Query and Document Understanding

Rishiraj Saha Roy

Ph.D. Student
under
Prof. Niloy Ganguly (IIT Kharagpur)

and

Dr. Monojit Choudhury (Microsoft Research India)

Pre-FIRE Workshop on Information Retrieval
Bengaluru, 15 - 17 June 2013
understanding [uhn-der-stan-ding]  

noun

1. mental process of a person who comprehends; comprehension; personal interpretation: My understanding of the word does not agree with yours.
What is “not” understanding?

Query: compare performance shikhar dhawan rohit sharma

Document: Shikhar Dhawan has much better shot placement than Rohit Sharma.
Much more to queries and documents than keywords and their frequencies!!!
Query: *create hyperlinks in excel*

- **Forums**
  - *create hyperlinks in* word .... Filters in excel have to be specified with...

- **Spam (?)**
  - Zingo.com – Your one stop tech quide. Best excel tips | Best hyperlinks in your page | Create your own blog today
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*
And much more!!!

- Term proximities
- Term dependencies
- Term and page annotations
- ...

- Endless research areas............
The mean length of (distinct) Web search queries is **increasing**

- **< 3 words** Short Queries (14%)
- **3 to 8 words** Medium Queries (80%)
- **> 8 words** Long Queries (3.2%)
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
Reordering, 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!!
Outline

- 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

\[ PMI(ab) = \log_2 \left( \frac{p(ab)}{p(a) \cdot p(b)} \right) \]

- Compute probabilities from any source – documents, queries, page titles, anchor text

- Microsoft Web n-gram services
Query Segmentation

- PMI measures strength of bonding – by chance or by choice?
- Meaningful 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
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?
Simplest – List based approach

Wikipedia titles:

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
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
Ps2’s accessories

Accessories of galaxy note

New Delhi is the capital of India

Paris is the capital of France

Manmohan Singh is the prime minister of India

??? is the prime minister of Pakistan
Identifying attributes

- **Challenge**
  - *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, microsoft share prices
- nikon camera prices, nikon camera models, nikon camera for sale, nikon camera lens
- 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
Questions??