priorities

- Priorities are assigned using an approximation of the Markovian SERP request model
- Each SERP different from the first one has a priority computed on historical queries (query log)
  - we cache pages that has follow-up queries more likely to be submitted. Why?

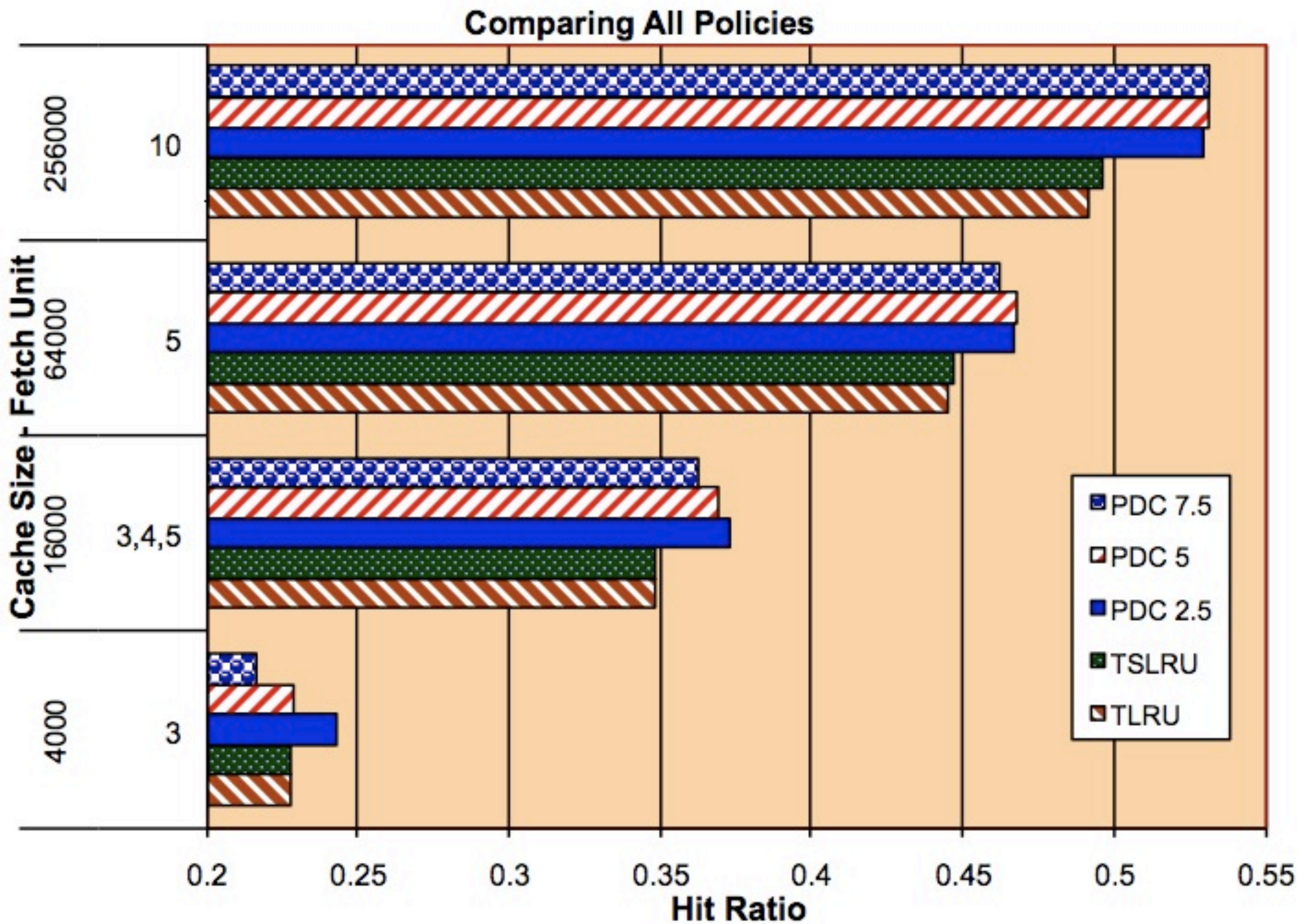
Lempel, R. and Moran, S. 2003. **Predictive caching and prefetching of query results in search engines.** In Proceedings of WWW '03. 19-28.

PDC and Prefetching

- in PDC results are organized according to "*Fetch Units*"
- When SERP i is requested for a query Q, we look up the cache to probe its presence.
- If i is not cached, we request SERP i, i + 1, ..., i + f
- That is we prefetch f SERPs.
- The fetch unit is of size f.

Lempel, R. and Moran, S. 2003. **Predictive caching and prefetching of query results in search engines.** In Proceedings of WWW '03. 19-28.

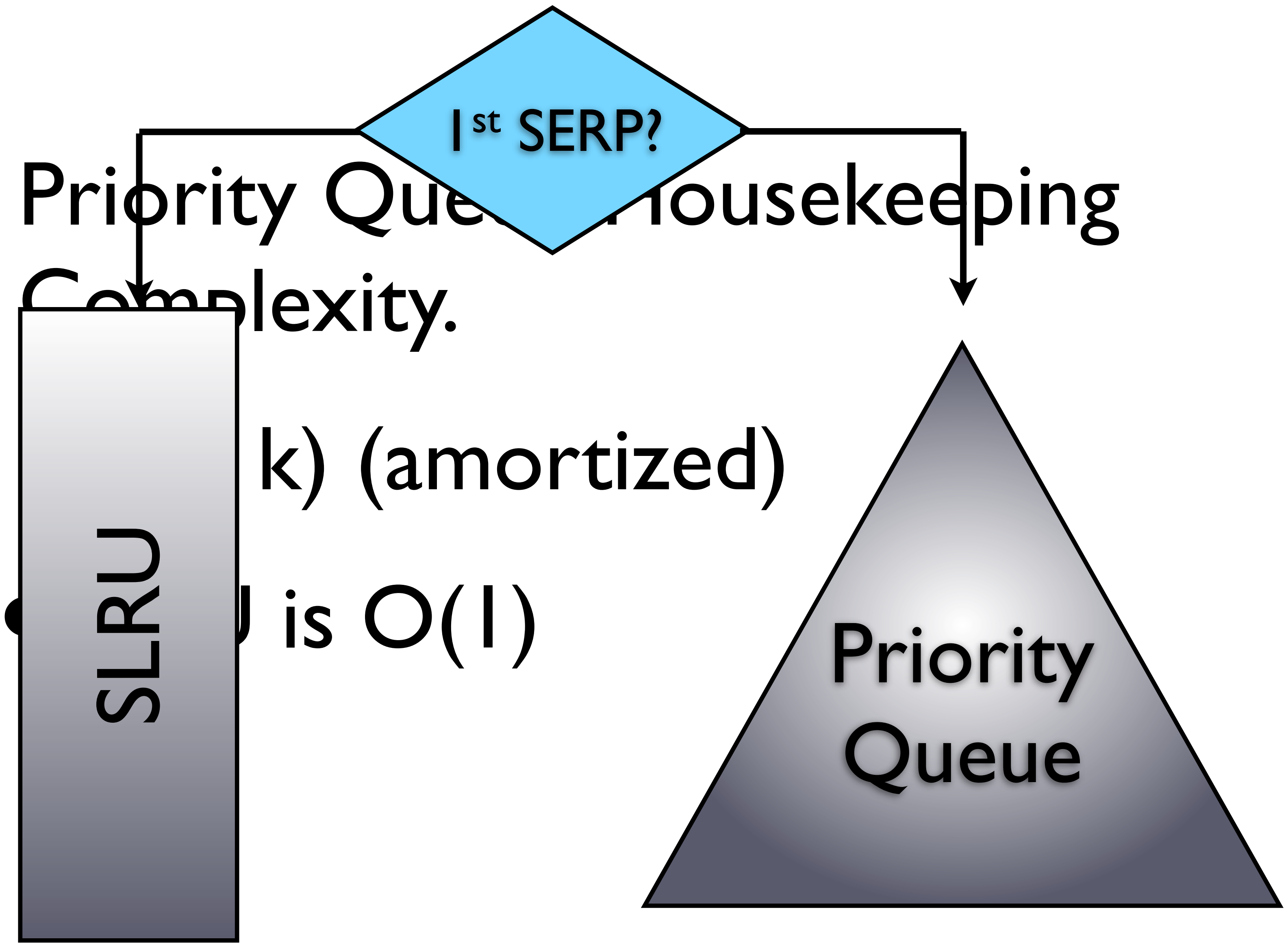
PDC Results



Lempel, R. and Moran, S. 2003. **Predictive caching and prefetching of query results in search engines.** In Proceedings of WWW '03. 19-28.

PDC's Main Drawback

- Priority Que… ousekeeping Complexity.
- … k) (amortized)
  - … U is O(1)



PDC's Main Lessons Learned

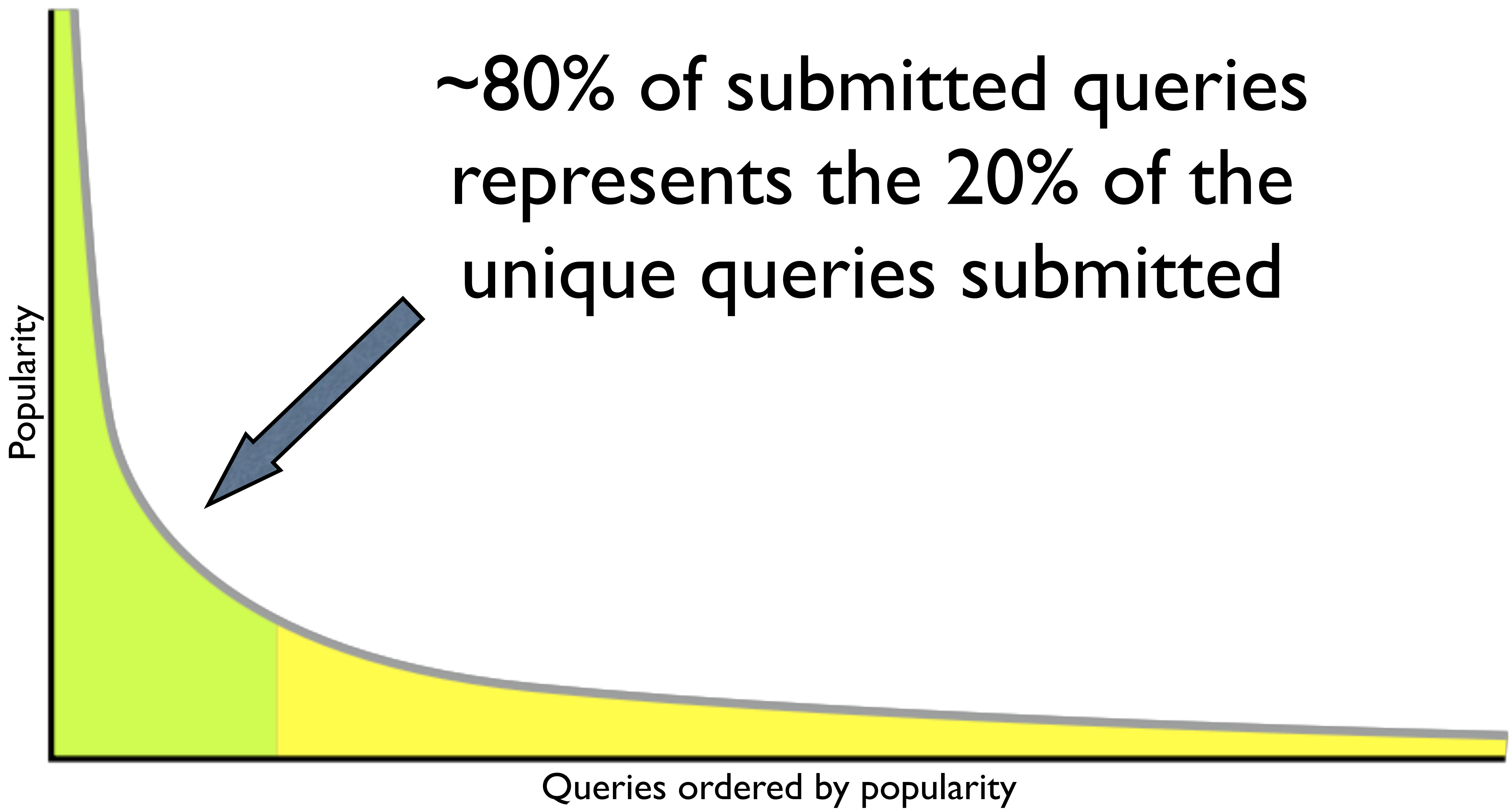
- Hit ratio benefits a lot from the use of historical data
- Prefetching helps a lot!
- Differently from previous caching policies, PDC not necessarily caches every submitted queries!!!

Lempel, R. and Moran, S. 2003. **Predictive caching and prefetching of query results in search engines.** In Proceedings of WWW '03. 19-28.

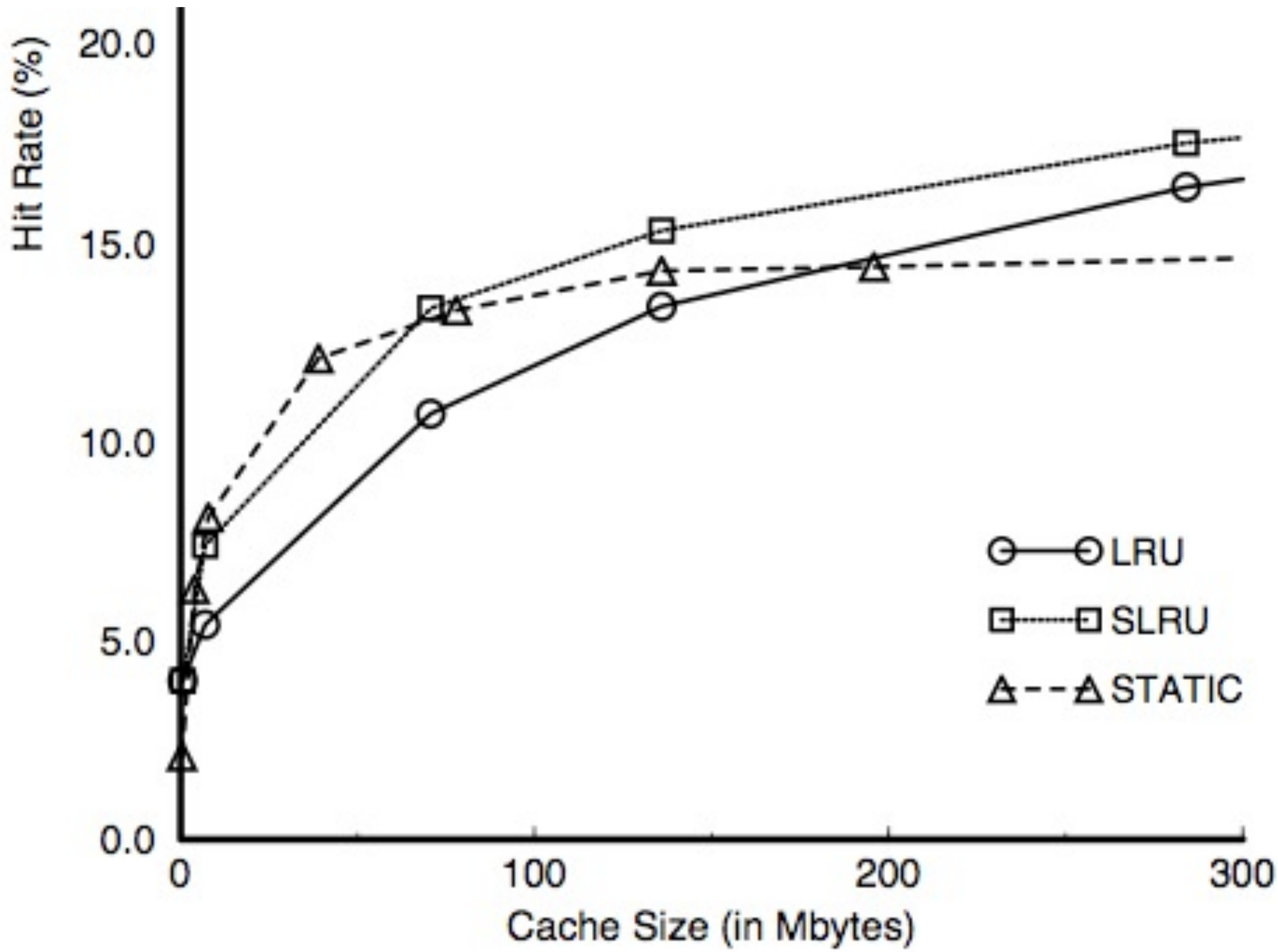
Overcoming PDC Complexity

- PDC uses query logs to estimate the likelihood of follow-up queries.
- Why not using query logs to estimate likelihood of resubmitting a query.
- Catching the head of the long tail distribution we might obtain high hit ratios

That is...


~80% of submitted queries represents the 20% of the unique queries submitted
Popularity
Queries ordered by popularity

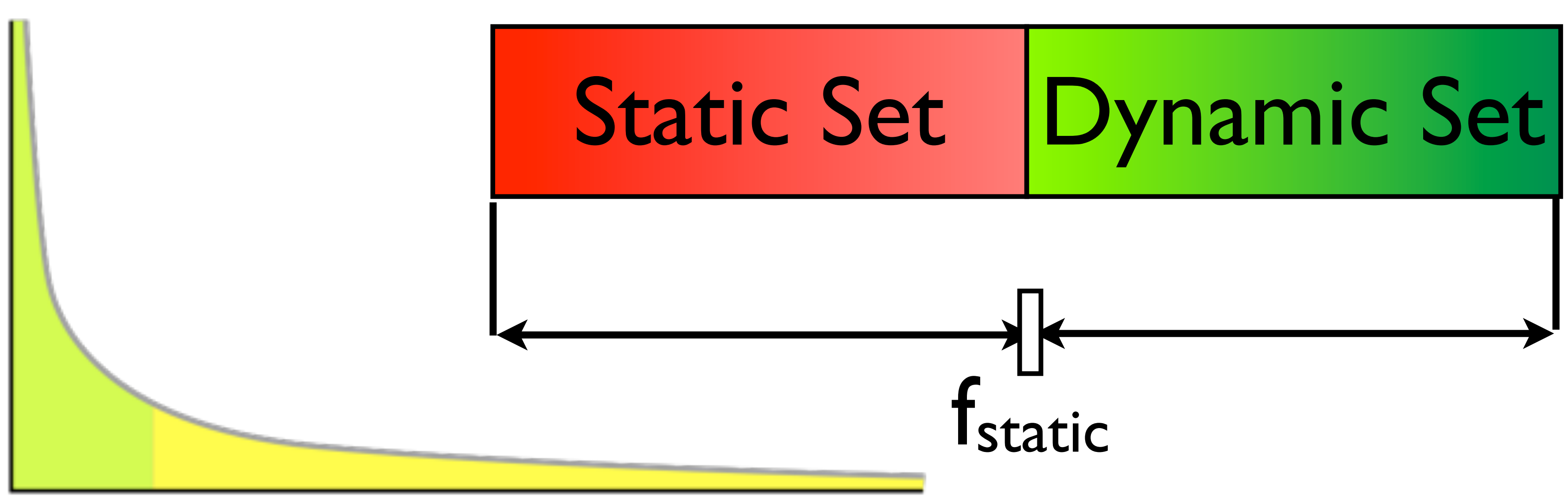
But...



Hit Rate (%)
20.0
15.0
10.0
5.0
0.0
0
100
200
300
Cache Size (in Mbytes)
LRU
SLRU
STATIC

Static Dynamic Caching

- SDC (Static Dynamic Caching) adds to the classical static caching schema a dynamically managed section.
- The idea:



- LRU
- SLRU
- PDC
- ...

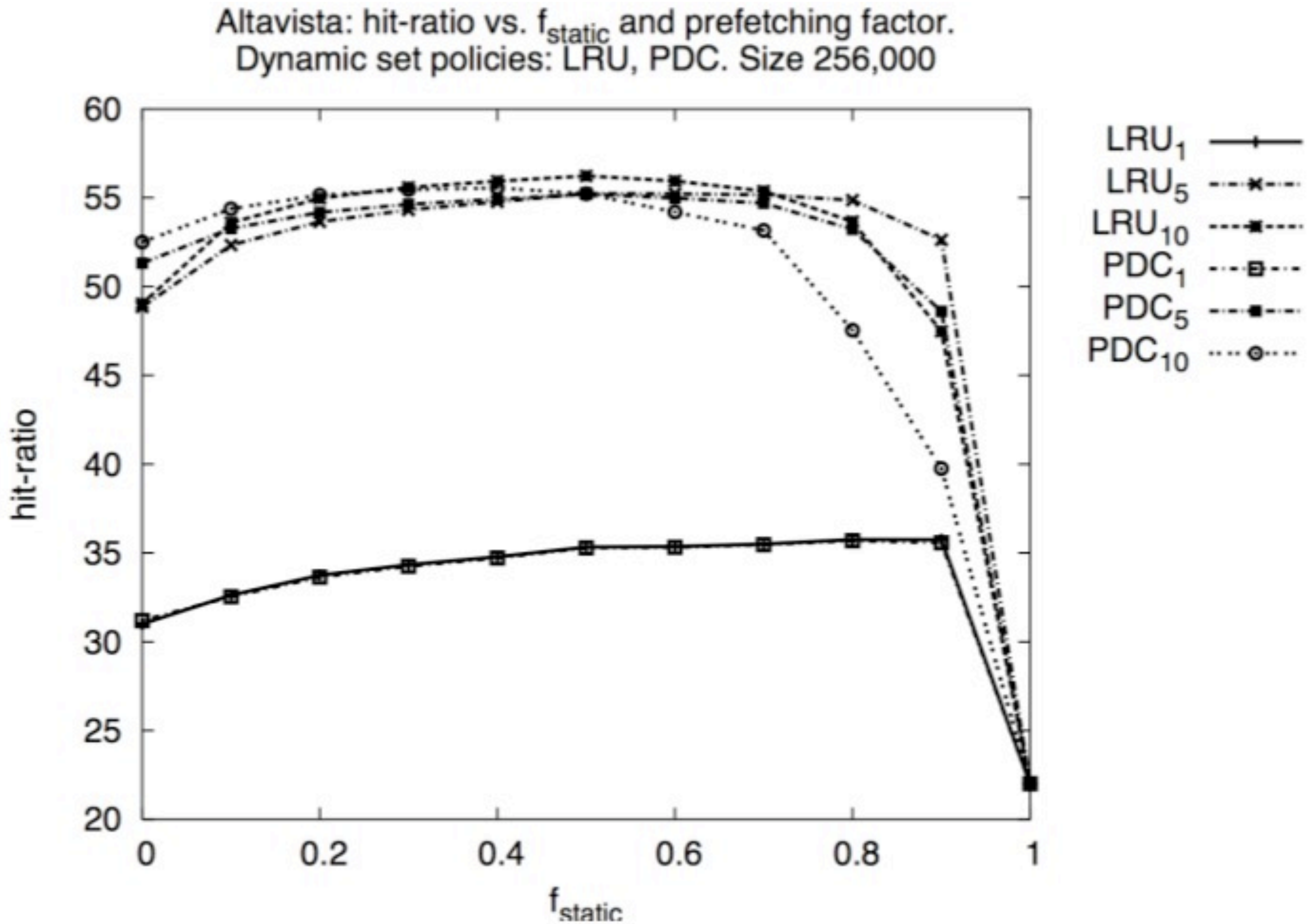
T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "**Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data**," ACM Trans. Inf. Syst., vol. 24, no. 1, pp. 51–78, 2006.

SDC and Prefetching

- SDC adopts an “adaptive” prefetching technique:
  - For the first SERP do not prefetch
  - For the follow-up SERPs prefetch f pages

T. Fagni, R. Perego, F. Silvestri, and S. Orlando, “**Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data**,” ACM Trans. Inf. Syst., vol. 24, no. 1, pp. 51–78, 2006.

SDC Hit-Ratios

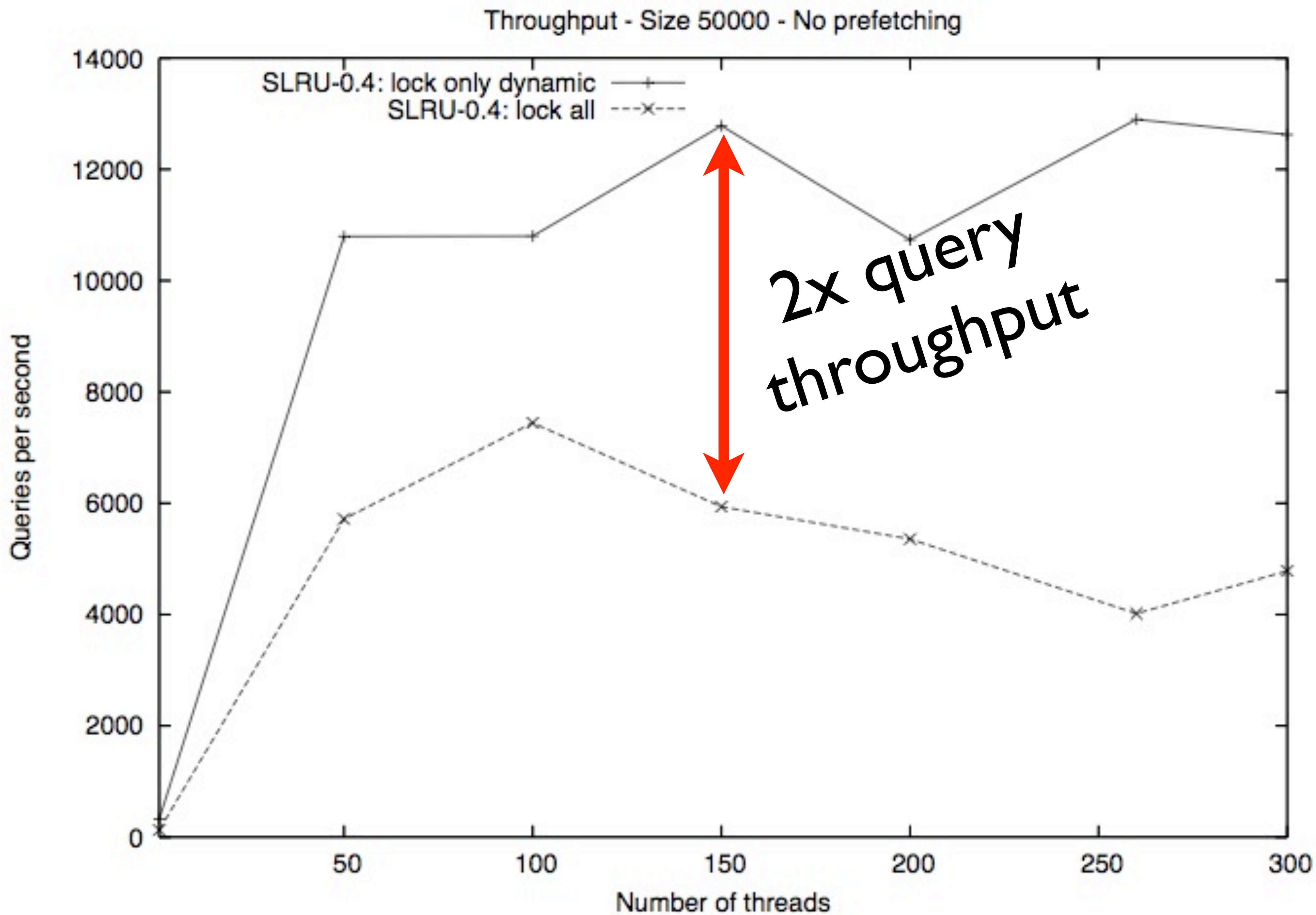


SDC's Main Lessons Learned

- Hit ratio benefits a lot from the use of historical data
- Prefetching helps a lot!
- Static caching alone is not useful, yet...
  - A good combination of a static and a dynamic approach helps a lot!!!

T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "**Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data**," ACM Trans. Inf. Syst., vol. 24, no. 1, pp. 51–78, 2006.

That's *not* All Folks!



T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "**Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data**," ACM Trans. Inf. Syst., vol. 24, no. 1, pp. 51–78, 2006.

Admission Control

- An interesting idea of SDC: frequent queries are cached permanently
- AC of Baeza-Yates et al. generalizes the idea by using two dynamically updated sets:
  - A Controlled Cache (**CC**)
  - An Uncontrolled Cache (**UC**)
- When a new query arrives an admission policy is applied to steer a query to the CC or to the UC.
- If the query is likely to be seen in the future move it to CC, otherwise send it to UC.

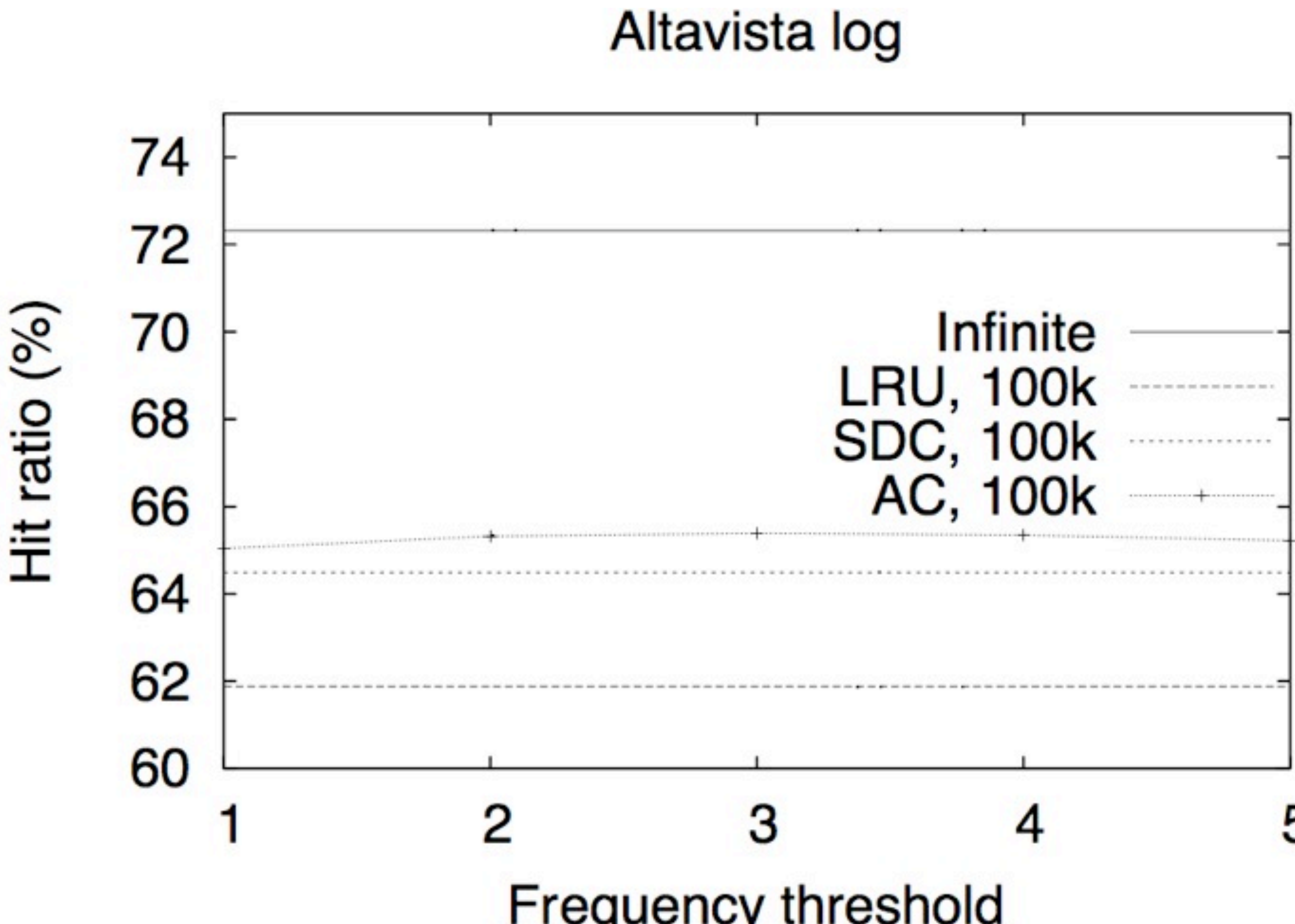
Ricardo Baeza-Yates, Flavio Junqueira, Vassilis Plachouras, and Hans Friedrich Witschel, "**Admission Policies for Caches of Search Engine Results**," SPIRE 2007, 74-85.

Admission Policy

- Makes use of features, e.g.:
  - Stateful features:
    - *PastF*: the frequency of the query in the (relatively recent) past
  - Stateless features:
    - *LenC*: the length of the query in characters
    - *LenW*: the length of the query in words

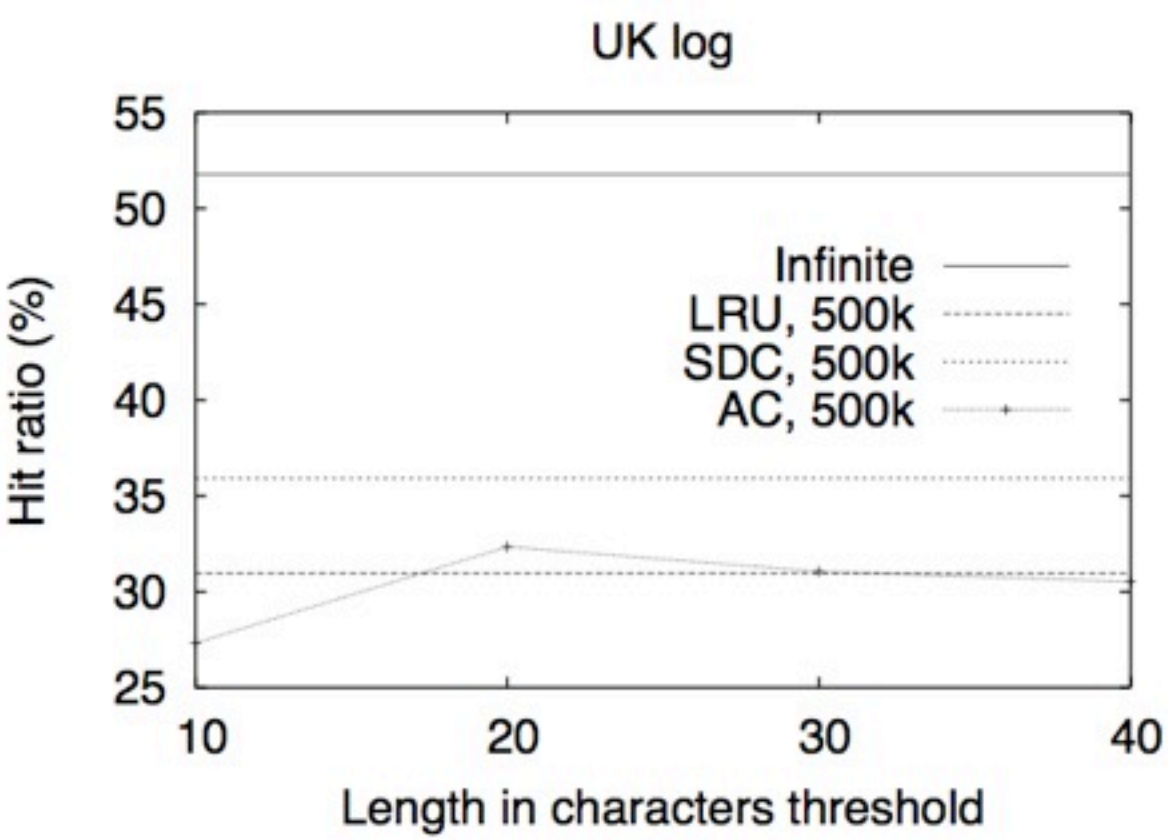
Ricardo Baeza-Yates, Flavio Junqueira, Vassilis Plachouras, and Hans Friedrich Witschel, "**Admission Policies for Caches of Search Engine Results**," SPIRE 2007, 74-85.

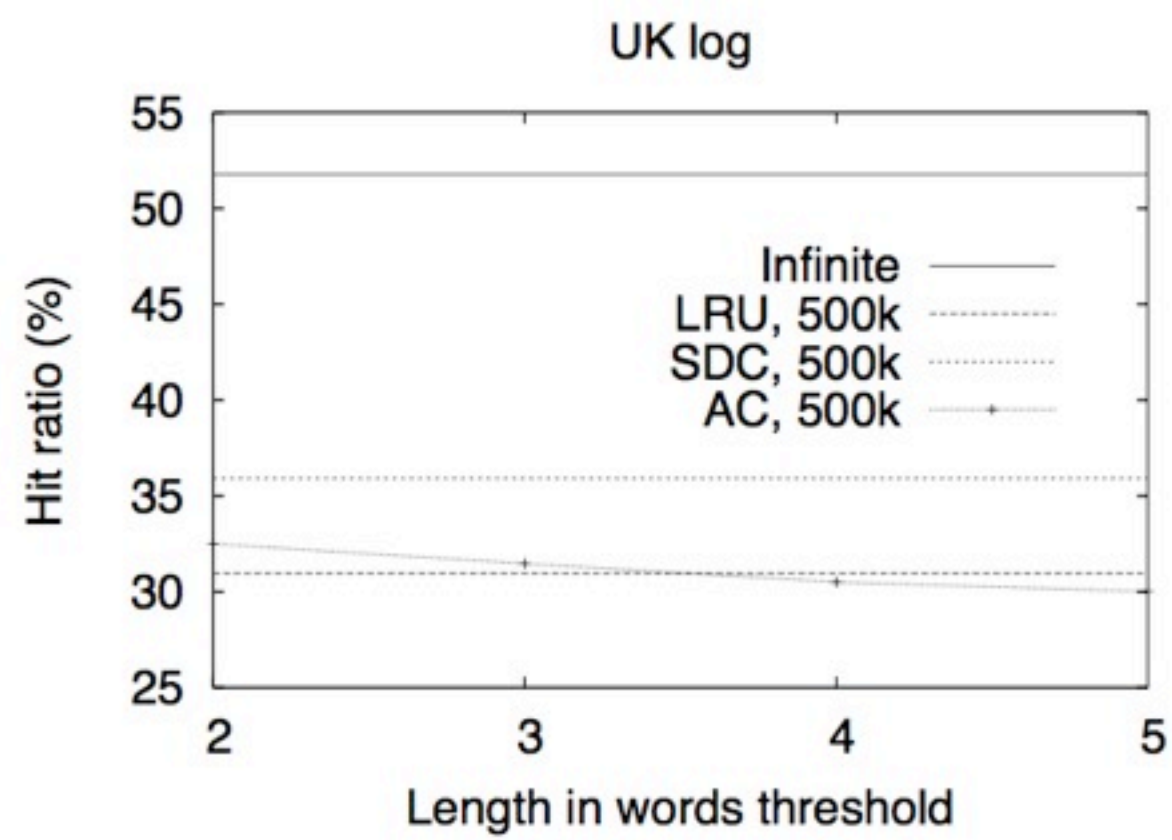
Hit-Ratio Results (Past1-5)



Ricardo Baeza-Yates, Flavio Junqueira, Vassilis Plachouras, and Hans Friedrich Witschel, "**Admission Policies for Caches of Search Engine Results**," SPIRE 2007, 74-85.

Hit-Ratio Results (LenC- LenW)





Ricardo Baeza-Yates, Flavio Junqueira, Vassilis Plachouras, and Hans Friedrich Witschel, "**Admission Policies for Caches of Search Engine Results**," SPIRE 2007, 74-85.

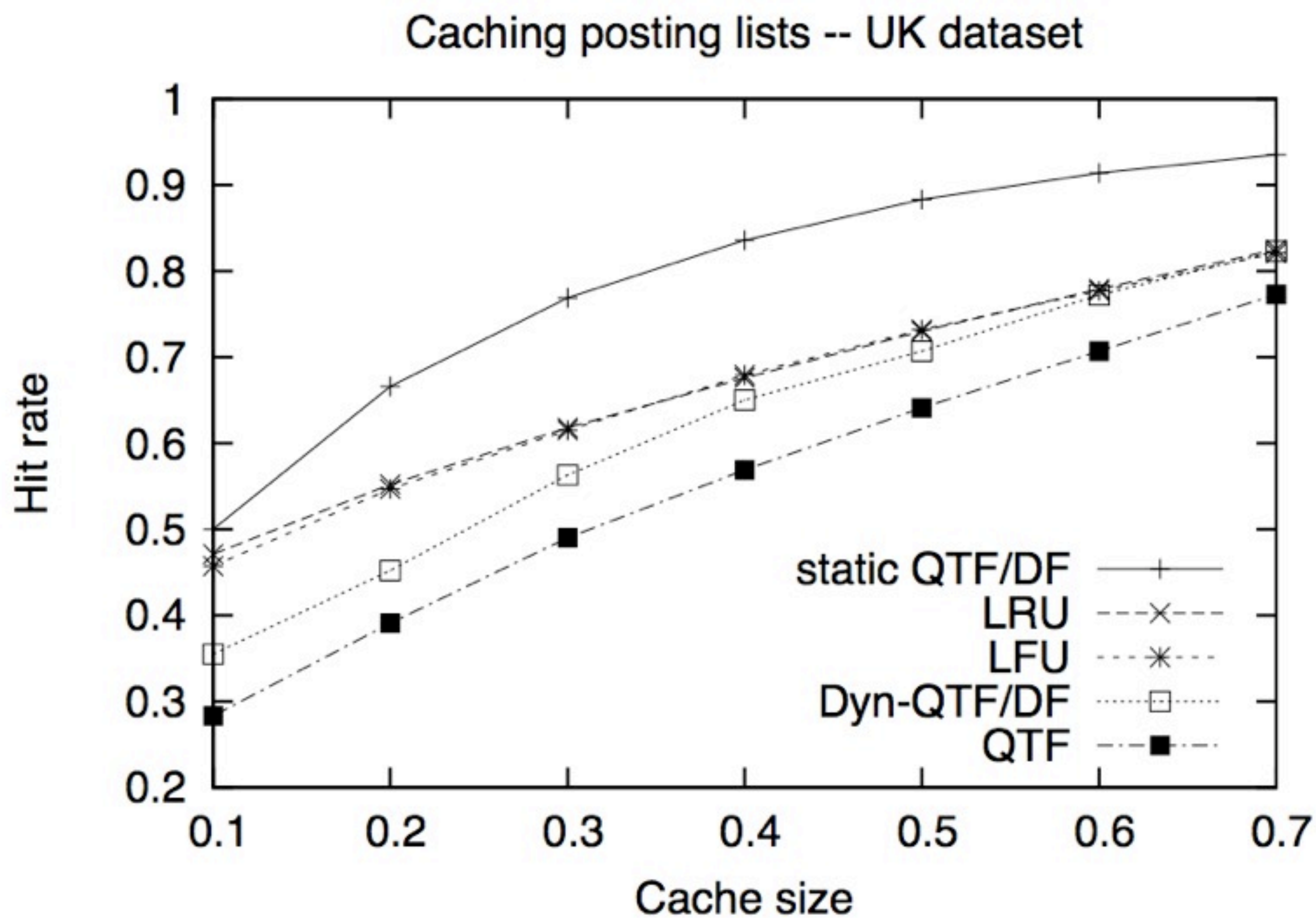
Caching Posting Lists

- SERP size is fixed
- Posting lists have different lengths.
- Posting list caching techniques adopt policies sensitive to list sizes.

$Q_{TF}D_F$ Policy

- Idea:
    - suppose you have 10 free slots and 3 postings lists to cache l_1, l_2, and l_3. l_1 appears 10 times and it is long 6 postings, l_2 and l_3 appear 6 times each and are long 5 postings.
- Traditional frequency-only-based policies will choose to cache l_1 filling up 6 slots and not leaving space for any of the two other lists.
- $Q_{TF}D_F$ decides to cache l_2 and l_3 since they optimize the ratio frequency/size instead of just frequency.
- Results:
    - Traditional static caching has a hit ratio of 10
    - $Q_{TF}D_F$ static policy has a hit ratio of 12

QtfDf Results



SDC-like QTFDF