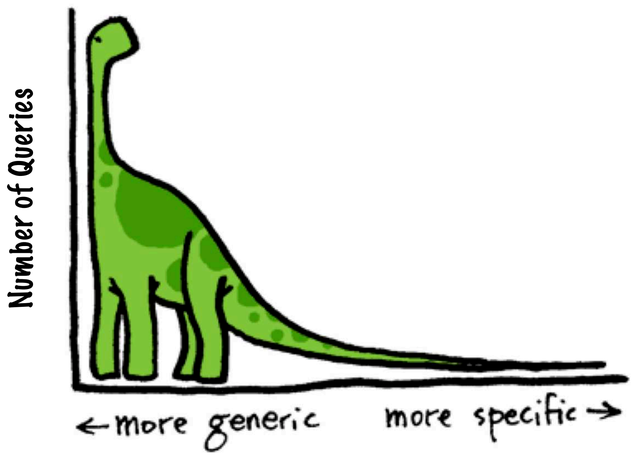
Bringing Head Closer to the Tail with Entity Linking

Manisha Verma[1], Diego Ceccarelli[2,3]
[1] UCL Department Of Computer Science [2] ISTI-CNR, Pisa, Italy [3] IMT Lucca

Introduction

Search queries follow a *Zipfian* distribution.

- ***Head queries*:** Few popular queries with a large volume;
- ***Tail queries*:** A significant percentage of queries that occur rarely.



Given a search log, there is **sufficient data** to analyze **head queries** but **insufficient data (low frequency, limited clicks)** to draw any conclusions about **tail queries**.

- Knowledge bases have enabled understanding of short or long unstructured text;
- Can we quantifying the extent of **overlap between long tail and head queries** by means of **entity linking**.

Research Questions

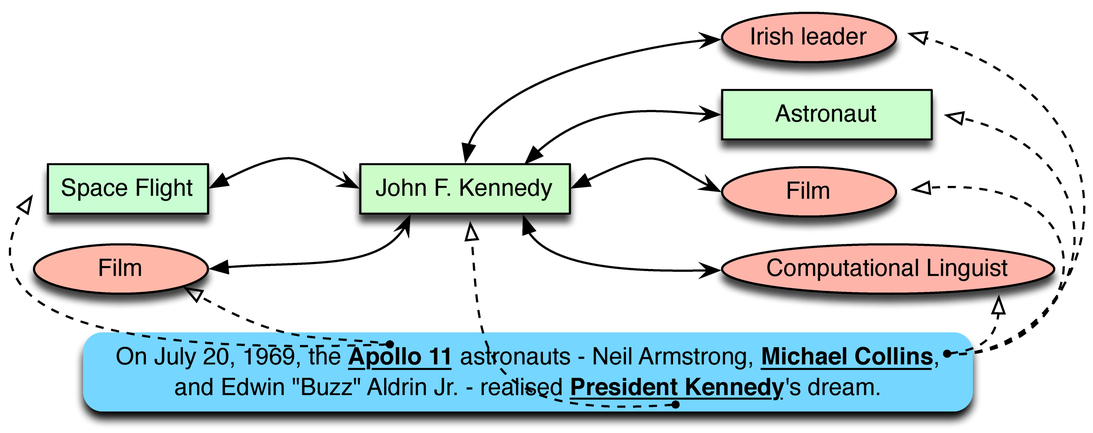
- Are tail queries a different means to inquire about entities mentioned in the head queries?
- Can we find tail queries about entities that are not searched in the head (*tail entities*)?
- Can we find a relationship between *tail entities* and *head entities*?

Entity Linking Problem

The annotation is usually organized in three subtasks:

1. **Spotting**: discover the fragments that could refer to an entity. A set of candidate mentions is detected, and for each mention a list of candidate entities is produces;
2. **Disambiguation**: for each spot associated with more than one candidate, a single entity is selected to be linked to the spot;
3. **Ranking**: the list of entites detected is ranked according to some policy, e.g. annotation confidence.

Our entity linker Dexter (`dxtr.it`) identifies at least one spot in 13,977 (70%) and 4,901,987 (63%) $\mathbf{Q_{head}}$ and $\mathbf{Q_{tail}}$ respectively.



Analysis

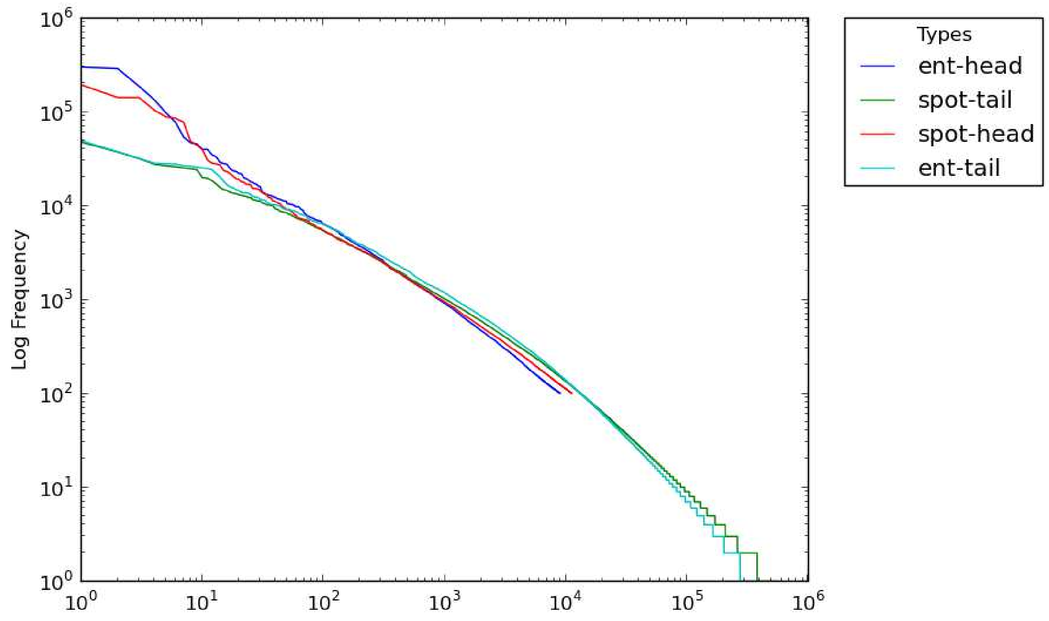
AOL log consists of approximately 20 million queries submitted by 650,000 users. There are in total 10,154,742 distinct queries. We extract 2 distinct sets from these queries:

$\mathbf{Q_{tail}}$: Tail queries with frequency *lower than or equal* to 2. Contains 7,746,607 distinct queries, i.e. 76% of distinct queries.

$\mathbf{Q_{head}}$: Head queries with frequency *greater than* 99. The set contains 19,953 distinct queries, i.e. 0.002%.

Although, the two sets differ in number of queries ($\sim$ 19K versus $\sim$ 7M), they **cover the same fraction of total queries issued** to the search engine.

Spot-Entity Distribution: While queries in head and tail follow totally different distributions, when we look at their spots/entities, the distributions are similar as shown below.



Popular Head and Tail Entities Overlap: On sorting the entities based on their frequency in $\mathbf{Q_{head}}$ and $\mathbf{Q_{tail}}$ respectively and comparing the ranked lists at different cutoffs, Jaccard distance between the ranked lists is 0.25 at 5000. At smaller cutoffs (top 50 entities) the Jaccard is 0.05.

$\mathbf{Q_{head}}$				$\mathbf{Q_{tail}}$			
S_{head}		E_{head}		S_{tail}		E_{tail}	
google	342,602	Google	349,337	florida	47,718	Florida	49,366
myspace	194,093	Yahoo	299,718	texas	37,388	Texas	37,526
yahoo	142,361	Myspace	289,353	ohio	31,861	Ohio	31,905
ebay	142,257	EBay	187,633	edu	26,641	New_York	28,396
yahoo.com	104,696	MapQuest	135,179	state	26,066	.edu	26,642
mapquest	88,617	Google_Search	98,112	california	25,233	U.S._state	26,392
google com	85,670	Hotmail	53,925	new york	24,865	California	25,859
www.yahoo.com	44,198	Craigslist	45,586	real estate	19,702	Myspace	24,998
internet	39,865	Ask.com	39,873	myspace	18,533	Floruit	24,207
ebay com	30,652	Internet	39,865	restaurant	17,065	Restaurant	21,996
hotmail.com	28,492	Pornography	35,089	michigan	15,635	Hotel	20,289
map quest	27,949	Tattoo	33,113	new jersey	14,813	Nudity	18,245
american idol	23,665	Yahoo!_Mail	28,238	black	13,921	Michigan	15,763

Queries with Multiple Entities:There are several queries in the log with multiple entities. On average tail queries have more entities than head. **Do tail queries inquire about entities in the head?** The percentage of tail queries containing only head entities, only tail entities ($\in E_{tail} \setminus E_{head}$) or both is shown below.