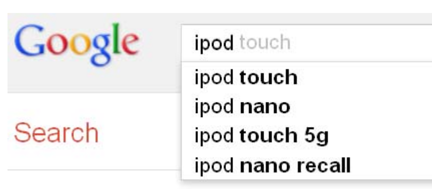
11-642:
Search Engines

Search Log Analysis

Jamie Callan and Jon Elsas
Carnegie Mellon University
callan@cs.cmu.edu

Lecture Outline

- **Query suggestions**
- **Query intents**
- **Click models**
- **Within-session learning**



Search Trails

A search trail is a single information seeking session

– One person, a series of queries, all within a short timespan
– Maybe the first query didn't find what the person wanted
– Maybe the sequence corresponds to query reformulations

Example	**Counterexample (?)**
ipod	ipod video discount
ipod nano	ipod video rebate
ipod nano recall	ipod video repair
ipod nano recall 2011	refurbished ipod video

3

Query Suggestions

Popular destinations for <u>frequent</u> queries can be looked up
… however, most queries are not frequent

- 57% of queries (20% of all searches) are unique
  97% of queries (66% of all searches) occur less than 10 times
  – *White, et al., 2007; 2008*

We cover two types of query suggestion approaches

- **Pseudo documents**
  – Works well for many (most?) queries
- **Co-occurrence statistics**
  – Works well for reasonably frequent queries

4

Query Suggestions #1: Pseudo Documents

Obtain <query$_i$, query$_{last}$> pairs from logs

- <"ipod", "ipod nano recall 2011">
- <"ipod nano", "ipod nano recall 2011">
- : : : :
- <"ipod repair", "ipod nano recall 2011">

Assume the last query in a session is successful

- Other success criteria is covered later in the course

Each query$_{last}$ is a candidate query suggestion

Query logs

: : : :
-- Session start --
ipod
ipod nano
ipod nano recall
ipod nano recall 2011
-- Session end –
: : : :
-- Session start --
ipod repair
ipod nano recall 2011
-- Session end –
: : : :

5

Query Suggestions #1: Pseudo Documents

Create a "pseudo document" for each candidate query suggestion

- The title is the query suggestion (query$_{last}$)
- The contents are the queries that preceded query$_{last}$ in the query log

Given a new query, rank the suggestion candidates

- Apply your favorite retrieval model to the pseudo document corpus

Pseudo documents are a very general technique

- Used to rank all kinds of things besides queries

Pseudo document for a suggestion

<TITLE>
ipod nano recall 2011
</TITLE>
<BODY>
ipod, ipod nano,
ipod shuffle,
music players,
small ipod, ipod,
ipod micro, nanno,
buy ipod, …
</BODY>

6

Query Suggestions #2: Co-occurrence Statistics

Select queries that contain the user query q as a substring

- **Set_1:** 100 most frequent queries with q as a substring
- **Set_2:** 100 most frequent queries that followed q in a search session
- All queries in Set_1 and Set_2 are potential query suggestions q_s

$$Score(q_s) = \frac{Count(q_s) + \lambda_1}{N_1 + \lambda_1} \times \frac{Count_{follows}(q, q_s) + \lambda_2}{N_2 + \lambda_2}$$

q_s is frequent **q_s frequently follows q**

Jamie's understanding of N_1 and N_2 (the papers are vague)

- N_1: Sum of Count(q_s) for all candidates
- N_2: Sum of Count(q_s, q) for all candidates

(White, et al., 2007; 2008)

7

Query Suggestions #2: Co-occurrence Statistics

Select queries that contain the user query q as a substring

- **Set_1:** 100 most frequent queries with q as a substring
- **Set_2:** 100 most frequent queries that followed q in a search session
- All queries in Set_1 and Set_2 are potential query suggestions q_s

$$Score(q_s) = \frac{Count(q_s) + \lambda_1}{N_1 + \lambda_1} \times \frac{Count_{follows}(q, q_s) + \lambda_2}{N_2 + \lambda_2}$$

q_s is frequent **q_s frequently follows q**

This baseline method does as well as two experimental methods

(White, et al., 2007; 2008)

8

Query Suggestions: Summary

Three methods for generating 'related query' suggestions

- Pseudo documents
  - Almost any query
- Co-occurrence #1
  - Mid-frequency queries
- Co-occurrence #2
  - Common queries

Lecture Outline

- Query suggestions
- **Query intents**
- Click models
- Within-session learning

Many Queries Have Multiple Intents

We have discussed the different intents behind some queries

- **jaguar:** A car, an animal, an operating system, …
- **flash:** Software, a superhero, part of a camera, …
- **mercury**: An element, a planet, a god, a car, …
- **michael jordan:** An athlete, a professor, a businessman, …
- **ai:** Artificial intelligence, Americal Idol, art institute, …

Query suggestions are one possible source of query intents

- Can they be inferred in other ways?

(Radlinski, et al., 2010)

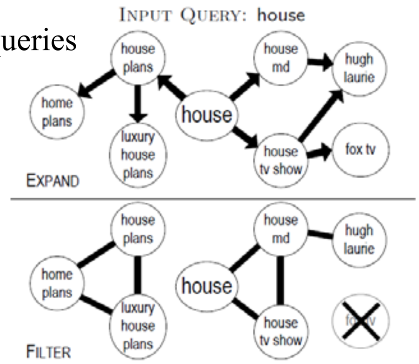
Identifying the Most Common Intent for a Query

For a query q

- Identify q's neighborhood **("expand step")**
  - Identify the 10 most common reformulations q' of q
  - Identify the 10 most common reformulations q'' of each q'
- Reduce the neighborhood to the most related queries **("filter")**
- **Cluster** the queries to find intent groups
- **Estimate the popularity** of each query and intent group
- The **name** of an intent group is its most popular query

(Radlinski, et al., 2010)

Identifying the Most Common Intent for a Query



Expand: Identify possibly-related queries

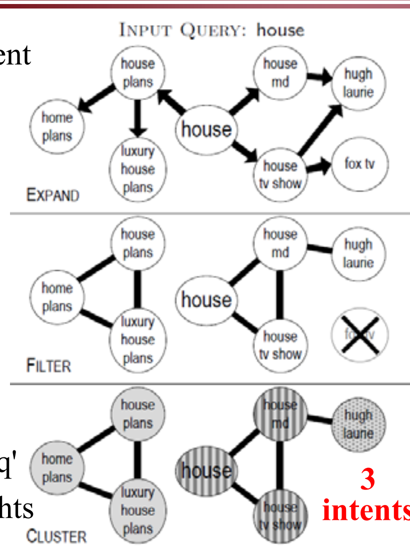
- (q, q') for at least 2 users
- $q'_{time} - q_{time} < 10$ min
- (q, q') must occur $\geq \delta$ times
  - Filters out frequent q' (e.g., hotmail)

Filter: Improve precision

- Connect (q, q') if often clicked for same d
  - Removes many q'
  - May add some new q'
- Remove components with < 2 members

(Radlinski, et al., 2010)

Identifying the Most Common Intent for a Query



Cluster: Find groups with same intent

- E.g., use your favorite algorithm
- Similarity metric: Random walk

Estimate popularity

- Random walk for two iterations

$$w_q = 1$$

$$w_{q'} = \frac{w^q \cdot count(q \to q')}{\sum_{r \in R(q)} count(q \to r)}$$

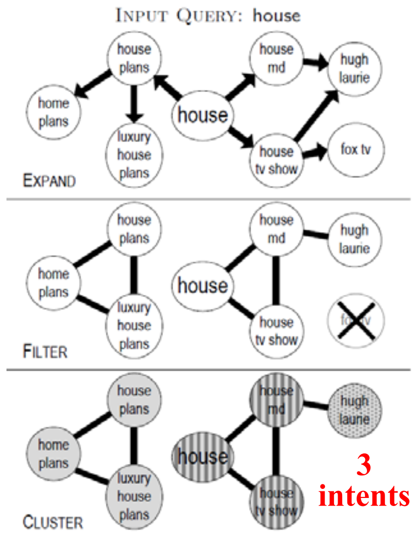
- R(q): the set of all related queries q'
- Cluster score: Sum of query weights

(Radlinski, et al., 2010)

Identifying the Most Common Intent for a Query

Name the intent group

- Use its highest-scoring query



15

(Radlinski, et al., 2010)

Identifying the Most Common Intents for a Query

juvenile delinquency

• juvenile delinquency	($w_c = 1.16$)	w_c: estimated relative popularity of each intent cluster
• causes of juvenile delinquency	($w_c = 0.50$)	
• delinquency prevention	($w_c = 0.25$)	
• definition of juvenile delinquency	($w_c = 0.20$)	
• articles on juvenile delinquency	($w_c = 0.18$)	
• reasons for juvenile delinquency	($w_c = 0.15$)	

Intents created by TREC analysts

- What are the rates of juvenile crime in various jurisdictions, what is the nature of the offenses, how are they punished, and what measures are taken for prevention?

16

(Radlinski, et al., 2010)

Identifying the Most Common Intents for a Query

physical therapists

• physical therapist	(w_c = 1.22)	w_c: estimated relative popularity of each intent cluster
• physical therapists salary	(w_c = 0.80)	
• how to become a physical therapist	(w_c = 0.21)	
• physical therapy schools in california	(w_c = 0.15)	
• physical therapist school of california	(w_c = 0.11)	
• physical therapist assistance programs	(w_c = 0.11)	

Intents created by TREC analysts

- How can I obtain information about training, licensing, and skills needed for physical therapists?

17

(Radlinski, et al., 2010)

Identifying the Most Common Intents for a Query

wireless communications

• wireless communications	(w_c = 1.07)	w_c: estimated relative popularity of each intent cluster
• what is wireless comm.	(w_c = 0.56)	
• wireless comm. systems	(w_c = 0.19)	
• history wireless technology	(w_c = 0.13)	
• wireless cell phone companies	(w_c = 0.13)	
• wireless broadband providers	(w_c = 0.10)	

Intents created by TREC analysts

- Information on existing and planned uses, research/technology, regulations and legislative interest.

18

(Radlinski, et al., 2010)

Identifying the Most Common Intents for a Query

Key ideas

- Reformulation patterns + click information can be combined to identify common intents for ambiguous queries
- An ambiguous query may have several common intents
  - Not a surprise ☺
  - An intent is expressed by a group of queries with the same goal
- Popular intents may differ from what well-informed people expect

Lecture Outline

- Query suggestions
- Query intents
- **Click models**
- Within-session learning

Click Models

Typical search behavior

- A person issues a query q
- A **s**earch **e**ngine **r**esult **p**age (SERP) is returned
- The person examines the SERP
  - – Maybe the person clicks on one or more links
- The person stops interacting with the SERP
  - – Perhaps issues a new query
  - – Perhaps moves on to a new task

(Chuklin, et al., 2016)

21

Click Models

Click models represent user search behavior as a sequence of observed and hidden events

- E: An item on the SERP is **e**xamined
  - – May depend on the document's rank (position)
- A: User is **a**ttracted by the item's representation
  - – May depend on the snippet quality
  - – May depend on the document's relevance to the query
- C: An item is **c**licked
- S: The information need is **s**atisfied

(Chuklin, et al., 2016)

22

Click Models

Click models define dependencies among events

- E.g., p(E | rank=r): Probability of examining page at rank r
- E.g., p(C | E): Probability that an examined page is clicked

Learn model parameters from a search log

(Chuklin, et al., 2016)

23

Random Click Model (RCM)

Random Click Model

- Assume each document (url) u is equally likely to be clicked

  $p(C_u) = \rho$ Probability of clicking document (url) u
- Learn ρ from data

Maximum likelihood estimate of ρ

$$\rho = \frac{1}{\sum_{s \in S} |s|} \sum_{s \in S} \sum_{u \in s} c_u^s$$

S: Search sessions

c_u^s: Document (url) u was clicked in session s

The RCM is not very accurate

(Chuklin, et al., 2016)

24

Click-Through Rate (CTR) Models: RCTR

Rank-Based CTR (RCTR)

- The click likelihood depends upon the rank *r*

  p (C_r) = ρ_r Probability of clicking document at rank *r*
- Learn ρ_r for each rank *r*

Maximum likelihood estimate of ρ_r

$$\rho_r = \frac{1}{|S|} \sum_{s \in S} c_r^s$$

S: Search sessions

c_r^s: Document at rank r was clicked in session *s*

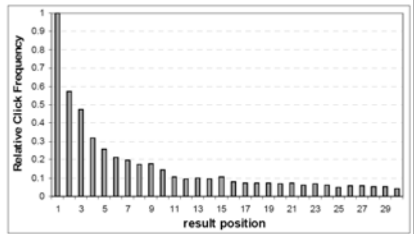


Figure 3.1: Relative click frequency for top 30 result positions over 3,500 queries and 120,000 searches.

(Agichtein, et al, 2006)
(Chuklin, et al., 2016)

25

Click-Through Rate (CTR) Models: DCTR

Document-Based CTR (DCTR)

- Estimate a click-through rate for every query-document pair

  p (C_u) = ρ_{uq} Probability of clicking document *u* for *q*
- Learn ρ_{uq} for each query-document pair

Maximum likelihood estimate of ρ_{uq}

$$\rho_{uq} = \frac{1}{|S_{uq}|} \sum_{s \in S_{uq}} c_u^s$$

S_{uq}: Sessions for query q that contain document u

c_u^s: Document u was clicked in session *s*

DCTR is prone to overfitting

- Little or no data for many document-query pairs

(Chuklin, et al., 2016)

26

Attractiveness

Some models incorporate the document attractiveness

- Usually attractiveness is a measure of snippet quality
- Usually $p(A_u)$ is independent of rank r

  $p(A_u) = \alpha_{uq}$ Probability that document u is considered attractive

(Chuklin, et al., 2016)

27

Position-Based Model (PBM)

Position-Based Model (PBM)

- Click probability depends on examination and attractiveness

  $p(C_u) = p(E_u) \cdot p(A_u)$

  $p(E_u) = \gamma_{r_u}$ Document at rank r_u is examined

  $p(A_u) = \alpha_{uq}$ Document u is attractive for query q

Model parameters can be estimated using EM

$$\alpha_{uq} \quad \alpha_{uq}^{(t+1)} = \frac{1}{|\mathcal{S}_{uq}|} \sum_{s \in \mathcal{S}_{uq}} \left(c_u^{(s)} + \left(1 - c_u^{(s)}\right) \frac{\left(1-\gamma_r^{(t)}\right)\alpha_{uq}^{(t)}}{1-\gamma_r^{(t)}\alpha_{uq}^{(t)}} \right),$$

where $\mathcal{S}_{uq} = \{s_q : u \in s_q\}$

$$\gamma_r \quad \gamma_r^{(t+1)} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(c_u^{(s)} + \left(1 - c_u^{(s)}\right) \frac{\left(1-\alpha_{uq}^{(t)}\right)\gamma_r^{(t)}}{1-\gamma_r^{(t)}\alpha_{uq}^{(t)}} \right)$$

(Chuklin, et al., 2016)

28

Cascade Model (CM)

Assume that the user examines the SERP from top to bottom until they find a relevant document

• $p(C_r) \leftrightarrow p(E_r) \cdot p(A_r)$	Clicked ↔ Examined & Attractive
• $p(E_r) = 1$	First rank is always examined
• $p(A_r) = \alpha_{r_u q}$	Attractiveness depends on rank & query
• $p(E_r \mid \neg E_{r-1}) = 0$	Sequential examination
• $p(E_r \mid E_{r-1}, \neg C_{r-1}) = 1$	Keep examining until click
• $p(E_r \mid C_{r-1}) = 0$	Stop examining after a click

Can only describe sequential examination with a single click

- Less general that the position-based model (PBR)

(Chuklin, et al., 2016)

Cascade Model (CM)

Maximum likelihood estimate for CM

$$\alpha_{uq} = \frac{1}{|S_{uq}|} \sum_{s \in S_{uq}} c_u^s$$

S_{uq}: Sessions for query q that contain document u

Each session *s* is truncated at its first click

c_u^s: Document u was clicked in session *s*

Click Models

There are many click models ...

- E.g., consider time spent examining a SERP item
  - – Time between click and subsequent click
- E.g., consider scrolling behavior
- E.g., consider eye movement behavior
- …

None seem dominant at this point

- DCTR is popular because ρ_{uq} can be treated as an (overfitted) relevance score for training LeToR
- Time-Aware Click Model (TACM) claims to correlate well with human assessors

(Chuklin, et al., 2016)

31

Click Models

How accurate are different models?

- WSCD 2012 dataset (Yandex), 1 million sessions

Model	Log-likelihood	Perplexity	Cond. perplexity	Time (sec)
RCM	−0.3727	1.5325	1.5325	**2.37**
RCTR	−0.3017	1.3730	1.3730	2.45
DCTR	−0.3082	1.3713	1.3713	9.39
PBM	−0.2757	1.3323	1.3323	77.95
CM	$-\infty$	1.3675	$+\infty$	12.17

The Position-Based Model (PBM) is best of this group

- On this dataset, the best models are about 10% better than PBM

(Chuklin, et al., 2016)

32

Click Models

Click models have several uses

- Improve understanding of user behavior
  - – E.g., model what SERP characteristics affect user behavior
- Guide development of better evaluation metrics
  - – E.g., something more realistic than NDCG
- Measure deviation of observed behavior from 'typical' behavior
  - – E.g., people click on this document much more than expected
- Generate realistic artificial data
  - – E.g., for testing software
  - – E.g., infer relevance from clicks
- Very important in web-based advertising

(Chuklin, et al., 2016)

33

Click Models

Click models can be used to generate artificial clicks

Input: Click model M, document ranking s

Output: Vector of simulated clicks $c_1 \ldots c_r$

Algorithm:

for rank = 1 to $|s|$

compute $p(C_r=1|C_1=c_1, \ldots, C_{r-1}=c_{r-1})$ given M and $c_1 \ldots c_{r-1}$

convert $p(C_r=1|C_1=c_1, \ldots)$ to $\{0, 1\}$ using a Bernoulli function

(Chuklin, et al., 2016)

34

Click Models

Click models can be used to generate artificial relevance judgments

- Any individual click is a noisy signal
- Use a click model such as DCTR to learn ρ_{uq} for each query-document pair *uq*
- Measure NDCG@k using ρ_{uq} as the relevance label

Lecture Outline

- Query suggestions
- Query intents
- Click models
- **Within-session learning**

Within-Session Learning

People gain domain knowledge as they search

- Through exposure to and interaction with new information
- Is it useful for the search engine to support this learning?
  - Should the search engine select information differently?

During the last several years, this has gained increasing attention

- We examine one line of work from Microsoft Research (MSR)

Two types of informational information needs were studied

- Acquisition of procedural knowledge ("how to do something")
- Acquisition of declarative knowledge ("knowing about a topic")

(Eickhoff et al., 2014)

37

Within-Session Learning

Identify procedural search sessions in a query log

- Select queries that are last in a session that ends at eHow.com
- Rank query n-grams by pointwise KL-divergence between the eHow query log and a general query log

$$\text{score}(t) = p_{\text{eHowLog}}(t) \log \frac{p_{\text{eHowLog}}(t)}{p_{\text{generalLog}}(t)}$$

t: A query n-gram (e.g., 'fix wifi')

- Select the top 10 n-grams (*indicator terms*)
- Select sessions that contain an indicator term (*procedural sessions*)
- Discard navigational sessions (proprietary classifier)

Repeat for declarative search sessions using Wikipedia.org

(Eickhoff et al., 2014)

38

Within-Session Learning

Indicator terms

Rank	Procedural	Declarative
1	to	what
2	how	what is
3	how to	who
4	how do	list of
5	to make	syndrome
6	how to make	biography
7	how do I	what is a
8	computer	about
9	can you	is the
10	change	history of

A query in a session needed to contain one of these terms

- General terms … but sessions also ended at eHow or Wikipedia

(Eickhoff et al., 2014)

Within-Session Learning

Example queries

Procedural		
#	Query	Clicked URL
1	Weak wireless signal	-
2	How to expand wifi range	http://www.repeaterstore.com/.../fg24008.php
3	Wifi repeater how to	-
4	Wifi repeater tutorial	http://forum.ubnt.com/showthread.php?t=13735
5	How to boost wifi signal	http://www.ehow.com/...boost-wifi-signal.html

Declarative		
#	Query	Clicked URL
1	What do sponges look like	a-z-animals.com/animals/sponge/
2	What do sponges feed on	tolweb.org/treehouses/?treehouse$_i d = 3431$
3	Sponges as pets	http://www.buzzle.com/articles/sponge-facts.html
4	How do sponges reproduce	http://answers.yahoo.com/...127140016AANRe91
5	Where do sponges live	http://en.wikipedia.org/wiki/Sponge

Manual inspection indicates 87% accuracy

(Eickhoff et al., 2014)

Within-Session Learning

Metrics previously used to describe user expertise and behavior

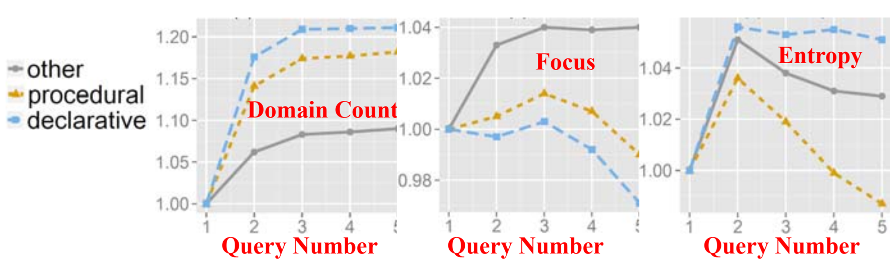
- **Domain count:** Number of unique domains in the SERPs
- **Focus:** % of SERP entries in the dominant category
- **Entropy:** Entropy of the SERPs topic distribution
  – Low entropy → search results cover few topics
- **Branchiness:** How often a person's browsing branches
- **Display time:** Average reading time
  – Experts typically have lower averages
- **Query complexity:** Reading difficulty of query terms

SERP: Search Engine Result Page

(Eickhoff et al., 2014)

Within-Session Learning

Results for sessions of length 5 (similar results for other lengths)

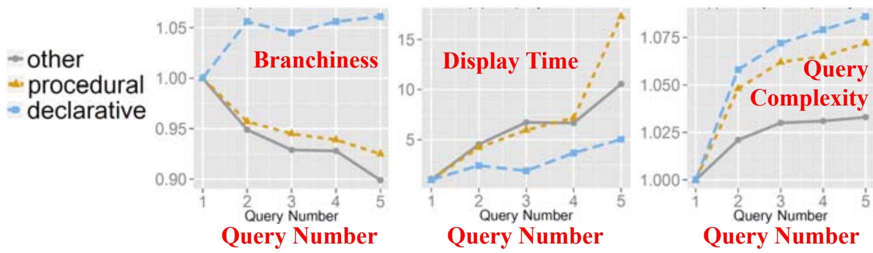


- Broader exploration for procedural and declarative searches
  – Procedural exploration becomes exploration within a topic

(Eickhoff et al., 2014)

Within-Session Learning

Results for sessions of length 5 (similar results for other lengths)



- Declarative searchers explore more paths
- Procedural searchers read more as they zero in on their topic
- Knowledge acquisition queries are more complex than others

(Eickhoff et al., 2014)

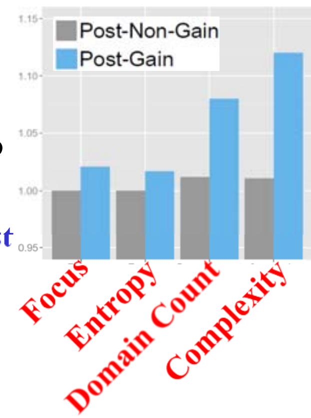
Within-Session Learning

Does expertise acquired in one session persist in later sessions?

- Compare people that gained expertise to people that did not
- Gains for focus and entropy don't persist
- Gains for domain count and complexity do

It's not clear why some gains do not persist

- An open research problem



(Eickhoff et al., 2014)

Within-Session Learning

What is the source of vocabulary in subsequent queries?

	Snippet	Short page visits	Long page visits ($t \geq 30$ sec)	None
D_{other}	0.27	0.13	0.41	0.52
D_{proc}	0.23	0.12	0.44	0.49
D_{decl}	0.26	0.10	0.45	0.49

Recall of subsequent query terms from different sources

Long page visits have almost 4× the influence of short page visits

- Perhaps query suggestions should be generated (mostly? only?) from queries that produce long page visits?

(Eickhoff et al., 2014)

45

Within-Session Learning

Can pages that support knowledge acquisition be identified?

Methodology

- Train a classifier to predict which pages will be followed immediately by an increase in query complexity and domain count

Features

- Text length, sentence length, term length
- Coverage of query terms in title, distribution of query terms
- Part of speech (POS) distribution
- Page complexity, page complexity vs. query complexity
  – Reading difficulty metrics

(Eickhoff et al., 2014)

46

Within-Session Learning

Classifiers: Support Vector Machines (SVM)

Accuracy: F_1=0.76

The best pages had a vocabulary that was a little more complex than the user's observed vocabulary

- Consistent with what is known about other forms of learning

(Eickhoff et al., 2014)

Within-Session Learning

Key ideas

- Mining search logs to identify very specific query types
  - Far beyond "informational, navigational, transactional"
- A sophisticated approach to measuring success
  - Measuring an increase in knowledge vs. measuring clicks
- Rediscovery of ideas known to learning scientists
  - The search engine should help you become smarter
- Advanced work done with familiar tools
  - Search logs, KL divergence, typical features, classification

Microsoft has studied these topics for several years

Lecture Outline

- **Query suggestions**
- **Query intents**
- **Click models**
- **Within-session learning**

For More Information

- E. Agichtein, E. Brill, S. Dumais, and R. Ragno. "Learning user interaction models for predicting Web search result preferences." *Proceedings of SIGIR 2006*. 2006.
- C. Eickhoff, J. Teevan, R. White, and S. Dumais. "Lessons from the journey: A query log analysis of within-session learning." In Proceedings of WSDM '14. 2014
- T. Joachims, L. Granka, and B. Pan. "Accurately interpreting clickthrough data as implicit feedback." *Proceedings of SIGIR 2005*. 2005.
- A. Chuklin, I. Markov, and M. de Rijke. Click models for web search. Morgan and Claypool. 2016.
- F. Radlinski, M. Szummer, and N. Craswell. "Inferring query intent from reformulations and clicks." WWW 2010.
- R. White, M. Bilenko, and S. Cucerzan. "Studying the use of popular destinations to enhance web search interaction." *Proceedings of SIGIR 2007*. 2007.