Query and Document Understanding

Rishiraj Saha Roy | Computer Scientist, Adobe Research Labs India | rroy@adobe.com

Overview

- Simple techniques in query and document understanding
- Lucene – A simple commercial text search library
- Take-home assignment on basic Information Retrieval
- Industry positions for text mining and IR skills

Basics

un·der·stand·ing [uhn-der-**stan**-ding] ? Show IPA

noun

1. mental process of a person who comprehends; comprehension; personal interpretation: *My understanding of the word does not agree with yours.*

Basics

- **What is "not" understanding?**
- **Query:** *compare performance shikhar dhawan rohit sharma*
- **Document:** Shikhar Dhawan has much better shot placement than Rohit Sharma.

Basics

Much more to queries and documents than keywords and their frequencies!!!

Basics

Query: *create hyperlinks in excel*

- Documents: Forums
  - **create hyperlinks in** word …. Filters in **excel** have to be specified with…
- Documents: Spam (?)
  - Zingo.com – Your one stop tech quide. Best **excel** tips | Best **hyperlinks in** your page | **Create** your own blog…

Basics

Query 1: *us open home page*

Query 2: *chrome cant open home page*

US open official site by IBM. **Cant** view **page** properly? Best viewed in Google **Chrome**.

Basics

- Relative word orders important

  china *detains* ***india*** *traders latest news*

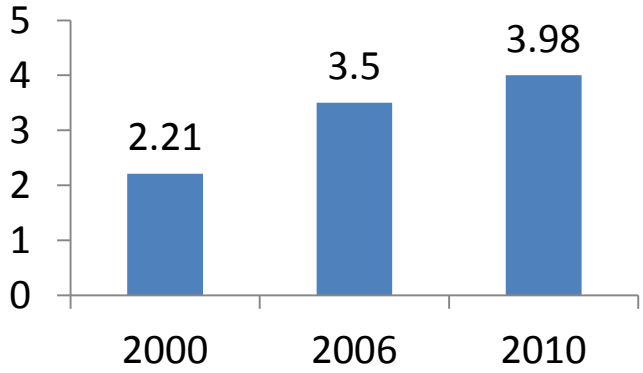
- Query segmentation
  - *glass office windows*
  - *open office windows*
- Entities, Attributes and Relations
  - *france capital, polio symptoms, bon jovi age*
  - *barclays capital, capital punishment?!*

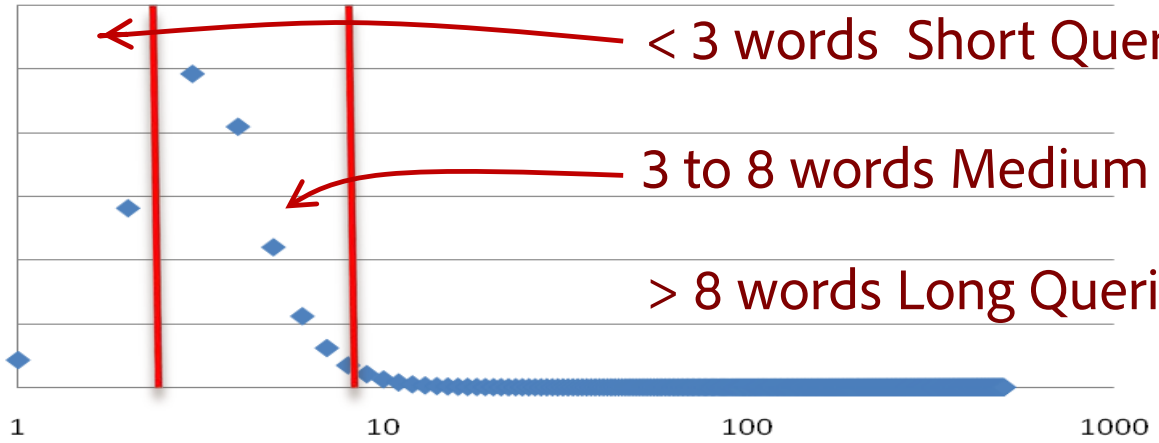
Basics

- And much more!!!
  - Term proximities
  - Term dependencies
  - Term and page annotations
  - …
- Endless research areas……..

Query Lengths

The mean length of Web search queries is increasing

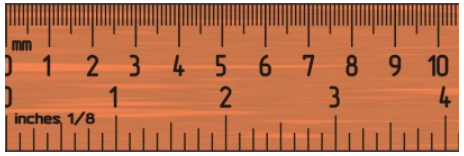




< 3 words Short Queries (14%)

3 to 8 words Medium Queries (80%)

> 8 words Long Queries (3.2%)



Motivation

- Query understanding: Why? How?
- Queries do not follow any formal grammar

"EMERGENCY HATCH PENGUIN EGGS HOW"

medicines for high pressure otc only

samsung galaxy gprs config at&t

(Some more) Motivation

- Reordering, no function words, multiword expressions, part NL
- Natural language processing (NLP) / Linguistics-based techniques fail!
- Computationally expensive!

About 22,500,000 results (0.22 seconds)

- Simple data-driven statistical approaches
- Empirical formulations
- Provide noticeable improvements!!

Query Segmentation

- Query segmentation
  - Why?
  - A simple how
- Extracting Entities and Attributes
  - Why?
  - Some simple hows

Query Segmentation

- Dividing a query into individual semantic units (Bergsma and Wang, 2007)
- Example
  - *australian open home page* →
  - *australian open* | *home page* 
  - *australian* | *open home* | *page* 

Query Segmentation

- Goes beyond multiword named entity recognition (*gprs config, history of, how to*)
- Helps in better query understanding
- Query expansion, query suggestions
- Can improve IR performance by increasing precision

north america versus *north of america*

Query Segmentation

- Simple algorithm – Pointwise Mutual Information

$$PMI(ab) = \log_2 \left(\frac{p(ab)}{p(a) * p(b)} \right)$$

- Compute probabilities from any source – documents, queries, page titles, anchor text
- Microsoft Web *n*-gram services
- http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx

Query Segmentation

- PMI measures strength of bonding – by chance or by choice?
- Meanigful bigrams have high PMI – *harry potter, blood pressure, jurassic park, difference between*
- Measure PMI of adjacent word pairs
- Fix significance threshold
- Insert boundary whenever PMI falls below threshold

Query Segmentation

- Input: *australian open home page*
- PMI(*australian*, *open*) = 15.89
- PMI(*open*, *home*) = 5.43
- PMI(*home*, *page*) = 13.92
- Threshold: 8.50
- Output: *australian open* | *home page*
- Problem: Not optimized over whole query!!

(Named) Entities

jetbeam rrt-01

- Where to buy? How to use? Life? Weight? ....

roger federer

- Return information in structured form

lotr cast

- Book? Movie? Game?

Detecting Entities

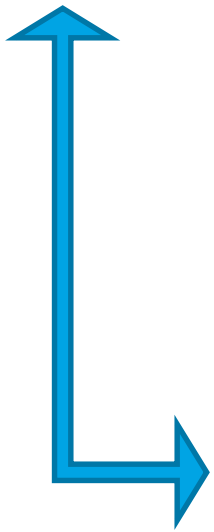
- Simplest – List based approach
- Wikipedia titles acts as a very good resource
- http://dumps.wikimedia.org/enwiki/latest/
- 5 million entries, 2 GB RAM, no problem

Detecting Entities

- Efficient data structures – Trie, Dictionary
  - Low memory
  - Fast search
- Lists work great, extensive commercial use
- Annotate both queries and documents

Detecting Entities

howard shore music director



```
howard shelley
howard sheppard
howard sherman sheehy
howard sherwin
howard shiplee
howard shipley
howard shore
howard shultz miller
howard siler
howard silverman
howard simms
howard simon
```

Detecting Entities

- Often need to view very large files – lists, logs
- LTF Viewer – An unsung hero
  - http://www.swiftgear.com/ltfviewer/features.html
- Vim, Cygwin, command-based
- Edit programmatically only

Problems

- More than one match
  - *the dark knight, the dark knight rises*
  - *tom **cruise ship** scene*

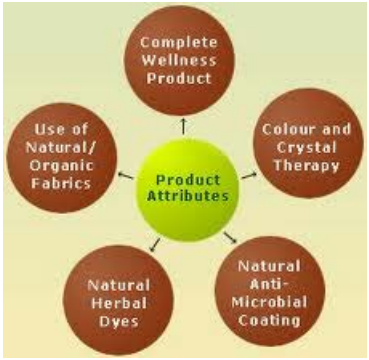
- False positives – Match, but not entity
  - *list of capitals*

Identifying Attributes

- Why?
- User wants specific results
  - *galaxy note **specs***
- Intent diversification
  - *galaxy note (What about it??)*
  - *Pictures, specs, stores, prices, accessories*

Identifying Attributes

- Using documents: Template based
  - *What is the A of I <what … A … I>*
  - *I's A*
  - *Who was A of I <who … A … I>*
  - *A of I*
  - *A in I*

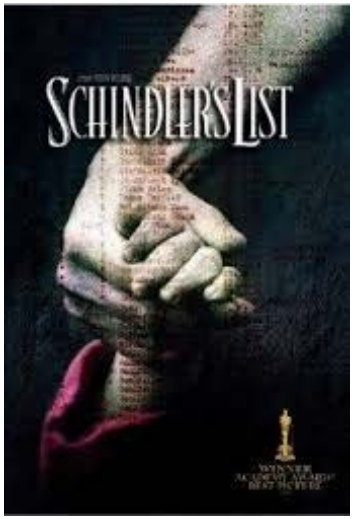


Identifying Attributes

- *Ps2's **accessories***
- ***Accessories** of galaxy note*
- *New Delhi is the **capital** of India*
- *Paris is the **capital** of france*
- *Narendra Modi is the **prime minister** of India*
- *??? is the **prime minister** of Pakistan*

Identifying Attributes

- Challenges
  - *Hall of fame*
  - *Wall of shame*

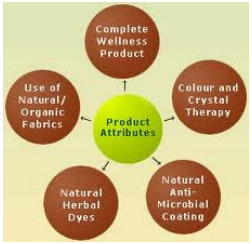


  - *Shindler's list*
  - *Beijing's mist*

Identifying Attributes

- Using query logs or documents – Co-occurrence counts

- Common wisdom: Attributes are **frequent** words
- More robust statistics: They co-occur with a higher number of distinct words



Identifying Attributes

- *nikon camera **prices**, winter coats **prices**, property **prices** in bengaluru*
- ***nikon camera** prices, **nikon camera** models, **nikon camera** for sale*
- Issues: Where to draw the line?

- *lyrics, recipe, cast*

- *after, test, centre, black, server*

Summary

- Keyword-based retrieval good, but not enough
- Query and document understanding are required to boost IR performance
- Methods used need to be fast and scalable
- Query segmentation is a first step towards better query representation
- Entities and attributes can be identified effectively using simple approaches
- References: http://bit.ly/19b2dMC

How to Use Lucene

Files: http://cse.iitkgp.ac.in/resgrp/cnerg/qa/ForLucene.zip

Basic IR Assignment

Open

Industry Scope

Questions?


Adobe