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

Learning to Rank: Neural Models
Search Log Analysis

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

Outline

- Introduction
- Deep Structured Semantic Models (DSSM)
- Deep Relevance Matching Model (DRMM)
- Kernel-based Neural Ranking Model (K-NRM)
- Convolutional Kernel-based Neural Ranking Model (Conv-KNRM)
- Re-ranking with BERT
- DeepCT
- **doc2query**
- Summary

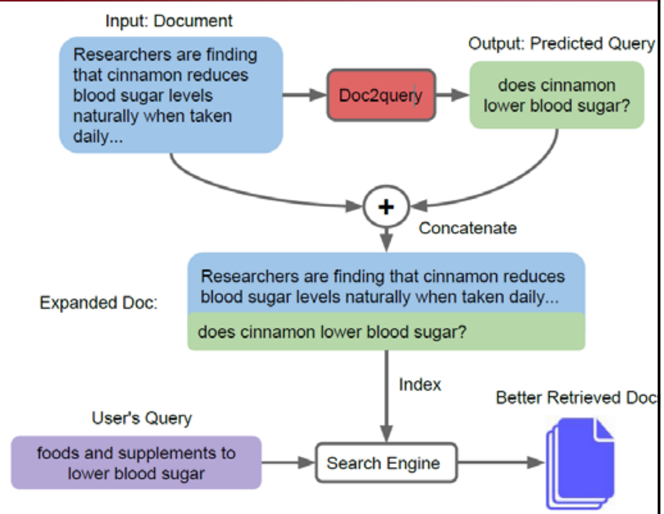
doc2query

For each document d in the corpus

- Automatically generate questions that d can answer
- Add these questions to d
  - Append to the end of d
  - *document expansion*

Use the expanded documents to build an ordinary inverted index

Use BM25 for first-stage retrieval



(Nogueira, et al., 2019)

3

doc2query: Examples

Document: July is the hottest month in Washington DC with an average temperature of 27C (80F) and the coldest is January at 4C (38F) with the most daily sunshine hours at 9 in July. The wettest month is May with an average of 100mm of rain.

Target query: what is the temperature in washington

Predicted query: weather in washington dc

Document: The Delaware River flows through Philadelphia into the Delaware Bay. It flows through and aqueduct in the Roundout Reservoir and then flows through Philadelphia and New Jersey before emptying into the Delaware Bay.

Target query: where does the delaware river start and end

Predicted Query: what river flows through delaware

(Nogueira, et al., 2019)

4

doc2query:
Examples

- why was the manhattan project important
- why was the manhattan charter
- what did the manhattan project do
- what was the importance of manhattan
- what was the importance of manhattan communication
- why was the manhattan project created
- what was the manhattan project
- why was the manhattan project important
- why was manhattan an important factor
- what was the result of the manhattan
- what is the manhattan project
- : : : : : : : : :

- what is vascular)
- what is vascular) material
- what is vascular) in plants
- what is a Phloem
- what is vascular) in photosynthesis
- what are vascular plants
- what is vascular plants
- what is a vascular plant
- what do vascular plants do
- what are vascular plants
- what is a vascular plant
- what are xylem and vascular plants
- : : : : : : : : :

(https://github.com/nyu-dl/dl4ir-doc2query)

5

doc2query:
Examples

- what was the impact of the civil war
- what was the impact of the american industrial
- what was the impact of the civil industrial
- what was the impact of the american civil war
- what was the result of the industrial of the industrial
- what did the treaty of Exclusion do
- what was the treaty of Exclusion
- what did the Congress treaty do
- what did the treaty of Exclusion immigration. do
- how much did the treaty of Exclusion affect
- : : : : : : : : :

- what is costa rica
- what is costa rica known for
- what is costa rica prime
- what is conducive
- what is costa rica known for?
- Rica: Medical
- what is Medical Medical
- what is Medical Medical made of
- what is Medical
- what is Medical in costa rica
- what is a medical services,
- what is Medical in services,
- : : : : : : : : :

(https://github.com/nyu-dl/dl4ir-doc2query)

6

doc2query

How are queries generated?

- Train: Use (q, $d_{relevant}$) pairs to train a sequence-to-sequence transformer model
  - $d_{relevant} \rightarrow q$
  - Datasets
    - » MS MARCO Dev set for training (6,900 queries)
    - » TREC CAR (3M queries)
- Test: Predict 10 queries per document
  - Use top-k sampling

A later version uses the T5 transformer, which generates better queries

- Better queries enables using 40 queries per document, which is much more effective

(Nogueira, et al., 2019)

doc2query

doc2query improves BM25 by about 15% (and docT5query is better – 25%?)

What are the effects?

- **Term reweighting:** tf is increased for some terms
  - 69% of the non-stopword terms in generated queries were already in the document
  - "Researchers find that living with cats reduces allergies in children." → "Do cats reduce allergies in children?"
- **Reduced vocabulary mismatch:** new terms are added to the document
  - 31% of the non-stopword terms in generated queries were not in the document
  - "Researchers find that living with cats reduces allergies in children." → "Are kittens healthy for kids?"

(Lin, et al., 2020)
(Nogueira, et al., 2019)

doc2query

What are the main sources of improvement?

	Method	MS MARCO Passage MRR@10	MS MARCO Passage Recall@1k
(1)	Original Text	0.184	0.853
(2a)	+ Expansion w/ New Terms	0.195	0.907
(2b)	+ Expansion w/ Copied Terms	0.221	0.893
(2c)	+ Expansion w/ Copied Terms + New Terms	0.277	0.944
(3)	Only Expansion Terms (Without Original Text)	0.263	0.927

The effect seems more complex than typical query expansion

- Not just term reweighting and adding related vocabulary
- The expansion queries appear to summarize the document quite well
- The effects are not well-understood, but they appear to be consistent

(Lin, et al., 2020)

doc2query

doc2query is compatible with other familiar techniques

- Use your favorite initial ranker (BM25, Indri, VSM)
- Pseudo relevance feedback
- BERT reranking
  – Better initial ranking produces better re-ranking

(Lin, et al., 2020)

Outline

Summary

Continuous representations are popular again

- Lexical (DSSM), conceptual (DRMM, K-NRM, Conv-KNRM)

Two main types of architectures

- Representation-based vs. interaction-based

Integration of exact-match and soft-match signals

- Older systems were discrete or continuous, not both
- The combination seems effective and reliable (robust)

Some architectures require much training data, some don't

- E.g., trained (much data) vs static (little data) embeddings

Summary

No feature engineering … but <u>much</u> network engineering

- Ignore the hype … not necessarily less work

Poor understanding of <u>how well</u> and <u>why</u> the system works

- Early neural rankers compared to weak baselines (i.e., not LeToR)
- What is the contribution of different parts of the network?
- Did the system learn (good), or did it memorize (less good)?
  - Neural ranking systems are good at memorizing data

Some research embeds familiar ideas in complicated networks

- log (tf), idf, proximity, multiple bags-of-words (title, body, …)

13

Summary

Better text understanding (DeepCT, doc2query) can improve older retrieval models

Why is this important?

- These models are still used widely
  - Alone, and as the first stage of re-ranking architectures
- Text understanding in these models hasn't changed in a long time
  - Much research, but little change in the state of the art
- Deep text analysis + efficient matching with inverted indexes
  - Moves a computationally complex task to indexing
  - Encourages us to explore hybrid indexing strategies

There is more opportunity here than many people realized

14

Lecture Outline

- **Introduction to search logs**
- **Users and tasks**
- **Segmenting search logs into sessions**

Search Logs

Most search engines save information about every search

- The query
- A timestamp
- The IP address of the search client
- Possibly an id recorded in a cookie or obtained another way
- Information about the operating system and browser
- …

Search engines can also collect information about which search results are clicked

- Clickthrough information

Tracking Clickthrough

A search result from a commercial search engine

Jamie Callan
www.cs.cmu.edu/~**callan**/ ▾ Carnegie Mellon University ▾
Jun 2, 2014 - SCS LTI Professor's research, teaching and publications.

This links to a Google service, not Jamie's web page

<a href="http://www.google.com/url?...
url=http%3A%2F%2Fwww.cs.cmu.edu%2F~callan%2F..."
…)">Jamie Callan

It logs the click and returns a page that redirects to Jamie's page

- User information: User's IP address, timestamp, browser information, …
- Result information: Query, clicked URL, position in the list, …

Publicly Available Web Search Logs

There are few publicly available web search logs

- The Excite log (1997)
  - 18,113 users, 51,473 queries
- The AOL log (2006)
  - More than 650,000 users
  - More than 20 million queries

Why aren't more search logs available?

- Competitive reasons
- Privacy reasons


cnet news
Search
Latest News | Crave | Webware | Business Tech | Green Tech | Wirele

Home > News >

August 21, 2006 12:38 PM PDT

Three workers depart AOL after privacy uproar

By Elinor Mills and Anne Broache
Staff Writers, CNET News

Print E-mail Share

Related Stories

AOL gaffe draws Capitol Hill rebuke
August 9, 2006

AOL's disturbing glimpse into users' lives

Two AOL employees have been fired, and its chief technology officer is resigning, after the release of Web search data from thousands of AOL members prompted widespread criticism of the company.

CTO Maureen Govern "has decided to leave AOL effective immediately," AOL Chief Executive Jon Miller wrote in an e-mail to employees dated Monday.

Govern could not be reached to comment.

Sensitive Information in Web Search Logs: One Individual's Queries

bladder infection	2006-05-13 09:22:53
cleveland ohio jobs	2006-05-15 07:45:51
cleveland plain dealer	2006-05-15 07:47:17
fitness job search	2006-05-15 07:53:46
ymca in cleveland ohio	2006-05-15 08:05:42
ymca jobs in cleveland ohio	2006-05-15 08:14:32
ymca in parma ohio	2006-05-15 08:23:01
united health care	2006-05-15 09:25:37
surgery for bladder	2006-05-15 10:23:07
incontinence surgery	2006-05-15 10:30:43
exercises for legs and abs	2006-05-15 19:26:20
free money for women starting a business	2006-05-16 09:36:40
…	

(AOL search log)

19

Web Search Logs: More Detail

	gout		2006-03-01 07:38:03
	chemotherapy		2006-03-01 07:41:04
	chemotherapy side effects		2006-03-01 07:42:36
Click on #1 result →	1	http://www.cancerhelp.org.uk	
	chemotherapy causing hearing loss		2006-03-01 07:45:23
	2	http://www.sciencedaily.com	
	kenny rogers songs		2006-03-02 06:05:40
	kenny rogers' song i cant unlove you		2006-03-02 06:06:58
Click on #4 result →	4	http://www.kennyrogers.com	
	kenny rogers' song i cant unlove you		2006-03-02 06:06:58
	3	http://www.cmt.com	
	kenny rogers' song i cant unlove you		2006-03-02 06:06:58
	6	http://www.lyricspremium.com	

(From AOL search log, part 9)

20

Inaccessible and Less Accessible Web Search Logs

Statistics about some web search logs have been published

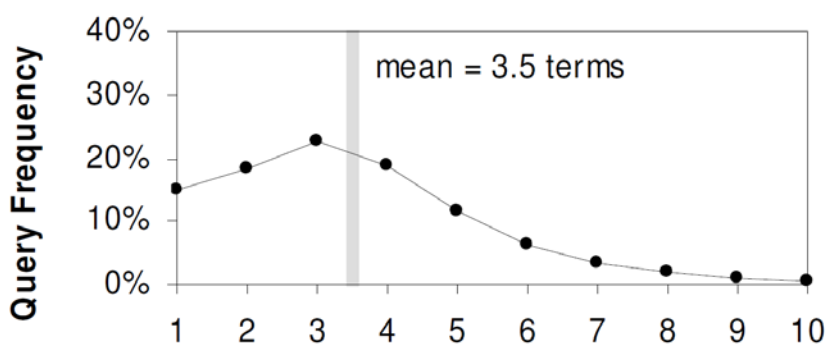
- **AltaVista (1999):** 285 million users, about 1 billion queries
- **AltaVista (2001):** Over 7 million queries

Some web search companies make search logs available for research use under a strict license

- These logs allow knowledge to be discovered and disseminated
- But … many researchers cannot get access

Web Search Query Length

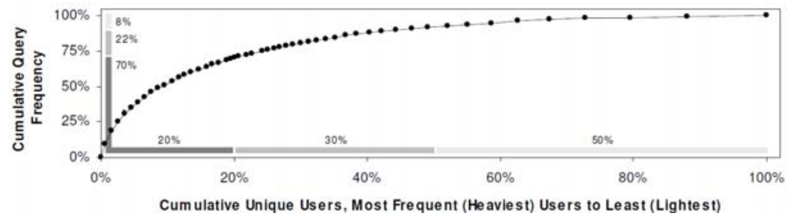
Web queries tend to be short



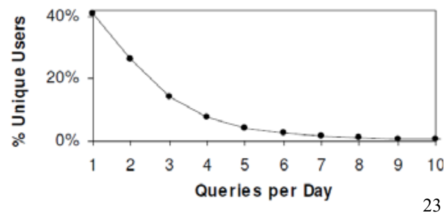
(Pass, et al., 2006)

Who Queries?

A few people issue most of the queries

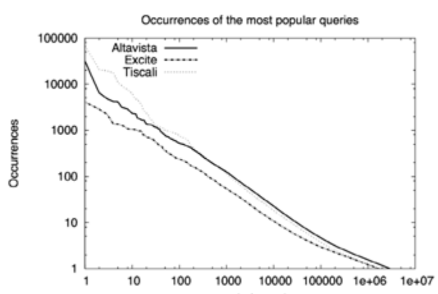


Most people don't search much



(Pass, et al., 2006)

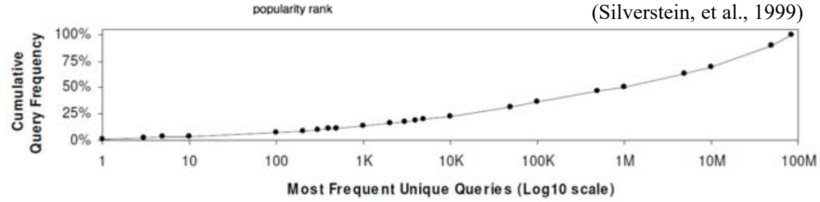
Query Frequency Follows a Power Law



AltaVista log

Freq	Percent
1	63.7%
2	16.2%
3	6.5%
> 3	13.6%

(Fagni, et al., 2006)
(Silverstein, et al., 1999)



(Pass, et al., 2006)

Query Frequency

Query frequency follows a power law

Frequency(q) = K × Rank(q)$^{-\alpha}$

K: Constant, positive

Rank(q): Popularity rank (r=1 is most popular)

α: Constant, about 2.4 for the Excite query log

Note the similarity to Zipf's law

- Same shape, different slope

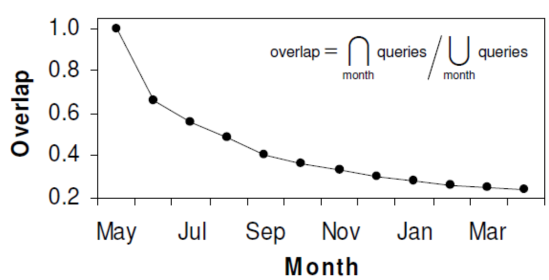
Implications

- A small percentage of the (unique) queries are very common
- Most (unique) queries occur very rarely

25

The Most Frequent Queries Vary Over Time

From month to month



From year to year

- Sex much more of a focus in the late 1990s than now

(Pass, et al., 2006)

26

Query Frequency

Two interesting statistics

- **20% of all queries seen each day have never been seen before (50% of all unique queries seen each day)**
  - *White, et al., 2007*
  - *Amit Singhal, Google, 2010*

    http://googlepolicyeurope.blogspot.com/2010/02/this-stuff-is-tough.html
- **8% of the queries are names**
  - *Amit Singhal, Google, 2010*

27

Queries Vary Geographically

(Pass, et al., 2006)

28

Lecture Outline

- Introduction to search logs
- **Users and tasks**
- Segmenting search logs into sessions

29

Who Uses Web Search for What? And How?

Web search behavior can be modeled along three dimensions

- **Query topics ("what?")**
  - E.g., topics (categories) in the Yahoo! Directory (*a controlled vocabulary*)
- **User demographics ("who?")**
  - E.g., provided by the user (age, gender)
  - E.g., inferred from the user's zip code
    - » income, educational level, political party affiliation
- **Session characteristics ("how?")**
  - E.g., Session length, number of queries/session
  - E.g., % of queries with low/high click entropy
    - » Variation in the documents people click on

(Weber and Jaimes, 2011)

30

The Yahoo! Directory

Who Uses Web Search for What? And How?

Data source:

- A large sample of a Yahoo! search engine query log (2008-2009)
- Registered Yahoo! users
  - U.S. users (user-provided information, U.S. search site)
  - Active users (> 100 queries during the sample period)
  - Not bots (proprietary algorithm)

Data size

- 2.3 million users

Cluster users based on the types of queries they issue

(Weber and Jaimes, 2011)

Who Uses Web Search for What? And How?: Representing Users

Get <user$_i$, query$_j$> pairs from logs

- <jackpgh98, “ingmar weber”>
- <jackpgh98, “search log analysis”>

Create pseudo documents for users

- **Title:** A user id
- **Contents:** The Yahoo! Directory categories of the top 10 documents for each query

Use your favorite similarity metric

- E.g., Jenson-Shannon Divergence, cosine correlation

Pseudo document

<DOC>
<TITLE> jackpgh98 </TITLE>
<BODY>
Computers and Internet / Information Technology,
Computers and Internet / People,
Higher Education / College and University Teaching,
Science / Information Architecture and Design,
…
</BODY>
</DOC>

(Weber and Jaimes, 2011)

Who Uses Web Search for What? And How?: Representing Users

q_{18} top 10 results:

1. d_{18}: c_{13}, c_{47}, c_{82}
2. d_{27}: c_{22}, c_{47}, c_{91}, c_{34}
3. d_{93}: c_{13}, c_{82}
4. …

q_{27} top 10 results:

1. d_{99}: c_{92}, c_{47}, c_{81}
2. d_{47}: c_{37}, c_{92}
3. …

c_i : Category i

Pseudo document

<DOC>
<TITLE> jackpgh98 </TITLE>
<BODY>
c_{13} c_{47} c_{82} c_{22} c_{47} c_{91} c_{34} c_{13} c_{82} …
c_{92} c_{47} c_{81} c_{37} c_{92} …
</BODY>
</DOC>

Who Uses Web Search for What? And How?: Finding Similar Users

Each user is represented by a pseudo document

jackpgh98

```
<DOC>
<TITLE> jackpgh98 </TITLE>
<BODY>
Computers and Internet /
Information Technology, Computers
and Internet / People,
Higher Education / College and
University Teaching,
Science / Information Architecture
and Design,
…
</BODY>
</DOC>
```

irguy214

```
<DOC>
<TITLE> irguy214 </TITLE>
<BODY>
Higher Education / College and
University Teaching,
Computers and Internet / Object-
Oriented Programming,
Science / Information Architecture
and Design,
Computers and Internet /
Linguistics,
…
</BODY>
</DOC>
```

kimfan1893

```
<DOC>
<TITLE> kimfan1893 </TITLE>
<BODY>
Television Shows / Reality
Television,
Television Shows / Society and
Culture,
Television Shows / Comedy,
Television Shows / Reality
Television,
…
</BODY>
</DOC>
```

…

Use your favorite similarity metric to find similar users

- E.g., Jenson-Shannon Divergence, cosine correlation, ….

These ideas are used repeatedly in search engines

- Product search, company search, people search, …

35

Who Uses Web Search for What? And How?: Finding Similar Users

Topics: Cluster users in the "what" dimension

- Representations are based on Yahoo Directory categories
  - i.e., controlled vocabulary terms

Use the other two dimensions to investigate the groups

- "Who": Demographic information
- "How": How people search

Manually label groups based on distinctive characteristics

- Manual, thus possibly subjective labels (but still useful)

(Weber and Jaimes, 2011)

36

Who Uses Web Search for What? And How?: Informational Users

What do they search for?

- Wide range of topics
  - Little interest in adult content

How do they search?

- More likely to issue non-navigational queries
- Less likely to have single-click sessions
- More likely to use query suggestions

Who is in this group?

- More likely to be well-educated
- More likely to have above-average income

(Weber and Jaimes, 2011)

37

Who Uses Web Search for What? And How?: Navigational Users

What do they search for?

- Dominated by popular websites (Facebook, YouTube, Craigslist)

How do they search?

- More likely to issue navigational queries
- More likely to have single-click sessions
- Less likely to use query suggestions

Who is in this group?

- Mostly representative of the entire population

(Weber and Jaimes, 2011)

38

Who Uses Web Search for What? And How?: Transactional Users

What do they search for?

- Shopping, adult content, gaming

How do they search?

- Somewhat similar to navigational users
  - But, multiple sites can perform the transaction
  - Diverse clicks
- Short interaction with search engine

Who is in this group?

- Depends heavily on the type of transaction
- Topic "recreation/games" associated with low income & education

(Weber and Jaimes, 2011)

39

Who Uses Web Search for What? And How?: Selected Groups

Baby boomers

- Who: 50 years old
- What: Interested in finance
- How: Simple navigational queries related to online banking

Challenged youth

- Who: Average age of 34
- Who: Low-income neighborhoods with low-level of education
- What: Interested in music
- How: Navigational sessions

(Weber and Jaimes, 2011)

40

Who Uses Web Search for What? And How?: Selected Groups

Liberal females

- Who: Mostly female from areas that voted Democratic
- What: Shopping queries
- How: Long sessions (browsing and comparison)

White conservatives

- Who: Mostly male from areas that voted Republican
- What: Interested in automotive, business, home & garden

(Weber and Jaimes, 2011)

Who Uses Web Search for What? And How?: Selected Groups

Older users: Health / diseases & conditions, gambling, travel

People in their late 20s: Health / fitness, reproductive health

Younger people: Games, education

Low income: Music, comics & animation, military

Asian descent: Computers & internet, programming & development

Is any of this surprising or useful?

(Weber and Jaimes, 2011)

Who Uses Web Search for What? And How?: Interplay Between What and How

Some topics typically receive few clicks
- News & media, society & culture, computers & internet

People are more likely to click on suggestions for some topics
- Health, science, arts

People with higher educational levels…
- Tend to have shorter sessions
- Click on query suggestions less often
- Are more likely to submit tail queries

(Weber and Jaimes, 2011)

Who Uses Web Search for What? And How?

Observations from query log analysis are useful for designing personalization strategies
- However, you have to figure out how to turn observations into useful strategies

(Weber and Jaimes, 2011)

Lecture Outline

- Introduction to search logs
- Users and tasks
- **Segmenting search logs into sessions**

45

Information Seeking in the Real World

Interpreting search logs is an open research problem

- d_1 is clicked at steps 2 and 4 … is it relevant to q_1?
- Are q_1, q_2, and q_3 about the same information need?
- Was the user satisfied with any of the search results?

How do we think about this sequence of interactions?

Search log

q_1
d_1
d_2
d_1
q_2
q_3
d_3
email site
:

q_i: Query
d_j: Clicked page

46

Information Seeking is a Dialogue Between a Person and a Search Engine

Ad-hoc search can be viewed as a *dialogue* about an information need

Person: query	Initial description
Engine: search results	Initial attempt to satisfy it
Person: reformulated query	Revised description
Engine: new search results	Revised attempt to satisfy it
…	

Viewing Search Logs as a Dialogue

Timeline (mm:ss)		Query
00:00	○	nursing registry
04:18	c	certified nursing assistant 1
08:48	c	nursing assistant registry
09:48	c	license look up for nursing assistants
10:06	c	nursing assistant 1 certification
11:42	c	nursing assistant 1 license look ups
12:18	c	nursing assistant 1 expiration look up
12:30	c	nursing registry in Raleigh
13:24	c	nursing aide registry of Raleigh
15:00	+	nursing aide registry of Raleigh website
16:06	<	nursing aide registry of Raleigh
19:48	c	north carolina board of nursing information for nursing assistant 1
22:24	c	license look up for nursing assistant 1
24:36	c	license information for nursing assistant 1 expiration
28:30	c	north carolina nursing assistant 1 license information

(Pass, et al., 2006)

Viewing Search Logs as a Dialogue

The first task is to distinguish the different dialogues

- Which queries address the same information need?

Originally, information need ≈ a search session

- **Session:** A sequence of user actions within a timespan
  - E.g., 30 minutes
- Perhaps an artifact of the experimental conditions
  - Much of the early work was done in a lab

Information need ≈ a search session is beginning to be challenged

- However, we start here because it is still the dominant view

Search log

$\boxed{q_1}$
d_1
d_2
d_1
$\boxed{q_2}$
$\boxed{q_3}$
d_3
email site
:

Viewing Search Logs as a Dialogue

gout	2006-03-01 07:38:03
chemotherapy side effects	2006-03-01 07:42:36
chemotherapy causing hearing loss	2006-03-01 07:45:23
kenny rogers songs	2006-03-02 06:05:40
commerce on line	2006-03-03 04:54:11
broadband internet	2006-03-06 05:32:28
middlesex county college nj	2006-03-06 16:55:56
kean college	2006-03-06 17:02:32
montclair college	2006-03-06 17:10:45
union county college	2006-03-07 04:49:23
rutgers	2006-03-07 05:10:17
kean college	2006-03-07 05:19:22
migraine headache	2006-03-10 06:02:55
new jersey income tax	2006-04-12 06:09:44

How would you segment this log into sessions?

(From AOL search log, part 9)

Segmenting Search Logs into Sessions: Simple Heuristics

Δ Time: Same session iff |timestamp (q_2) – timestamp| (q_1) < Δ

- Often Δ = 30 minutes, but many values have been tried
- Radlinski found 30 minutes to be effective in a library setting
- Jones found no value that is better than random on the web

Common term: Same session iff $q_1 \cap q_2 \neq \emptyset$

- Probably high Precision, low Recall

Rewrite classes: Common reformulation patterns

- E.g., term added, deleted, or replaced
- Probably high Precision, low Recall

Segmenting Search Logs into Sessions: Simple Heuristics

gout	2006-03-01 07:38:03	CT, RC
chemotherapy side effects	2006-03-01 07:42:36	
chemotherapy causing hearing loss	2006-03-01 07:45:23	ΔT, CT, RC
kenny rogers songs	2006-03-02 06:05:40	ΔT, CT, RC
commerce on line	2006-03-03 04:54:11	ΔT, CT, RC
broadband internet	2006-03-06 05:32:28	ΔT, CT, RC
middlesex county college nj	2006-03-06 16:55:56	
kean college	2006-03-06 17:02:32	
montclair college	2006-03-06 17:10:45	ΔT
union county college	2006-03-07 04:49:23	CT, RC
rutgers	2006-03-07 05:10:17	CT, RC
kean college	2006-03-07 05:19:22	ΔT, CT, RC
migraine headache	2006-03-10 06:02:55	ΔT, CT, RC
new jersey income tax	2006-04-12 06:09:44a	

(From AOL search log, part 9)

Segmenting Search Logs into Sessions: Other Features

gout	2006-03-01 07:38:03	
chemotherapy side effects	2006-03-01 07:42:36	--- CT, RC
chemotherapy causing hearing loss	2006-03-01 07:45:23	
kenny rogers songs	2006-03-02 06:05:40	ΔT, CT, RC
commerce on line	2006-03-03 04:54:11	ΔT, CT, RC

What other features could be used to segment a log?

- Edit distance between queries
- Co-occurrence (e.g., PMI, χ^2) of queries in a query log
- Queries have co-occurring clicks in a query log
- ODP or Yahoo page category overlap of top 10 results
- JSD similarity of top 10 results
- …

Challenges to Recognizing Information Needs In Search Engine Logs

A person's information need may span days or weeks

- E.g., writing a paper, searching for colleges, medical problems

People routinely interleave tasks

- E.g., writing a paper, but take a break to make dinner plans

Typical search behavior reflects tasks and subtasks

- The subtasks may appear distinct when they are actually related

Missions and Goals
(Tasks and Subtasks)

An information need is a single, well-defined goal
- It is represented by a group of queries

A mission is a set of related information needs
- An extended or higher-level information need

Example:
- **Mission:** Find information on hiking in the Pittsburgh area
- **Goal:** Getting to the Laurel Highlands Hiking Trail
- **Goal:** Getting to the Rachel Carson Trail

(Jones and Klinker, 2006)

55

Challenges to Recognizing Information Needs
In Search Engine Logs

Can queries from the same information need or mission be identified automatically?
- **Boundary task:** Given a pair of sequential queries **(easier)**
  – Are they from the same information need ("goal")?
  – Are they from the same information seeking mission?
- **Same task:** Given a pair of queries **(harder)**
  – Are they from the same information need ("goal")?
  – Are they from the same information seeking mission?
- **Note:** We do not know what the goals or missions are
  … but we can still recognize queries that belong together

(Jones and Klinker, 2006)

56

Missions and Goals (Tasks and Subtasks)

	the who, wikipedia	Boundary	Mission: Old music. Goal: The Who
Same	toronto	(mission)	Mission: Toronto. Goal: ?
mission	toronto tourism		Mission: Toronto. Goal: Things to do
	toronto blue jays		Mission: Toronto. Goal: Things to do
	toronto zoo	Boundary	Mission: Toronto. Goal: Things to do
	toronto hotels	(goal)	Mission: Toronto. Goal: Hotels
	usair 2130		
Same	toronto hotel deals		Mission: Toronto. Goal: Hotels
goal	toronto hotels downtown		Mission: Toronto. Goal: Hotels
	sigir 2014		
	toronto restaurants		Mission: Toronto. Goal: Restaurants
	toronto second city		Mission: Toronto. Goal: Things to do
	toronto yorkville		Mission: Toronto. Goal: Things to do
	toronto yorkville hotels		Mission: Toronto. Goal: Hotels
	toronto yorkville restaurants		Mission: Toronto. Goal: Restaurants

57

A Classification-Based Approach to Detecting Pairs of Related Queries

Heuristics work surprisingly well

Sequential queries | Pairs of queries

Features	Goals		Missions	
	Boundary	Same	Boundary	Same
Baseline	63.1%	94.8%	59.9%	70.5%
30 minute	57.2%	90.9%	73.8%	74.4%
Trained time	69.5%	92.6%	75.8%	74.4%
commonw	80.7%	94.9%	79.3%	78.9%
commonw+prisma+time	84.0%		82.1%	

- **Baseline:** Always predicts majority class ('no boundary' or 'different goal')
- **Trained time, goals:** 1.5 min for boundary, 17.2 min for same
- **Trained time, missions:** 6 min for boundary, 47 min for same

(Jones and Klinker, 2006)

58

A Classification-Based Approach to Detecting Pairs of Related Queries

Features

- **Temporal**
  - ≤ {5, 30, 60, 120} minutes, Δ time, are_sequential
- **Edit distance**
  - Several character and token-based metrics
- **Query log**
  - Various types of <q_1, q_2> co-occurrence in a larger query log
- **Web search**
  - Cosine distance of top 50 search results for each query ("prisma")

(Jones and Klinker, 2006)

59

A Classification-Based Approach to Detecting Pairs of Related Queries

A trained classifier is somewhat more effective than heuristics

Features	Goals		Missions	
	Boundary	Same	Boundary	Same
Baseline	63.1%	94.8%	59.9%	70.5
Commonw+cosine+time	84.0%		82.1%	
All features	87.3%	97.1%	84.4%	88.4%
Levenshtein distance	85.0%	95.2%	78.2%	77.0%
commonw+time	81.5%	95.3%	79.3%	78.9%

Metric: Classifier accuracy. Differences are statistically significant.

(Jones and Klinker, 2006)

60

Segmenting and Organizing Query Logs

There is more recent work, but the main message hasn't changed

- **Predict whether two queries are for the same information need**
  - Adjacent queries: 85-90% accuracy
  - Any pair of queries: 95-97% accuracy
    - » Higher because the negative class is very common
- **Classifiers are best, but the best heuristics aren't far behind**
  - Edit distance is very effective
  - Cosine distance among results is effective
  - Time alone is primitive
    - » But effective in combination with other heuristics
    - » Still a very commonly-used heuristic

Lecture Outline

- **Introduction to search logs**
- **Users and tasks**
- **Segmenting search logs into sessions**

For More Information

- Z. Dai and J. Callan. Deeper text understanding for IR with contextual neural language modeling. SIGIR 2019.
- Z. Dai and J. Callan. Context-aware sentence/passage term importance estimation for first stage retrieval. arXiv:1910.10687. 2019.
- J. Dalton, C. Xiong, and J. Callan. CAsT 2019: The Conversational Assistance Track overview. TREC 2019.
- J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arxiv: 1810.04805. 2018.
- J. Lin, R. Nogueira, and A. Yates. Pretrained transformers for text ranking: BERT and beyond. arxiv 2010.06467. 2020.
- R. Nogueira, W. Yang, J. Lin, and K. Cho. Document expansion by query prediction. arxiv 1904.08375. 2019.
- T. Fagni, R. Perego F. Silvestri, and S. Orlando. "Boosting the performance of Web search engines: Caching and prefetching query results by exploiting historical usage." *ACM Transactions on Information Systems (TOIS)*, 24(1), pp 51-78. 2006.
- R. Jones and K.L. Klinker. "Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs." In *Proceedings CIKM '08*. 2008.
- R. Jones, B. Rey, O. Madani, and W. Greiner. "Generating query substitutions." WWW 2006.
- G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *InfoScale '06: Proceedings of the 1st International Conference on Scalable Information Systems.* 2006.
- F. Radlinski, M. Szummer, and N. Craswell. "Inferring query intent from reformulations and clicks." WWW 2010.
- C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. "Analysis of a very large web search engine query log." *SIGIR Forum*, 33(1), pp 6-12. 1999.
- I. Weber and A. Jaimes. Who uses web search for what: and how. WSDM 2011. 2011.
- R. White, M. Bilenko, and S. Cucerzan. "Studying the use of popular destinations to enhance web search interaction." *Proceedings of SIGIR 2007*. 2007.