

11-442 / 11-642 / 11-742:
Search Engines

Information Needs and Queries

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

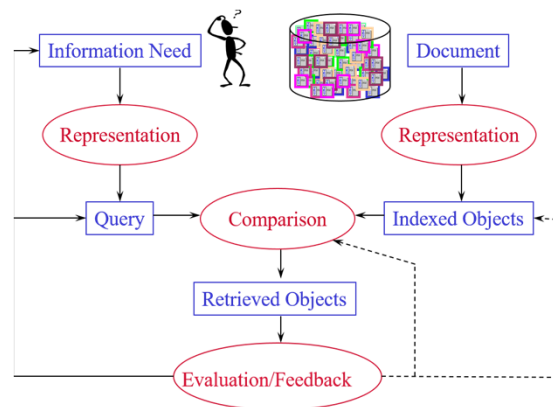
1

Outline

Information needs

Queries

Query processing and query reformulation



2

© 2021, Jamie Callan

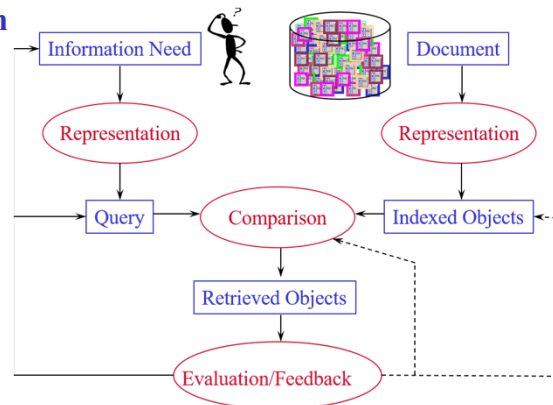
2

Information Needs

A person begins a search with an information need in mind

The information need is implicit and unknown

- The query describes the information need
...but it may not be an accurate description



3

© 2021, Jamie Callan

3

Real Web Queries

- virginia beach
- city of virginia beach
- geico
- map quest
- ringworm
- images of scalp ringworm
- netflix
- three laws of motion
- brain teasers
- origin of 'picnic'
- colleges in georgia
- bad credit
- blackwater
- diplomat security
- fedex logo
- lose weight fast
- danica patrick
- bikinis
- expeditor airlines
- bathroom ventilation fans
- black models agency

4

© 2021, Jamie Callan

4

Information Needs

Often people don't describe their information needs well

- Librarians are trained to elicit information needs
 - Describe the information need in narrative (sentences, paragraph) form
 - Identify the key topic(s), and synonyms for key topic(s)
 - Identify supporting or related concepts, and their synonyms
 - Identify other constraints (e.g., genre, type of material, format, ...)
 - ...
- Much of what is known about this topic is from Library Science
 - How well does this information apply to the web?

5

© 2021, Jamie Callan

5

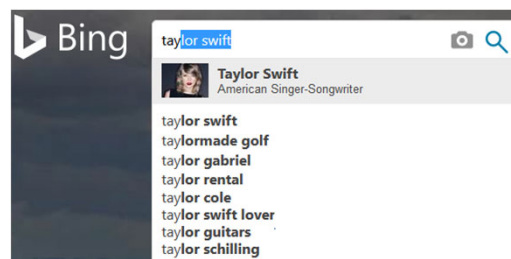
Information Needs

Search engines also elicit information needs

- How do they compare to elicitation by librarians?
- Initial elicitation



- Subsequent elicitation



6

© 2021, Jamie Callan

6

Specifying an Information Need: TREC Blog Track Topic 1105

Query: parenting

Description: I am looking for blogs that provide advice, counseling, and information on parenting.

Facet: personal

Narrative: Relevant blogs include those from parents, grandparents, or others involved in parenting, raising, or caring for children. Blogs can include those provided by health care providers if the focus is on children. Blogs that serve primarily as links to other sites, or that of themselves, market products related to children and their caregivers, are not relevant.

7

© 2021, Jamie Callan

7

Specifying an Information Need: TREC Topics

Why are TREC topics elaborate? Why not just use a query?

They are influenced by how librarians elicit information needs

- Gather information from multiple perspectives
- Gather information at various levels of detail

Why would this be a good idea for TREC?

- Greater consistency in making relevance judgments
- Support research on advanced methods of creating queries
 - Narratives may be more like what people would say if our systems were smarter

8

© 2021, Jamie Callan

8

Information Needs

There are many different kinds of information needs

- **Known item:** I've seen it before, but I can't find it now
- **Known attribute:** I know something about it
- **General content search:** Find something about the topic
- **Exhaustive literature review:** Find everything about the topic
- : : : : :

Different types of information needs require different methods

- Not a lot is known about effective strategies for different needs
- Major focus of research and commercial activity

9

© 2021, Jamie Callan

9

Common Web Information Needs

Informational (39%): “iphones”, “San Francisco”

- User wants to learn about the topic
- Find information on a topic

Transactional (36%): “shopping”, “buying airline tickets”

- User has a task, but no specific destination (website) in mind
- Find a site to carry out a transaction

Navigational (25%): “Megabus”, “Dell”

- User has a specific destination (website) in mind
- Find a specific location (website)

(Broder, 2002)

10

© 2021, Jamie Callan

10

Common Web Information Needs

Five intents from a more recent study

- **Informational** 27-42%
- **Navigational:** Purpose is to reach a particular site 11-39%
- **Transactional:** The intent is to complete a transaction 22%
- **Commercial:** Motivated by commercial interest 19-46%
- **Local:** The query has a local focus 9-26%

A query can be in more than one category

(Lewandowski, 2012)

11

© 2021, Jamie Callan

11

Query Intents

Five sub-intents for shopping related queries

- **Buying guide:** Factors to consider when buying a product type
- **Reviews:** Ratings, recommendations, comparisons
- **Support:** Manuals, troubleshooting, tutorials, warranties
- **Official product homepage**
- **Shopping site/Purchase:** Places where the product can be bought

(Chapelle, et al., 2011)

12

© 2021, Jamie Callan

12

Information Needs and Queries

Information needs are expressed as queries
... what do we know about queries?

13

© 2021, Jamie Callan

13

Outline

Information needs

Queries

Query processing and query reformulation

14

© 2021, Jamie Callan

14

Web Queries

Typically, web queries are 1-3 words long (average is 2.x)

- Because people can't form longer queries?
- Because people don't need longer queries?
- Because Web search engines discourage longer queries?

15

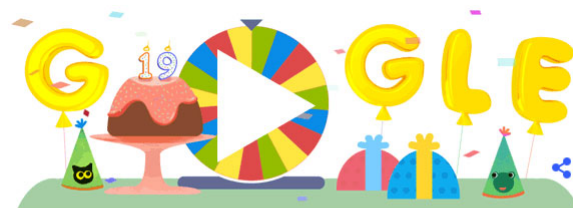
© 2021, Jamie Callan

15

Information Needs and Manual Queries

The user interface plays a large role in how people express their information need

- A small box encourages short queries



**Big picture
(entertainment)**

**Small search box
(the task)**

16

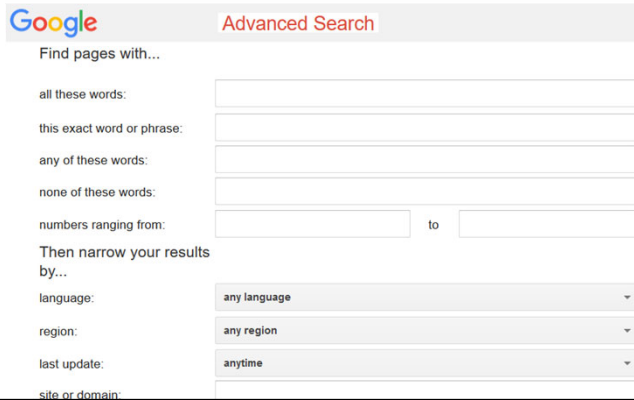
© 2021, Jamie Callan

16

Information Needs and Manual Queries

The user interface plays a large role in how people express their information need

- A small box encourages short queries
- A form encourages more detail



Google Advanced Search

Find pages with...

all these words:

this exact word or phrase:

any of these words:

none of these words:

numbers ranging from: to

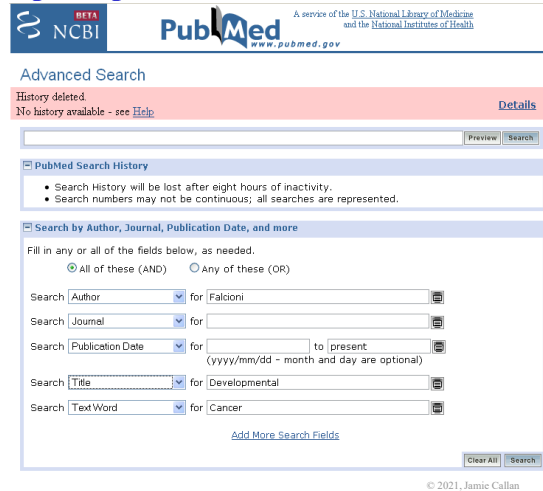
Then narrow your results by...

language:

region:

last update:

site or domain:



NCBI PubMed A service of the U.S. National Library of Medicine and the National Institutes of Health www.pubmed.gov

Advanced Search

History deleted
No history available - see [Help](#)

Details

PubMed Search History

- Search History will be lost after eight hours of inactivity.
- Search numbers may not be continuous; all searches are represented.

Search by Author, Journal, Publication Date, and more

Fill in any or all of the fields below, as needed.

☒ All of these (AND) ☐ Any of these (OR)

Search for

Search for

Search for to
(yyyy/mm/dd - month and day are optional)

Search for

Search for

[Add More Search Fields](#)

© 2021, Jamie Callan

17

Information Needs and Manual Queries

User-training plays a large role in how people express their information needs

- WestLaw queries are 10-12 words long
 - » Professional searchers

WestLaw example:

- **Information need:** Requirements for disabled people to be able to access a workplace
- **Query:** disab! /p access! /s work-site work-place (employment /3 place)

(Manning, et al., 2008)

18

© 2021, Jamie Callan

18

TREC Legal Track: Adversarial Production Requests

Production Request 52: Please produce any and all documents that discuss the use or introduction of high-phosphate fertilizers (HPF) for the specific purpose of boosting crop yield in commercial agriculture.

Negotiated Query: ((“high-phosphat! fertiliz!” OR hpf) OR ((phosphat! OR phosphorus) w/15 (fertiliz! OR soil))) AND (boost! OR increas! OR rais! OR augment! OR affect! OR effect! OR multipl! OR doubl! OR tripl! OR high! OR greater) AND (yield! OR output OR produc! OR crop OR crops)

Query language details

- ! matches different stems (e.g., fertilize, fertilizers, fertilized, ...)
- w/15 is NEAR/15
- “ ” is a phrase operator

(TREC 2007 Legal Track)

19

© 2021, Jamie Callan

19

Outline

Information needs

Queries

Query processing and query reformulation

20

© 2021, Jamie Callan

20

Query Languages to Query

People can manually form structured queries

- Most people don't do this well (but experts do very well)

21

© 2021, Jamie Callan

21

Query Languages to Query

People can manually form structured queries

- Most people don't do this well (but experts do very well)

The search engine can automatically form a structured query

- **Query-processing:** Transformations to individual query terms
- **Query reformulation:** Transformations to the query as a whole

Goal: Improve the match between query and relevant documents

Historically this has been very important

22

© 2021, Jamie Callan

22

Query Processing

Case conversion: Virginia → virginia

Stopword removal: city of virginia beach → city virginia beach

Stemming:

- Stemmed index: apples → apple
- Unstemmed index: apples → #synonym (apple, apples)

Whatever was done to create the index, also do it for queries

23

© 2021, Jamie Callan

23

Query Processing

Phrases:

- die-cast → #NEAR/1 (die cast)
- virginia beach → #NEAR/1 (virginia beach)
- barack obama → #NEAR/3 (barack obama)

Abbreviations: virginia → #synonym (virginia, va)

Spelling correction:

- brittany spears → britney spears
- brittany spears → #synonym (brittany, britney) spears

24

© 2021, Jamie Callan

24

Query Reformulation: Multiple Representations

User query: **The Time Traveler's Wife**

A web search engine might transform the query to match against multiple document representations (fields)

- Each query term is mapped to a subquery
 - $q_i \rightarrow \#SUM(q_i.field_1 \dots q_i.field_n)$
 - $q_i \rightarrow \#WSUM(weight_1 q_i.field_1 \dots weight_n q_i.field_n)$
 - » Weights indicate the relative importance of each field

25

© 2021, Jamie Callan

25

Query Reformulation: Multiple Representations

User query: **The Time Traveler's Wife**

A web search engine might transform the query to match against multiple document representations (fields)

#AND (
 #WSUM(0.1 time.url 0.2 time.title 0.3 time.inlink 0.4 time.body)
 #WSUM(0.1 traveler.url 0.2 traveler.title 0.3 traveler.inlink 0.4 traveler.body)
 #WSUM(0.1 wife.url 0.2 wife.title 0.3 wife.inlink 0.4 wife.body))

26

© 2021, Jamie Callan

26

Query Reformulation: Sequential-Dependency Models

The sequential dependency model (SDM) converts unstructured queries to structured queries

A sequential dependency model query has three parts

- Bag of words matches
 - #AND ($q_1 q_2 \dots q_n$)
- N-gram matches (ordered, phrase-like)
 - #NEAR/1 ($q_1 q_2$) #NEAR/1 ($q_2 q_3$) ... #NEAR/1 ($q_{n-1} q_n$)
- Short window matches (unordered, sentence-like)
 - #WINDOW/8 ($q_1 q_2$) ... #WINDOW/8 ($q_{n-1} q_n$)
 - Note: Window sizes are $4 \times$ number of terms in window

**Very
important!**

27

© 2021, Jamie Callan

27

Query Reformulation: Sequential-Dependency Models

User Query: **The Time Traveler's Wife**

A sequential dependency model query

#wand (
0.7 #and (time traveler wife)
0.2 #and (#near/1 (time traveler) #near/1 (traveler wife))
0.1 #and (#window/8 (time traveler) #window/8 (traveler wife)))

Probabilistic #and

Bag of words: Pretty much guaranteed to find something

#NEAR/1: Extra weight for matching n-grams

#WINDOW/n: Extra weight for matching window constraints

28

© 2021, Jamie Callan

28

Query Reformulation: Sequential-Dependency Models

User Query: a b c d e

A sequential dependency model query

```
#wand (  
  0.7 #and (a b c d e)  
  0.2 #and ( #near/1 (a b)    #near/1 (b c)  
            #near/1 (c d)    #near/1 (d e) )  
  0.1 #and ( #window/8 (a b) #window/8 (b c)  
            #window/8 (c d) #window/8 (d e) ) )
```

29

© 2021, Jamie Callan

29

Query Reformulation

User query **The Time Traveler's Wife**

Combining multiple representations with sequential dependency

```
#wand (  
  0.6 #and (  
    #wsum(0.1 time.url    0.2 time.title    0.3 time.inlink    0.4 time.body)  
    #wsum(0.1 traveler.url 0.2 traveler.title 0.3 traveler.inlink 0.4 traveler.body)  
    #wsum(0.1 wife.url    0.2 wife.title    0.3 wife.inlink    0.4 wife.body))  
  0.4 #wand (  
    0.5 #and (time traveler wife) sequential dependency model  
    0.3 #and (#near/1 (time traveler) #near/1 (traveler wife))  
    0.2 #and (#window/8 (time traveler) #window/8 (traveler wife)))
```

30

© 2021, Jamie Callan

30

Query Processing and Query Reformulation

Query processing and reformulation are found in many systems

- Simple, carefully-tuned heuristics
- Mostly designed for “common” scenarios

Usually improves retrieval accuracy significantly

- Good “average case” performance
 - Some queries are hurt, but most will be improved
 - Win / loss ratio

31

© 2021, Jamie Callan

31

Query Classification and Reformulation

When a query is received by the search engine

- Use classification to identify the query intent (e.g., “local”, “navigational”, ...)
- Use a query structure (“template”) designed for that type of information need

User query ———> Select intent ———> Reformulated query

Navigational	Template ₁
Informational	Template ₂
:	:
Shopping	Template _n

This is done by Web search engines

- But not a lot is known about how they do it

(E.g., U.S. Patent 20060190439)

32

© 2021, Jamie Callan

32

Query Reformulation on the Web

Jon Pederson (Bing) says...

- **Query understanding is critical to web search**
 - Affects most queries
 - Can radically improve results
- **Trade-off between relevance and efficiency**
 - Rewrites can produce costly queries
 - Win/loss ratio is the key metric (# queries improved / # queries that get worse)
- **Especially important for tail queries**
 - No meta-data to guide matching and ranking

(Pederson, 2010)

33

© 2021, Jamie Callan

33

Outline

Information needs

Queries

Query processing and query reformulation

34

© 2021, Jamie Callan

34

For More Information

- O. Chapelle, S. Ji, C Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. “Intent-based diversification of web search results: Metrics and algorithms.” *Information Retrieval*. Springer. 2011.
- D. Lewandowski, J. Drechsler, and S. von Mach. “Deriving query intents from web search engine queries.” *Journal of the American Society for Information Science and Technology*. 2012.