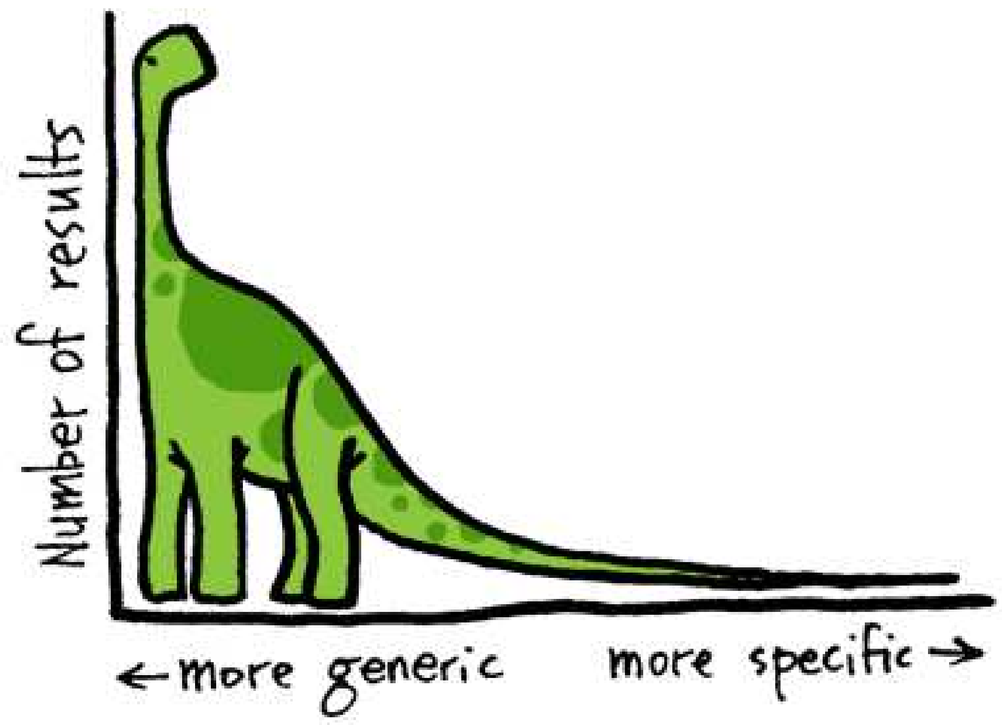
Bringing the Head Closer to the Tail with Entity Linking

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

With the creation and rapid development of knowledge bases, it has become easier to understand the **underlying semantics of unstructured text** (short or long) on the web. In this work we especially look at the impact of entity linking on search logs. Search queries follow a *Zipfian* distribution wherein other than few popular queries (*head queries*), a significant percentage of queries (*tail queries*) 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. In this work we focus on quantifying the extent of **overlap between long tail and head queries** by means of entity linking. We specifically analyze the frequency distribution of entities in head and tail queries.



Research Questions

we are interested in studying the relationship between the head and the tail queries through the entities they contain. Our primary research questions are:

- 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 (we will call them *tail entities*)?
- Can we find a relationship between *tail entities* and *head entities*?

AOL Query Log

We perform our analysis on the AOL query log, since it is publicly available[1]. AOL log consists of approximately 20 million queries submitted by 650,000 users from March to May 2006. Queries are normalized (text lowercased, non ascii characters removed) and there are in total 10,154,742 distinct queries. We extract 2 distinct sets from these queries:

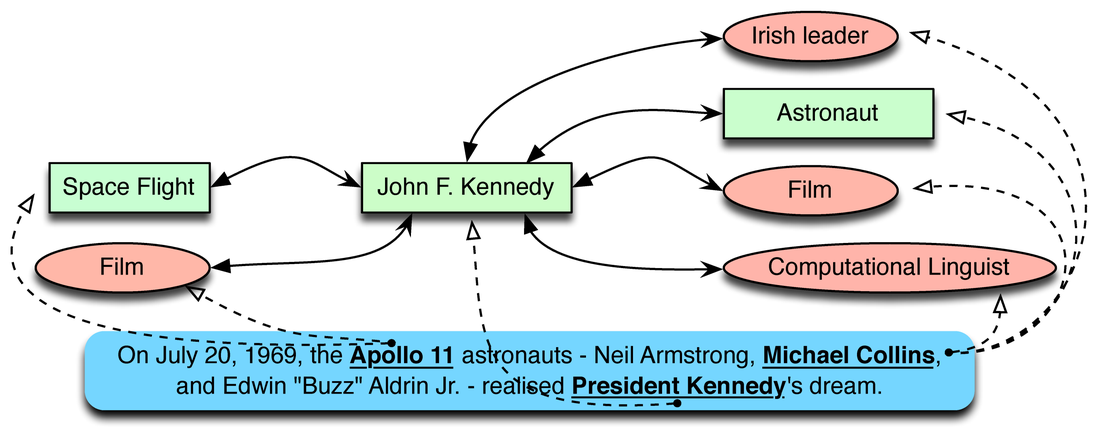
Q_{tail} The set of queries in the long tail, i.e. queries that appear in the log with a frequency *lower than or equal* to 2. The set contains 7,746,607 distinct queries, i.e. 76% of distinct queries, but it is 26% of the total volume of the queries.

Q_{head} The set of queries in the head. It contains queries that appear with a frequency *greater than* 99. The set contains 19,953 distinct queries, i.e. 0.002% if we look at the distinct queries, but still these queries represent 26% of total query volume.

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. All our analysis are performed on these two sets.

Entity Linking Problem

The entity linking task aims at identifying, given a plain document, the small fragments of text (interchangeably called *mentions* or *spots*) referring to any *named entity* that is listed in a given knowledge base, e.g. Wikipedia. The ambiguity of natural language makes it a non trivial task. The same entity can be in fact mentioned with different text fragments, e.g., "President Kennedy" or "John F. Kennedy". On the other hand, the same mention may refer to different entities, e.g., "Michael Collins" may refer to either the well known astronaut, or to the Irish leader and president of the Irish provisional government in 1922.



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%) Q_{head} and Q_{tail} respectively.

Analysis

Q_{head}				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
my space	48,401	Bank_of_America	46,922	hotel	20,018	Real_estate	25,232
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
craigslist	27,222	American_Idol	28,890	georgia	14,525	United_States	16,680
american idol	23,665	Yahoo!_Mail	28,238	black	13,921	Michigan	15,763

ESAIR '14 – 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval. November 7, 2014. Shanghai, China.

[1] `http://www.gregsadetsky.com/aol-data/`