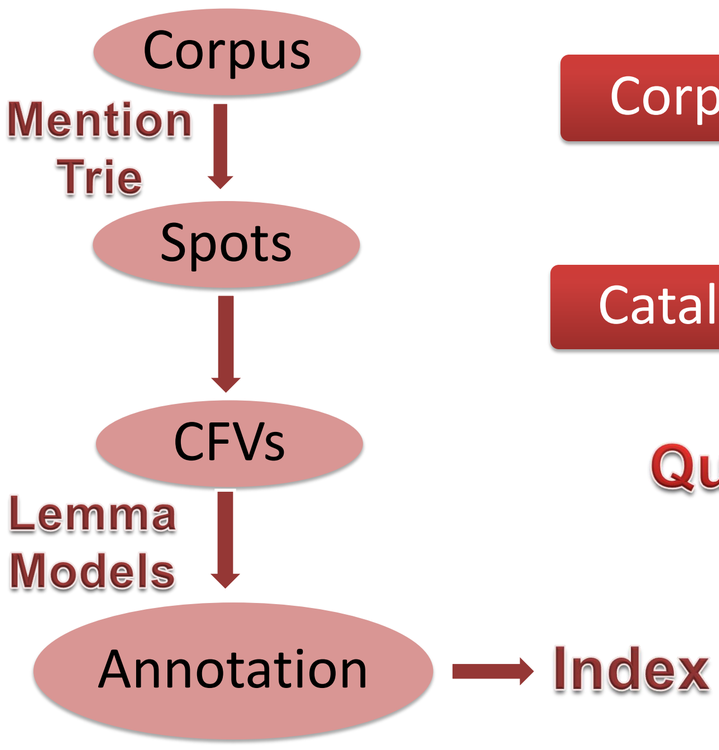
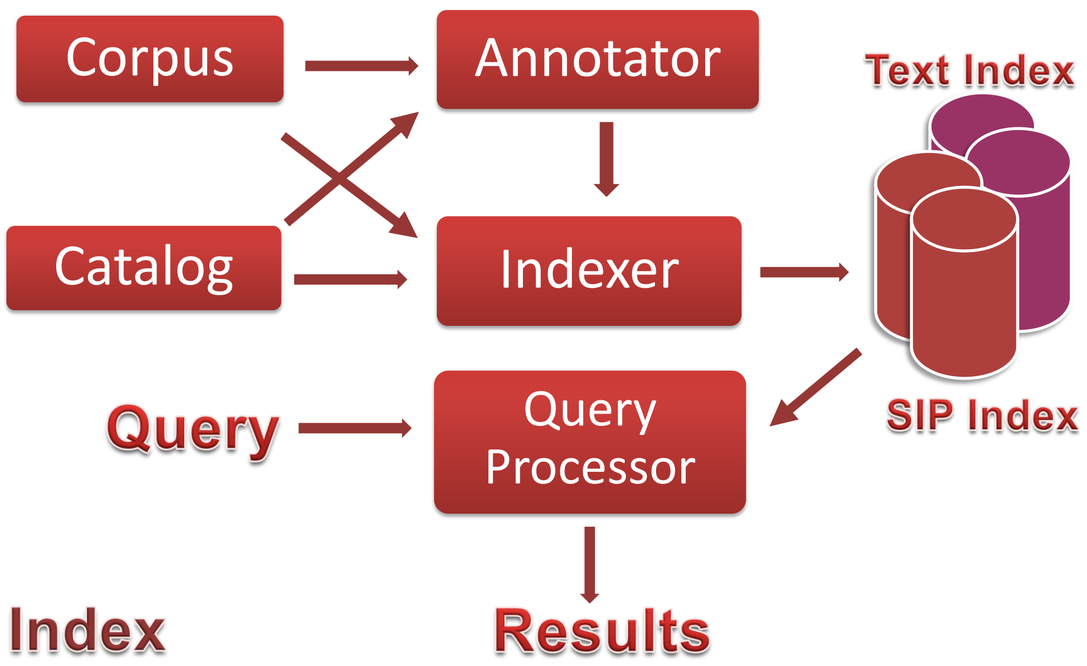
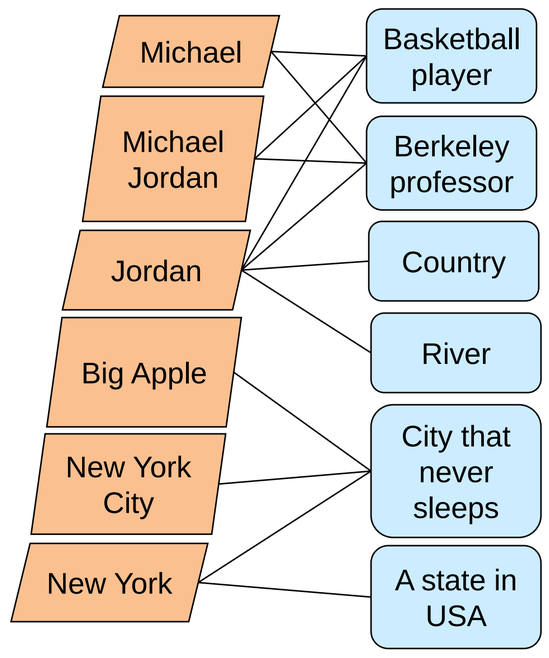
Web Scale Entity Disambiguation

OVERVIEW

Web-scale open domain entity search system that enables semi-structured search over a corpus annotated with entities and types from large catalogs







OVERVIEW

PROBLEM

Scalable?

- Rigorous compression techniques for lemma models bring down their space requirements from >20 GB to 1.15 GB
- Hence, today all lemma models fit in RAM
- With all models in RAM, system can disambiguate at the rate of 0.6 ms/spot
- But, when we migrate from 2-3 Million lemmas in YAGO to over 40 Million in Freebase, even compressed lemma models won't fit in RAM
- Now, lemma disambiguation models will have to be distributed/partitioned

Baseline: Bin Packing

- Partition lemma models into minimum number of disjoint subsets
- Each partition must fit in RAM
- Make one pass through (local part of) corpus per partition
- Advantage
  - No extra network traffic
  - Generate CFVs on the fly; no buffering overhead
  - Statistically load balanced
- Potential problem: too many passes

PROBLEM

CHALLENGES

Map Reduce

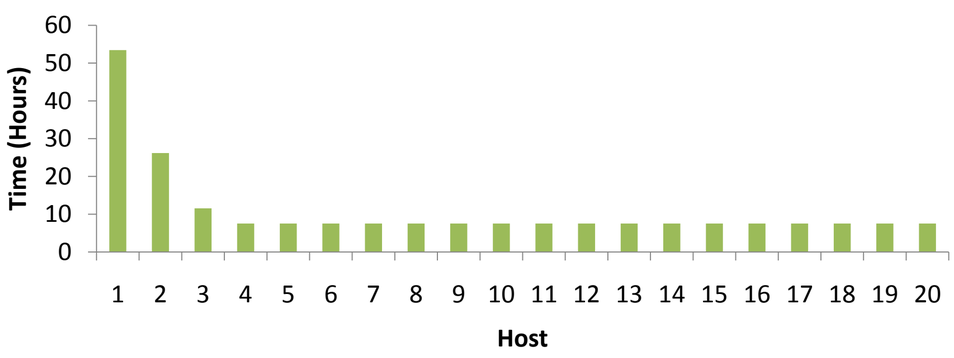
- How might a Hadoop expert code this up?
  - Scan 1B doc in mappers (one per CPU core)
  - Emit ~100B CFVs keyed on lemma, no reduce
    - We do not care where CFV are stored, if at all
    - If materialized on HDFS, quite some extra bulk
  - Second set of 40M mappers, one per lemma
  - Each mapper loads its lemma model in RAM
    - Not as good compression as running them together
  - CFVs find their way to their "home" model
  - Annotations keyed on entity ID, reduce=indexing
- Is this the best possible implementation?

What Might Go Wrong

- Total work per lemma follows skewed distribution.
  - Greedy packing to 160 cores → 79.5% idle CPU
  - Must fragment/replicate lemmas with most work

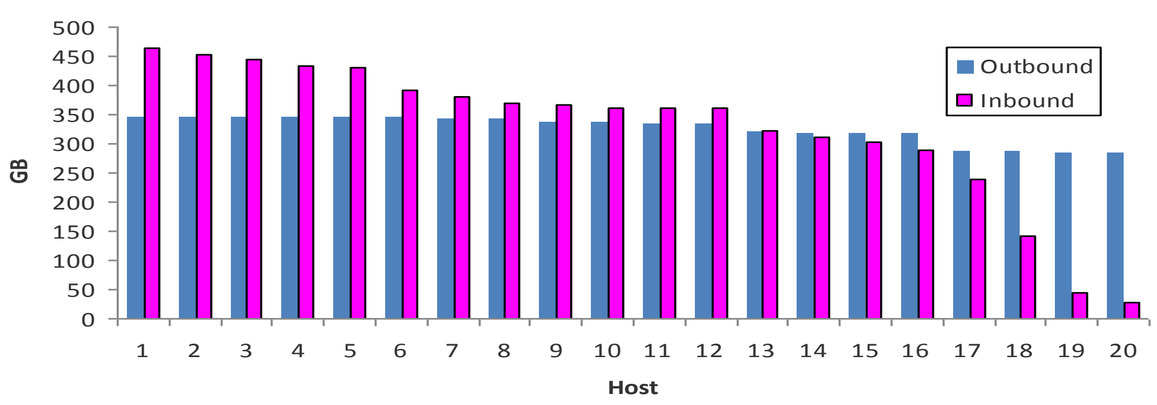
Lemma	Ambiguity	#CFV,M	Time
"2005"	486	710	2d
"2004"	742	230	1d
"2003"	517	140	12h
"2002"	332	110	6h
"2000"	331	90	4.5h
"7"	71	250	3h
"2001"	209	90	3h
"12"	51	240	2h
"2006"	484	30	2h

Computation Imbalance



- All CFVs for one lemma goes to one host
- Otherwise packed for load balance
- Hot lemmas lead to chunky jobs
- Large idle times, must chop up (how?)

Communication Imbalance



- Source corpus, outbound traffic balanced
- CFV destinations show hotspots
- Can Hadoop optimally reorganize CFVs and pack second set of maps with RAM constraint?

Soumen Chakrabarti, Shashank Gupta IIT Bombay