IIIT-H

Web Mining
Lecture 21: Introduction to Query Log Mining

Manish Gupta

23rd Oct 2013

Slides borrowed (and modified) from
Daxin Jiang, Jian Pei, Hang Li. Web Search/Browse Log Mining: Challenges, Methods, and Applications. Tutorial at WWW 2010

Recap of Lecture 20: Entity Semantics Mining (Part 2)

- Entity Set Expansion
- Entity Acronym Expansion
- Entity Actions
- Entity Tagging

Announcements

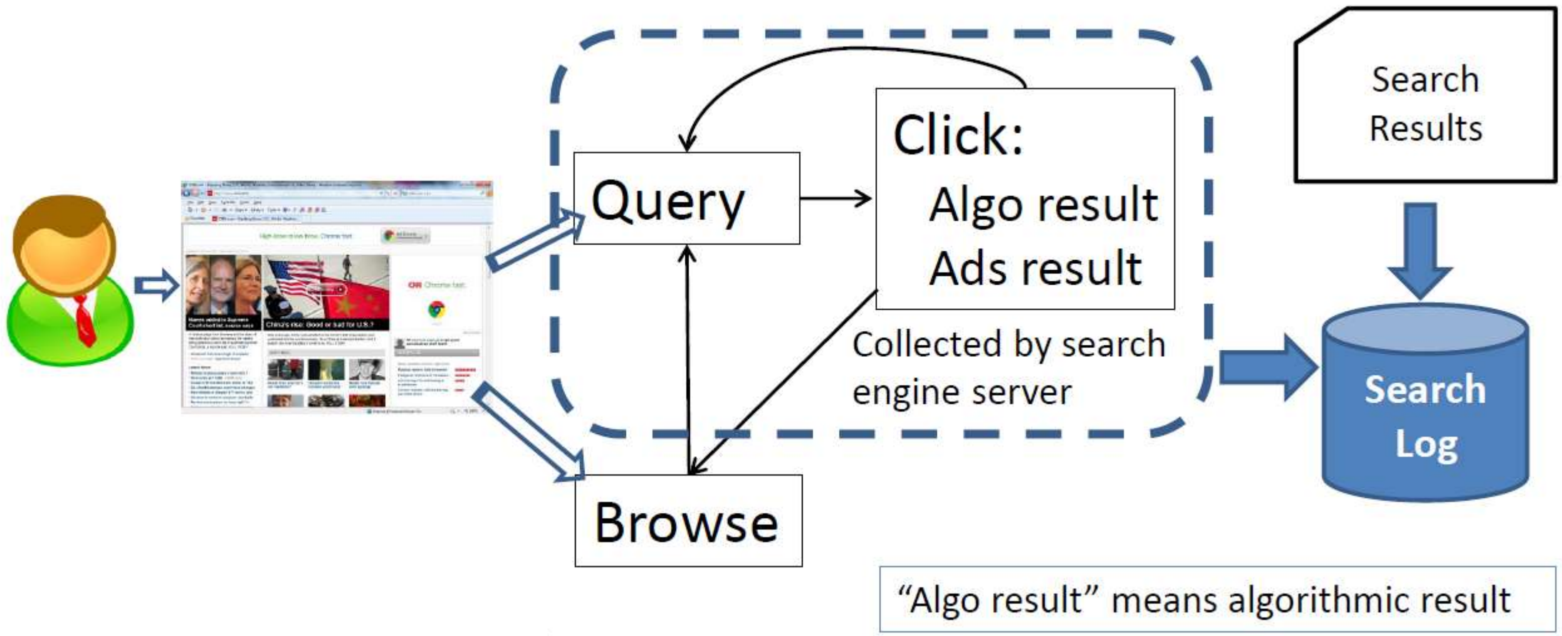
Today's Agenda

- Search and browse logs
- Log mining applications
- Four data structures
- Query Statistics
- Query Classification

Today's Agenda

- **Search and browse logs**
- Log mining applications
- Four data structures
- Query Statistics
- Query Classification

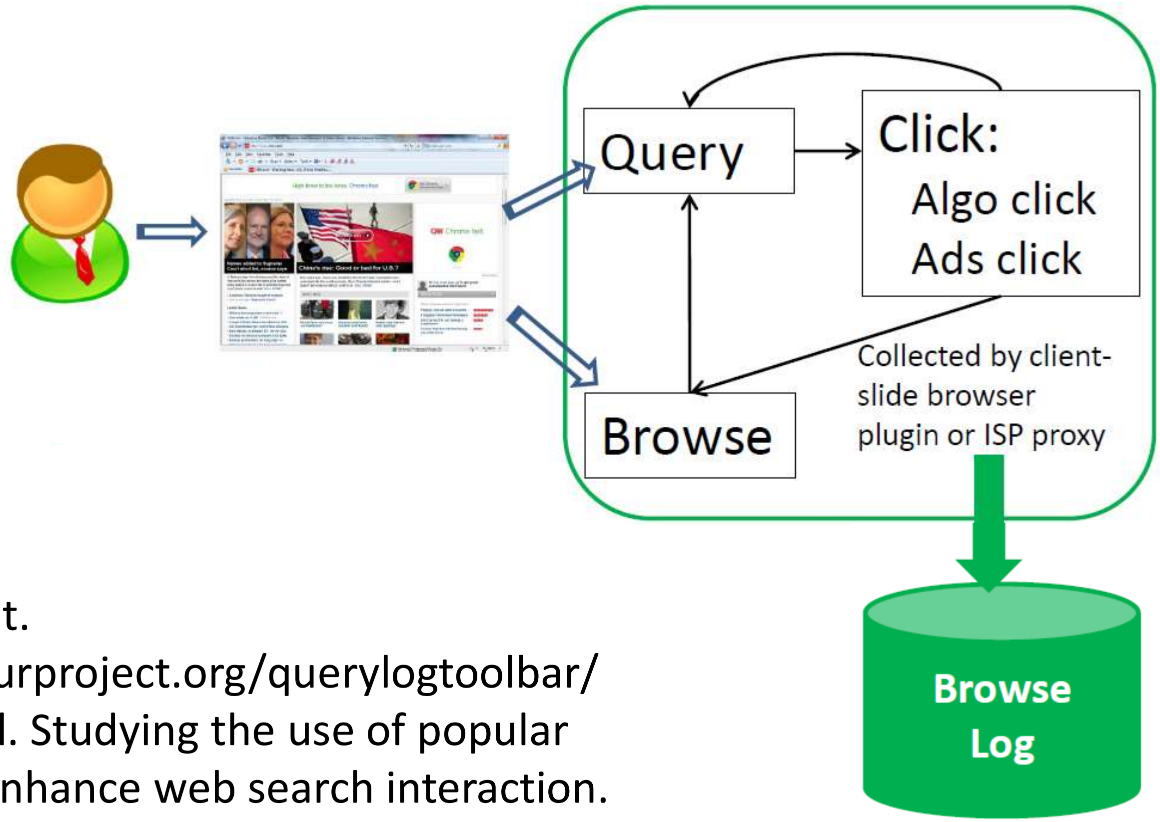
Different Types of Log Data: Search Logs



- Search Logs
  - Collected by search engine server
  - Record the user queries, clicks, as well as the search results provided by the search engine

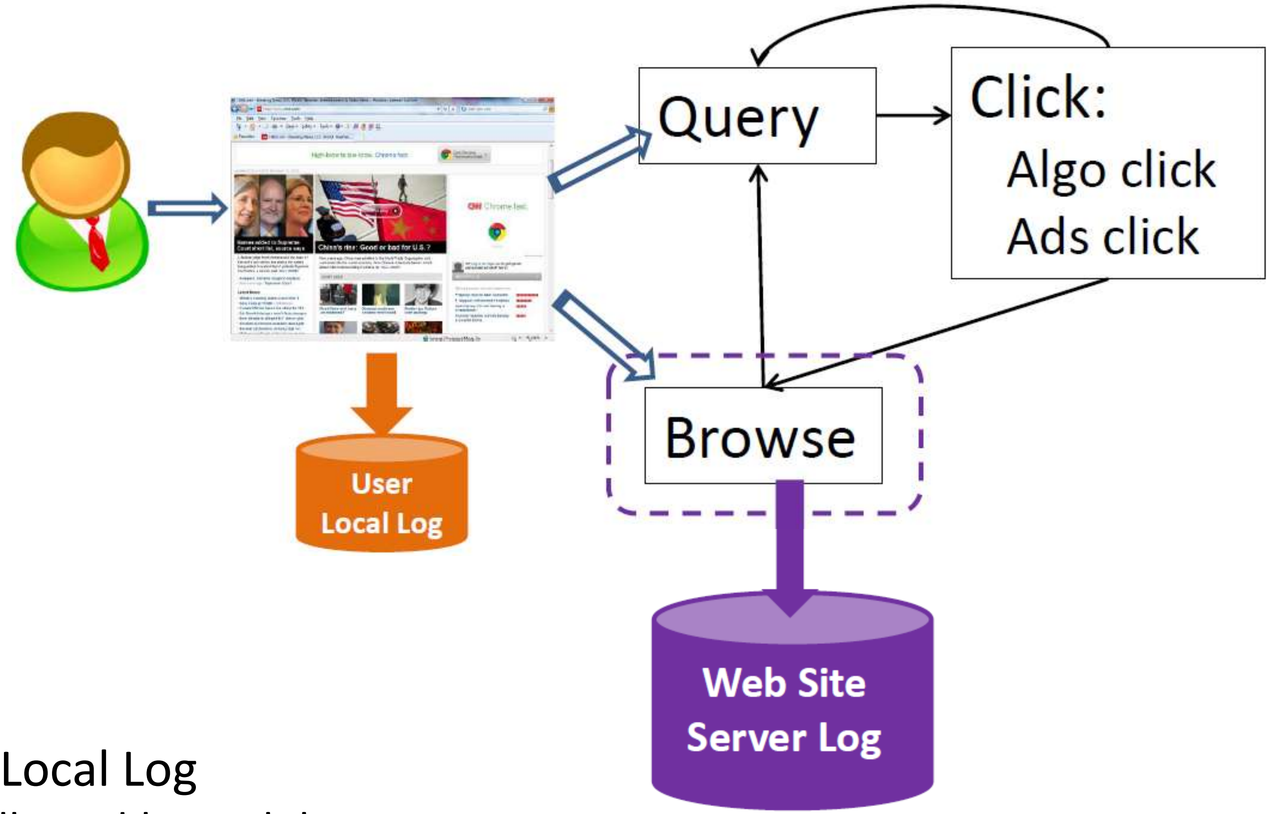
Different Types of Log Data: Browse Logs

- Browse Logs
  - Collected by client-slide browser plugin or ISP proxy
  - Store the user's queries, clicks, and the browsed URLs



- The Lemur toolkit. http://www.lemurproject.org/querylogtoolbar/
- White, R.W., et al. Studying the use of popular destinations to enhance web search interaction. SIGIR'07.

Different Types of Log Data: Other Logs



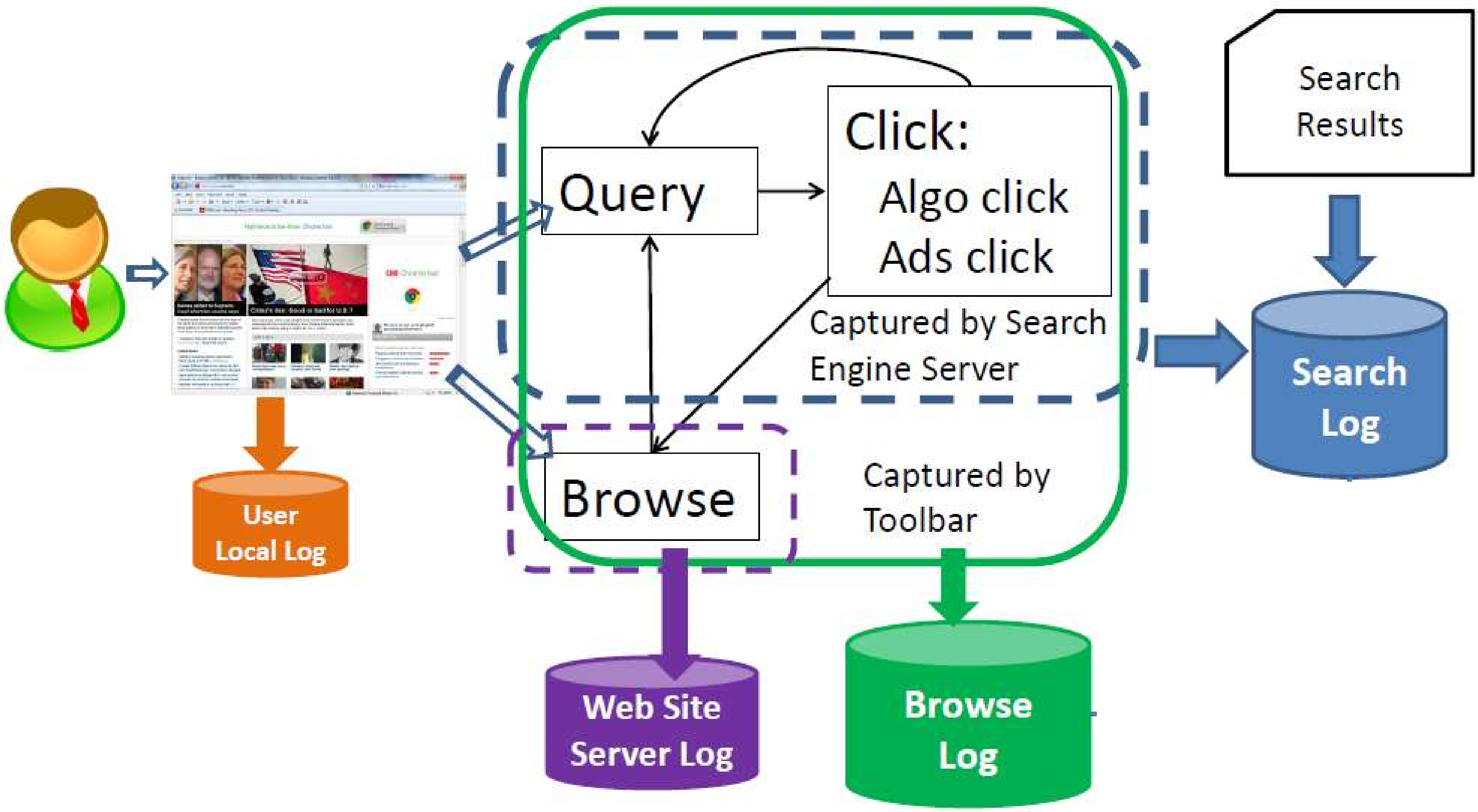
User Local Log

- Collected by Web browser
- Stored on user's local machine
- Contains richer information, e.g., user's every input in browser

Web Site Server Logs

- Each site has its own server logs
- Record how users visit the site

Putting them Together



Major Information in Search Logs

- Recorded by search engine severs
- Four categories information
  - User info: user ID & IP
  - Query info: query text, time stamp, location, search device, etc
  - Click info: URL, time stamp, etc
  - Search results
  - Algo results, Ads results, query suggestions, deep links, instant answers, etc.

Joined to derive the position and type of clicks

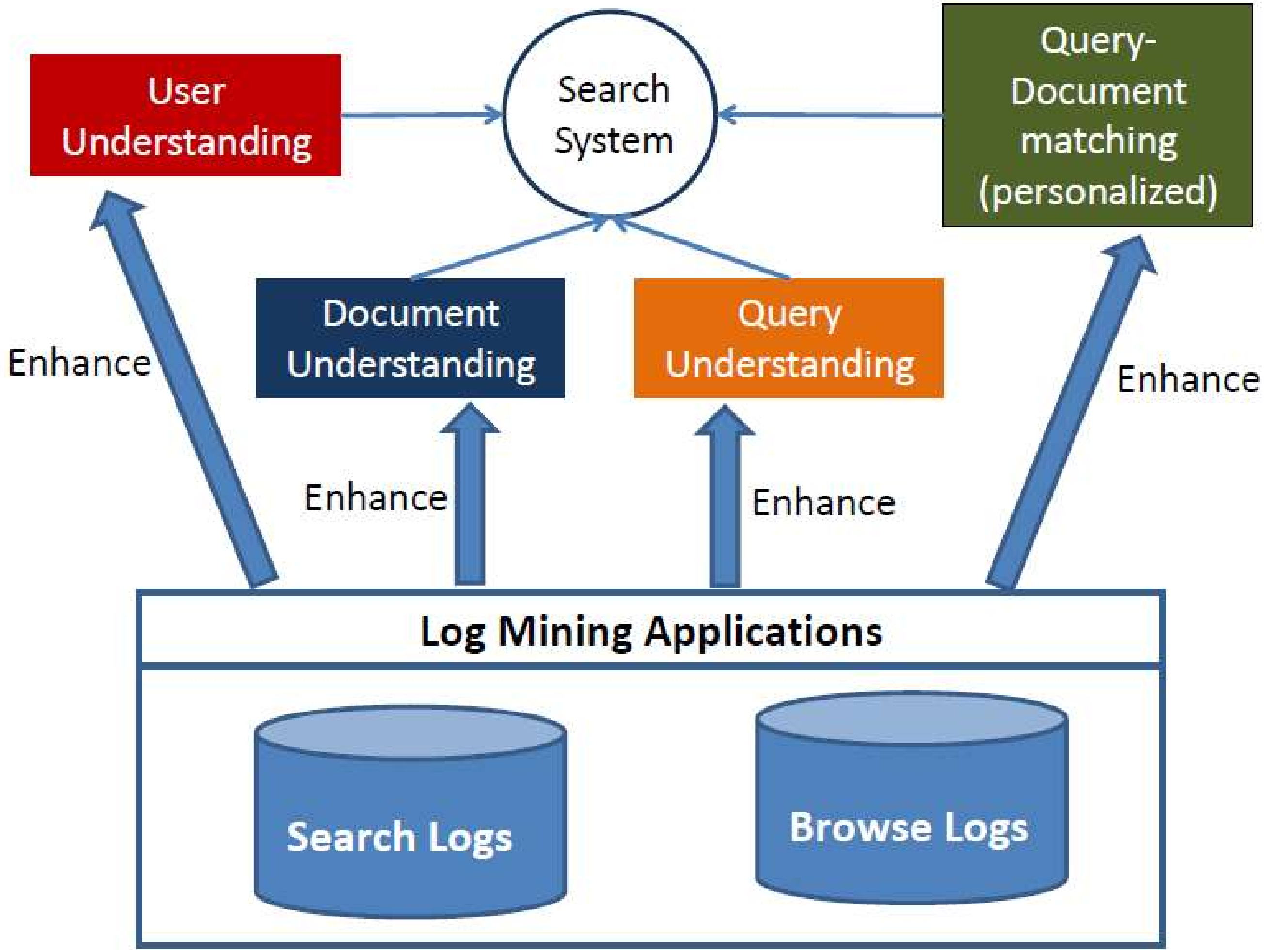
Major Information in Browse Logs

- Captured by client-side browser plug-in or ISP proxy
- Major information
  - User ID & IP, query info, click info
  - Browse info: URL, time stamp
- Client-side browser plug-in has to follow strict privacy policy
  - Only collects data when user's permission is granted
  - User can choose to opt-out at any time

Log Mining Applications

- According to Silvestri and Baeza-Yates [Silvestri09]
  - Enhancing efficiency of search systems
  - Enhancing effectiveness of search systems
- Fabrizio Silvestri and Ricardo Baeza-Yates. Query Log Mining. WWW'09 tutorial
- We only focus on the effectiveness part
- A search system provides both algo and Ads results

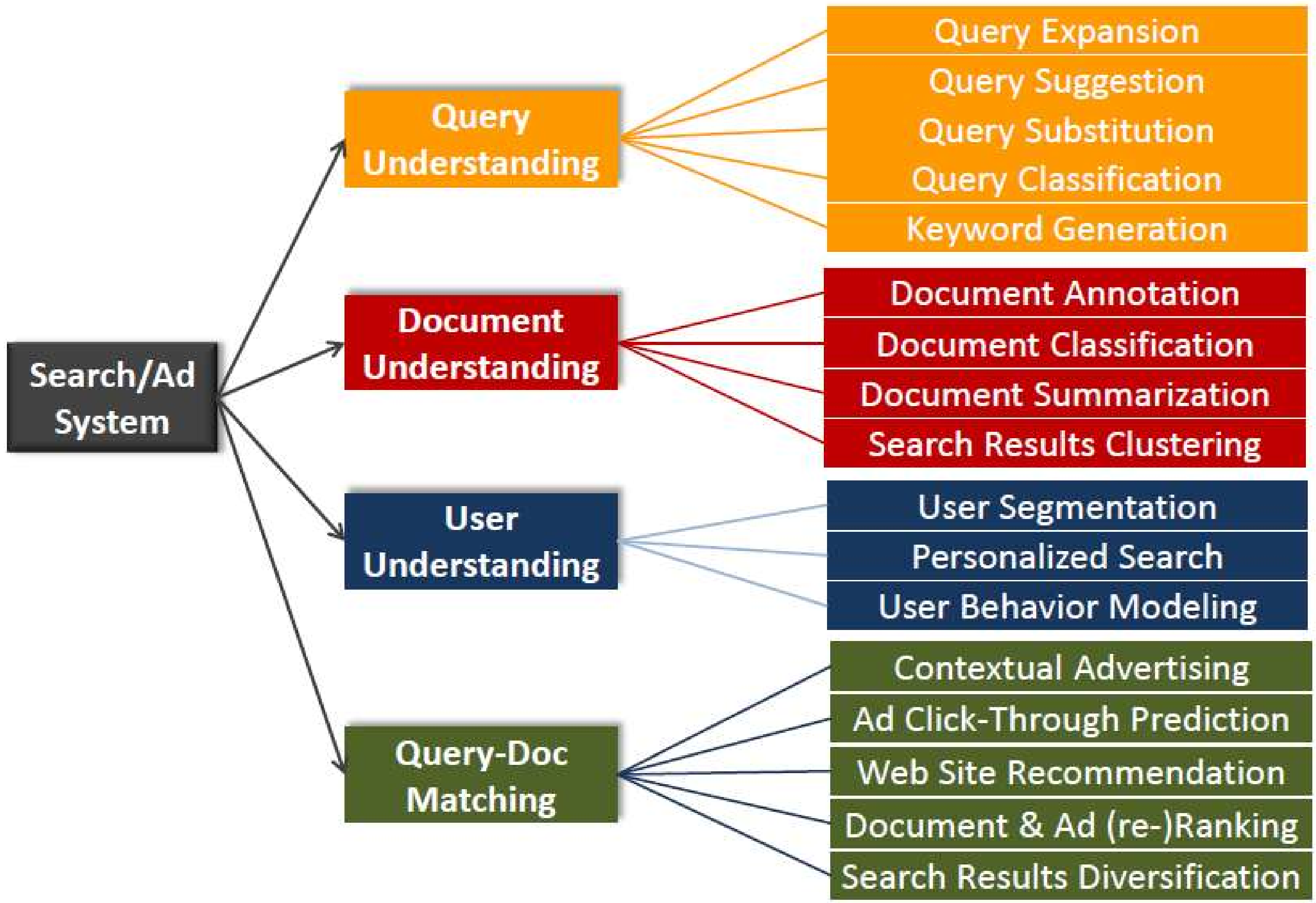
A View of Search System



Today's Agenda

- Search and browse logs
- **Log mining applications**
- Four data structures
- Query Statistics
- Query Classification

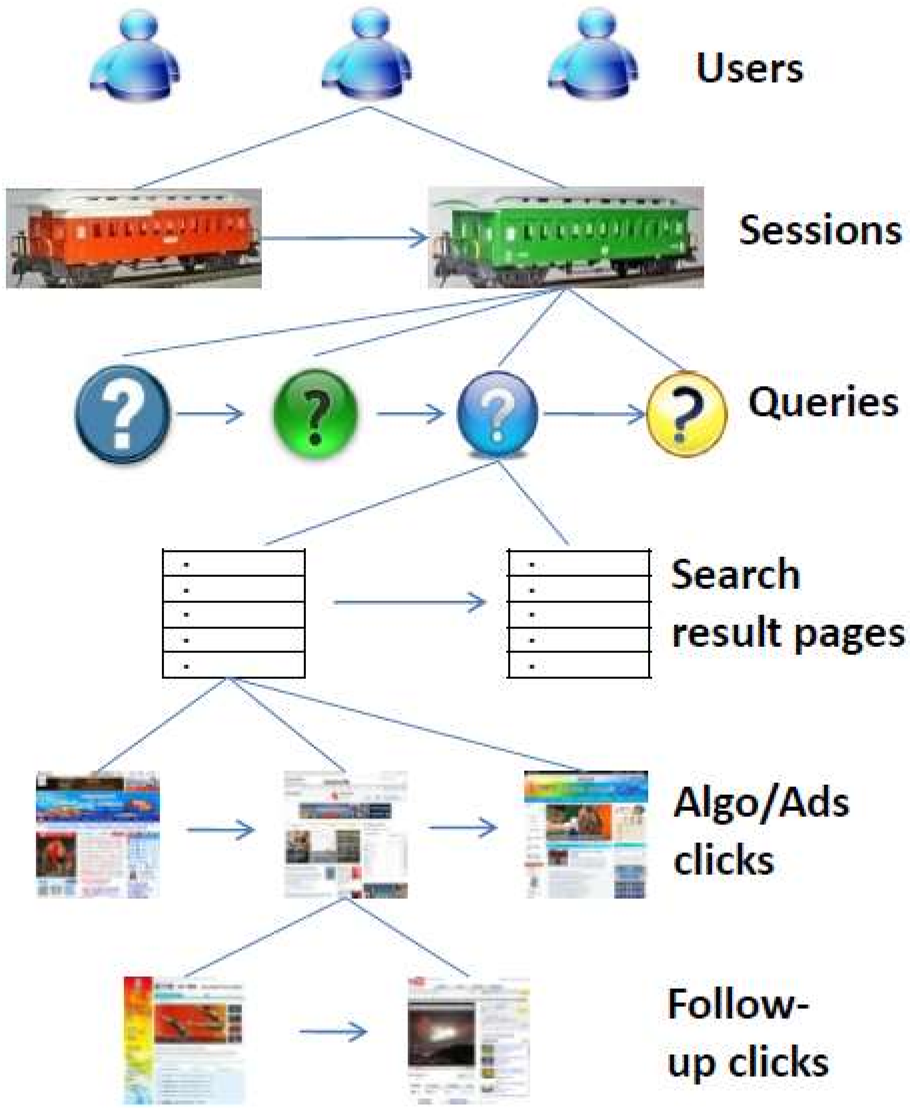
Log Mining Applications



Organizing Raw Logs by Common Data Structures

- Raw log data are stored in the format of plain text: unstructured data
- Can we summarize some common data structures from the textual logs to facilitate various log mining applications?
- Challenges: complex objects, complex applications

Complex Objects



- Various types of data objects in log data
- Complex relationship among data objects
  - Hierarchical relationship
  - Sequential relationship
- How to describe the various objects as well as their relationships?

Complex Applications

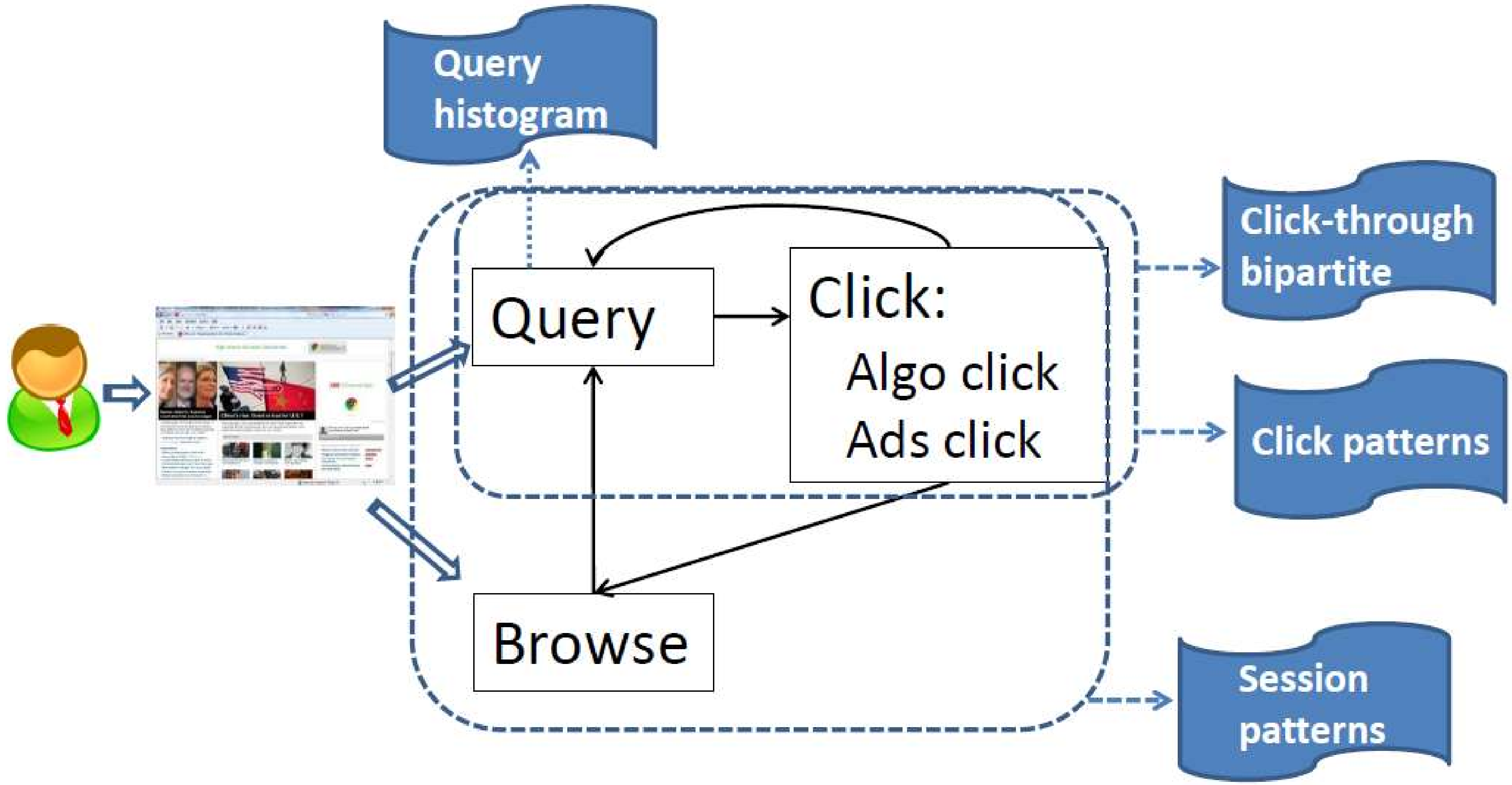
- Query understanding
  - Given a query q, what are the top-K queries following q in the same session?
- Query-Document matching
  - Given a query q, what are the top-K clicked urls?
  - Given a url u, what are the top-K queries lead to a click on u?
- Document understanding
- User understanding

- How to summarize the common data structures to support various applications?

Today's Agenda

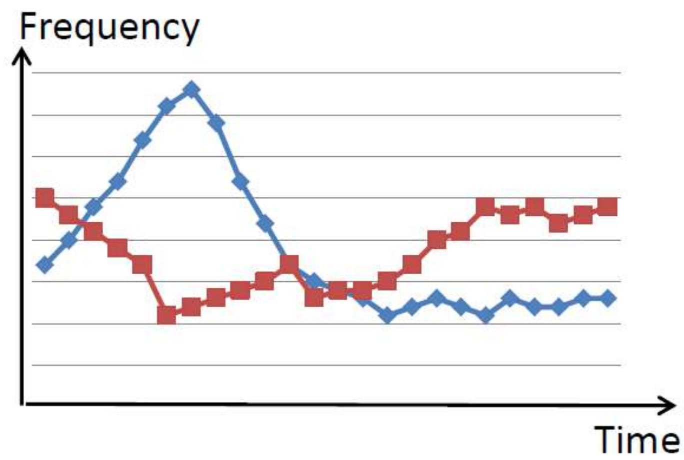
- Search and browse logs
- Log mining applications
- **Four data structures**
- Query Statistics
- Query Classification

Major Data Structures in Log Mining



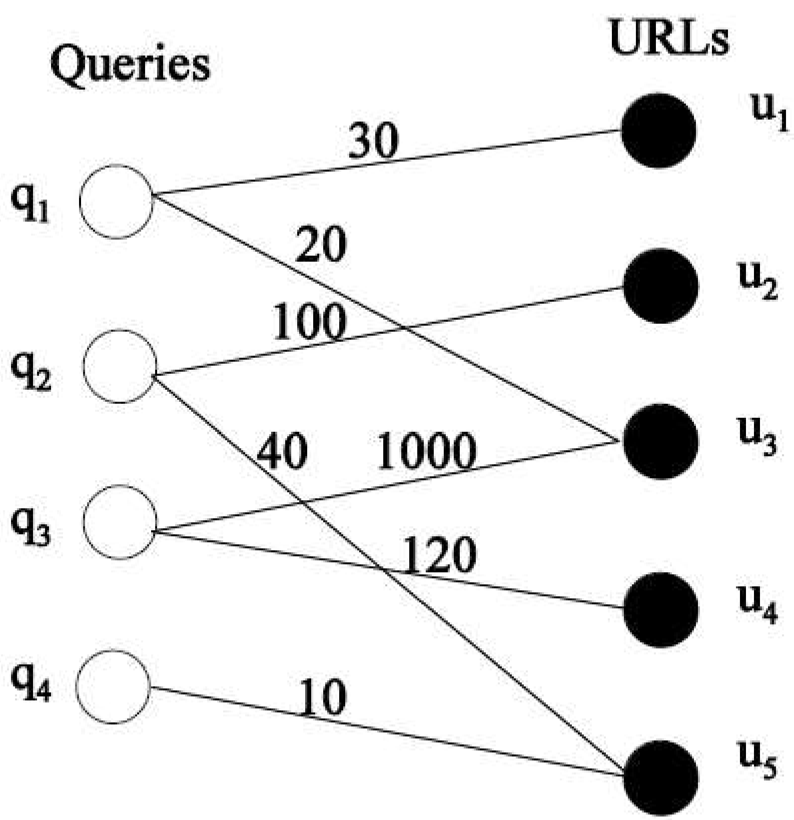
Data Structure: Query Histogram

Query String	Count
facebook	3,157 K
google	1,796 K
youtube	1,162 K
myspace	702 K
facebook com	665 K
yahoo	658 K
yahoo mail	486 K
yahoo com	486 K
ebay	486 K
facebook login	445 K



- Example applications
  - Query auto completion
  - Query suggestion: given query q, find the queries containing q
  - Semantic similarity & event detection: temporal changes of query frequency

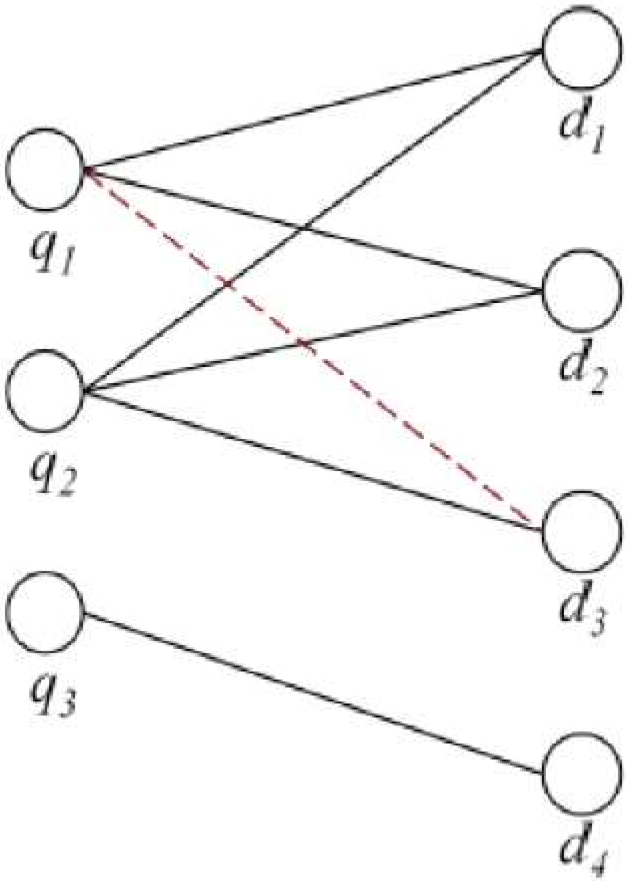
Data Structure: Click-through Bipartite



An example of click-through bipartite

- Example applications
  - Document (re-) ranking
  - Search results clustering
  - Web page summarization
  - Query suggestion: find similar queries

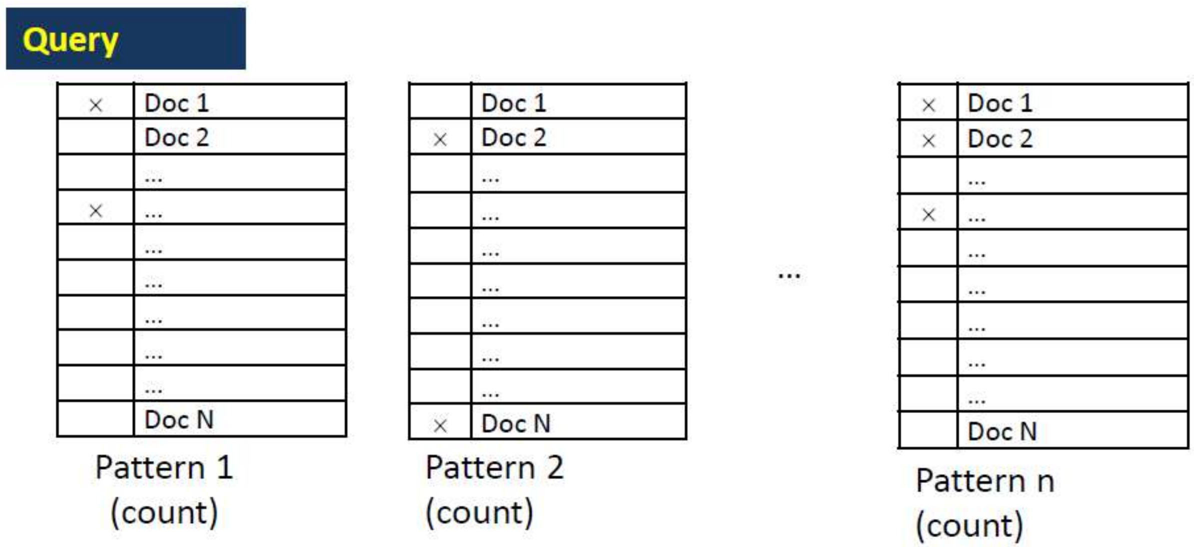
Random Walk



- Construct matrix $A_{ij} = P(d_i|q_j)$ and matrix $B_{ij} = P(q_i|d_j)$
- Random walk using the probabilities
- Before random walk, document d_3 is connected with q_2 only; after a random walk expansion, d_3 is also connected with q_1, which has similar neighbors as q_2

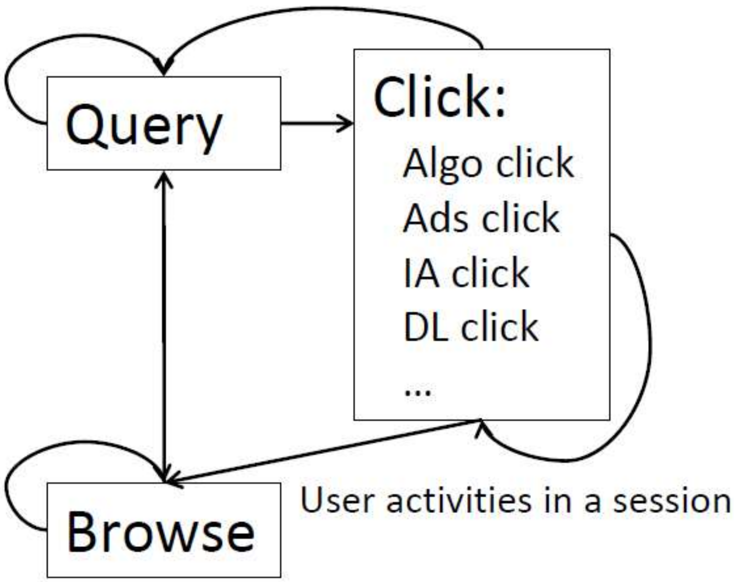
Gao, J., et al. Smoothing clickthrough data for web search ranking. SIGIR'09.

Data Structure: Click Pattern



- More information than click-through bipartite
  - Relationship between a click and its position
  - Relationship between the clicked docs with un-clicked docs
- Example applications
  - Estimate the “true” relevance of a document to a query
  - Predict users’ satisfaction
  - Classify queries (navigational/informational)

Data Structure: Session Patterns



Algo click: algorithmic click
AD click: advertisement click
IA click: instant answer click
DL click: deep link click

- Sequential patterns
  - E.g., behavioral sequences
    - SqLrZ [Fox05]
    - S: session starts; Q: query L: receives a search result page R: click; Z: session ends
- Example applications
  - Doc (re-)ranking
  - Query suggestion
  - Site recommendation
  - User satisfaction prediction

Fox et al. Evaluating implicit measures to improve web search. TOIS, 2005.

Session Segmentation

- Problem: given a sequence of user queries, where to cut the session boundary
- Features for session segmentation
  - Timeout threshold (e.g., [Silverstein99])
- 30 minutes timeout is often used
  - Common words or edit distance between queries (e.g., [He02])
  - Adjacency of two queries in user input sequences (e.g., [Jones08])
  - Similarity between the top K search results of two queries (e.g., [Radlinski05])
- Tradeoff between cost and accuracy

Today's Agenda

- Search and browse logs
- Log mining applications
- Four data structures
- **Query Statistics**
- Query Classification

Data Sets

Region	Data	Engine	Date	# of queries	# of sessions	Reference
US	Excite97	Excite	16 Sept, 1997	51 K	18K[1]	Jansen00, Jansen01, Spink02, Jansen06
	Excite99	Excite	1 Dec 1999	1M	326 K	Wolfram01, Spink02, Jansen06
	Excite01	Excite	30 Apr 2001	1M	262K	Spink02, Spink02a, Jansen06
	AV98	AltaVista	2 Aug- 13 Sept, 1998	575 M	285 M	Silverstein99, Jansen01, Jansen06
	AV02	AltaVista	8 Sept, 2002	1 M	369 K	Jansen04, Jansen06
Europe	FB98	Fireball	1-31 Jul. 1998	16 M	-	Holscher00, Jansen01, Jansen06
	BWIE00	BWIE	3-18 May, 2000	72K	83K	Cacheda01a, Cacheda01b, Jansen06
	FAST01	FAST	6 Feb, 2001	452 K	153K	Spink02a
	ATW01	AlltheWeb	6 Feb 2001	452K	153K	Jansen05, Jansen06
	ATW02	AlltheWeb	28 May 2002	957K	345K	Jansen05, Jansen06

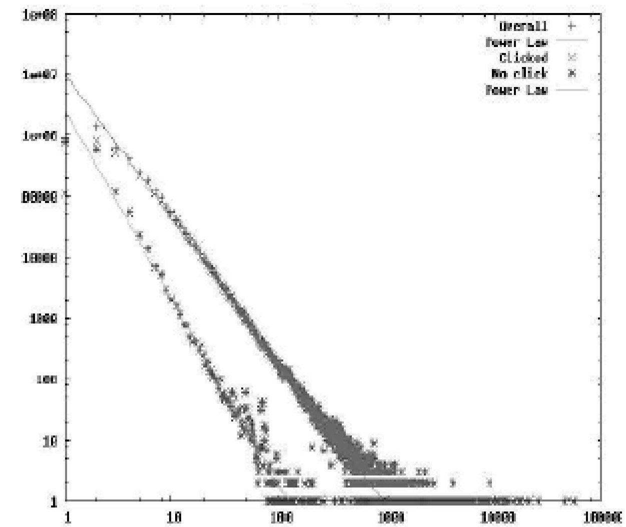
Query Length

Region	Data	Avg	1	2	3	>3	Reference
US	Excite97	2.21	31%	31%	18%	15%	Jansen00
	Excite99	2.4	29.8%	33.8%	36.4%		Wolfram01
	Excite01	2.6	26.9%	30.5%	42.6%		Spink02
	AV98	2.35	25.8%	26.0%	15.0%	12.6%	Silverstein99
	AV02	-	20%	-	-	-	Jasen06
Europe	FB98	1.66	54.6%	30.8%	10.4%	4%	Jansen01
	FAST01	2.3	25%	36%	39%		Spink02a
	ATW01	2.4	25.1%	35.8%	22.4%	15.9%	Jansen05
	ATW02	2.3	33.1%	32.6%	18.9%	15.1%	

- Average length: 1.66~2.6 words
- Much shorter than in traditional IR (6-9)
- Average length remains constant over time and across regions

Query & Term Frequencies

Region	Data	Head	Tail	Reference
US	Excite97 (term)	0.34% unique terms (occurrence >100) account for 18.2% traffic	44.8% unique terms (occurrence=1) account for 8.6 traffic	Jansen00
	Excite99 (term)	Top 100 terms account for 19.3% traffic	61.6% unique terms occurs only once	Wolfram01, Spink02
	Excite01	Top 100 terms account for 22.0% traffic	61.7% unique terms occurs only once	Spink02
	AV98	Top 25 queries account for 1.5% traffic	63.7% unique queries occur only once	Silverstein99
Europe	BWIE00	-	23.4% unique queries only occur once	Cacheda01a
	FAST01	Top 100 terms account for 14% traffic	-	Spink02a
	ATW01	Top 100 terms account for 15% traffic	7% unique queries only occur once	Jansen05
	ATW02	Top 100 terms account for 14% traffic	10% unique queries only occur once	



- Head and tail parts are highly skewed
  - Head: few queries/terms account for large traffic
  - Long tail: consists of large percentage of unique queries/terms
- Middle region follows Zipf distribution (the distribution of words in long English texts)

Number of Viewed Search Result Pages

	Data	Avg	1	2	3	>3	Source
US	Excite97	1.7	58%	19%	9%	_	Jansen00 Wolfram01
	Excite99	1.6	42.7%	21.2%	36.1%		Wolfram01
	Excite01	1.7	50.5%	20.3%	29.2%		Spink02
	AV98	1.39	85.2%	7.5%	3.0%	4.3%	Silversten99
	AV02	<2	73%	_	_	_	Jasen06,
Europe	FAST01	2.2	_	_	_	_	Spink02a
	FB98	<2	59.5%	-	-	-	Jansen01,
	BWIE00	<2	67.9%	13.2%	6.0%	_	Cacheda01a
	ATW01	<2	83.5%	9.6%	3.0%	_	Jansen05
	ATW02	<2	76.3%	13.1%	3.9%	_	

- On average, users view less than two search result pages
- Over half of users do not access result beyond first page
- Relevance of top 10 search results is critical

Session Length

Region	Data	Avg	1	2	3	>3	Source
US	Excite97	1.6	67%	19%	7%	7%	Jansen00 Jansen01,
	Excite99	1.7	60.4%	19.8%	19.8%		Wolframe01
	Excite01	2.3	55.4%	19.3%	25.3%		Spink02 Spink02a
	AV98	2.02	77.6%	13.5%	4.4%	4.5%	Silverstein99
	AV02	~2	47%	-	-	-	Jansen06
Europe	FAST01	2.9	53%	18.9%	29%		Spink02a
	ATW01	3.0	52.9%	18.3%	9.4%	19.4%	Jansen05
	ATW02	2.8	58.7%	16.1%	7.9%	17.3	

- Average session length is around 2-3 queries
- More than half of sessions consist of only one query
- Europe sessions are longer than US sessions

Topic Distribution

Name	People & Place	Commerce	Health	Entertainment	Internet & Computer	Porn	Source
Excite97	6.7% (6)	13.3% (3)	9.5% (5)	19.9% (1)	12.5% (4)	16.8% (2)	Wolfram01
Excite99	20.3% (2)	24.4% (1)	7.8% (4)	7.5% (6)	10.9% (3)	7.5% (5)	Wolfram01
Excite01	19.7% (2)	24.7% (1)	7.5% (6)	6.6% (7)	9.6% (4)	8.5% (5)	Spink02,
AV02	49.3% (1)	12.5% (2)	7.5% (4)	4.6% (6)	12.4% (3)	3.3% (7)	Jasen06
FAST01	22.5% (1)	12.3% (3)	7.8% (6)	9.1% (5)	21.8% (2)	10.8% (4)	Spink02a
ATW01	22.5% (1)	12.3% (3)	7.8% (6)	9.1% (5)	21.8% (2)	10.8% (4)	Jansen05
ATW02	41.5% (1)	12.7% (3)	4.9% (5)	9.5% (4)	16.3% (2)	4.5% (6)	

- Top six topics are same over time and across regions
- Percentage of individual topics change over time
  - In US, percentage of porn searches decreases, while percentages of commerce and people & place increase

Summary of Query Statistics

- Web search is quite different from traditional IR

	Traditional IR	Web search
Query length	6-9	2-3
Query frequency	Zipf distribution	Zipf distribution + skewed head and tail
Num. of viewed result page	~10	1-2
Session length	7-16	1-2
Topics	More focused	Diverse

Today's Agenda

- Search and browse logs
- Log mining applications
- Four data structures
- Query Statistics
- **Query Classification**

Query Classification Tasks

- Queries can be categorized on multiple dimensions
  - Task (navigational, informational)
  - Topics (ODP categories, auto-created concepts)
  - Entity and Attribute (e.g., ‘avatar game’)
  - Time-sensitiveness (e.g., ‘WWW conference’)
  - Location-sensitiveness (e.g., ‘pizza’)
  - Data Source (e.g., wiki, image, video)
- Intent of query can be represented by the categories
- Applications of query classification
  - Relevance ranking
  - Faceted search or categorized search
  - Online advertisement

Challenges in Query Classification

- Queries are
  - Usually very short
  - Often ambiguous
  - Meaning changes over time and location

Search Tasks

- High level task categories [Broder02]
  - Navigational: to reach particular site
  - Informational: to acquire some information assumed to be present on one or more web pages.
  - Transactional: to perform some web-mediated activity.
- Distribution
  - Varies according to different studies
  - Navigational: 20%, Informational: 48%, Transactional: 30%

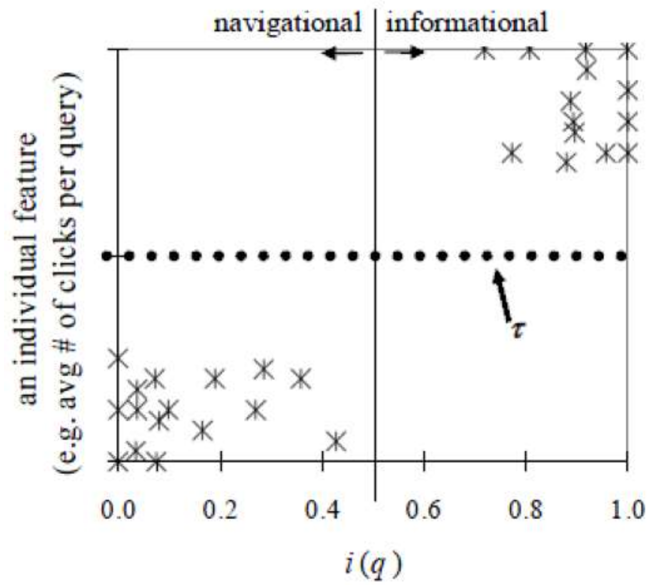
Methods for Query Task Classification

- Using web pages
  - [Kang03]
- Using click-through data and anchor text data
  - [Lee05]

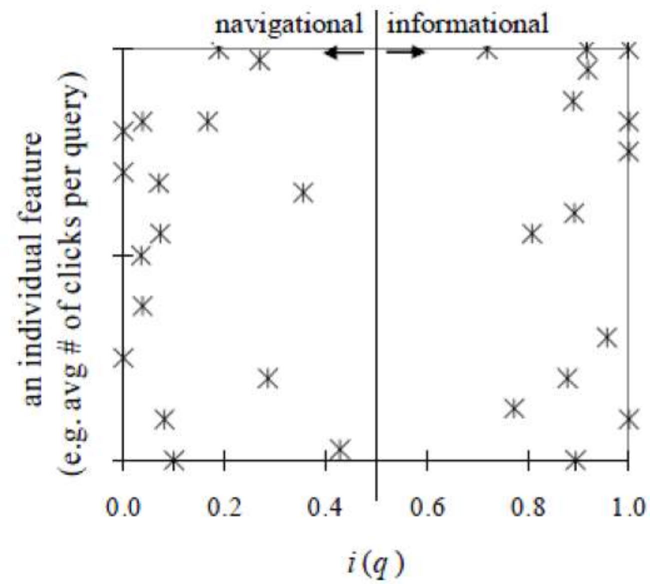
Query Task Classification Using Click-through and Anchor Text Data [Lee05]

- Only two categories considered, i.e., navigational and informational
- Basic idea
  - Navigational query ⇔ click distribution is skewed
  - Navigational query ⇔ anchor text distribution is skewed
- Method
  - Using mean, median, skewness, and kurtosis to characterize distributions of clicks and anchor texts
  - Linear combination of features
- Accuracy: 90%
- Challenge: difficult for tail queries

Result of Single Feature



An Effective Feature



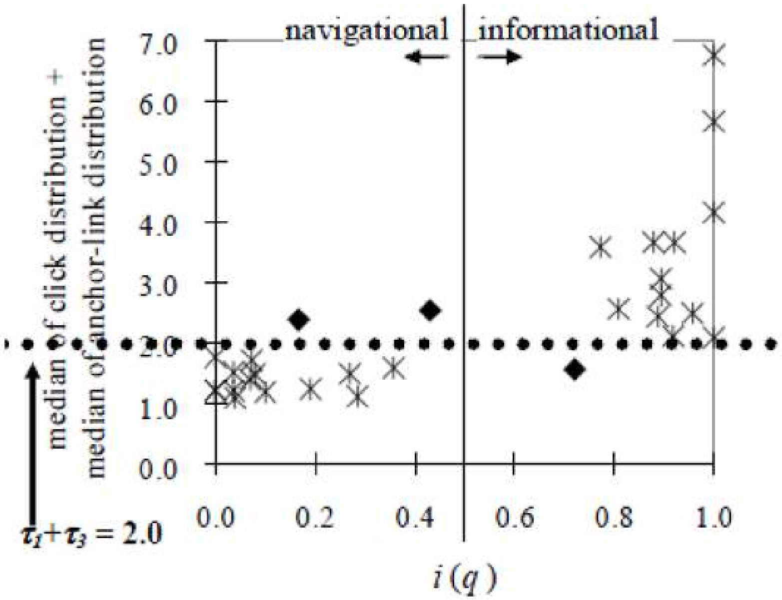
An Ineffective Feature

- 50 head queries labeled by 28 graduate students
- Each point represents query
- i(q) is percentage of informational labels of query q
- A feature is effective if we can set horizontal bar (i.e., threshold) to separate navigational queries from informational queries

Result of Linear Combination

- Linear combination
  - $f = w_1 . f_1 + w_2 . f_2 + \cdots + w_n . f_n$
- A simple combination shows a better accuracy
- Combines two features
- Equal weights
- Accuracy reaches 90%

f= (median of click distribution)
+(median of anchor distribution)



Search Topics

- ODP categories
- Automatically constructed concepts (clusters)
- Query can have multiple topics (is ambiguous)
  - e.g., ‘Jaguar’ [car][animal]

Methods for Query Topic Classification

- Directly applying text classification techniques
- Using search results of query [Shen05]
- Using search log data
  - Using query log data [Beitzel07]
  - Using click-through data [Fuxman07], [Li08]

Query Topic Classification Using Query Log Data [Beitzel07]

- Four methods of classification
  - Exact-match lookup
  - N-gram lookup
  - Perceptron
  - Selectional Preference
- Combination of four methods
  - exact-match lookup first, followed by the perceptron, 4-gram lookup, and selectional preferences
- Accuracy: F1 score = 0.25

Selectional Preference: Step 1

- View query as pair of lexical units
  - <head, tail>
  - Queries with *n* terms form *n-1* pairs
  - Example: “directions to DIMACS” forms two pairs
    - <directions, to DIMACS> and <directions to, DIMACS>
  - Only applicable to queries of 2+ terms

Selectional Preference: Step 2

- Manually label some words with categories
- Check head and tail of each pair to see if they appear in manually labeled set
- Convert each <head, tail> pair into:
  - <head, CATEGORY> (*forward* preference)
  - <CATEGORY, tail> (*backward* preference)

Selectional Preference: Step 2

Query Log

Yahoo mail
interest rates
EBay motor

→

<Finance> Category

banks
bonds
stocks
rates
savings
loan

→

Forward Preference:
Head Tail

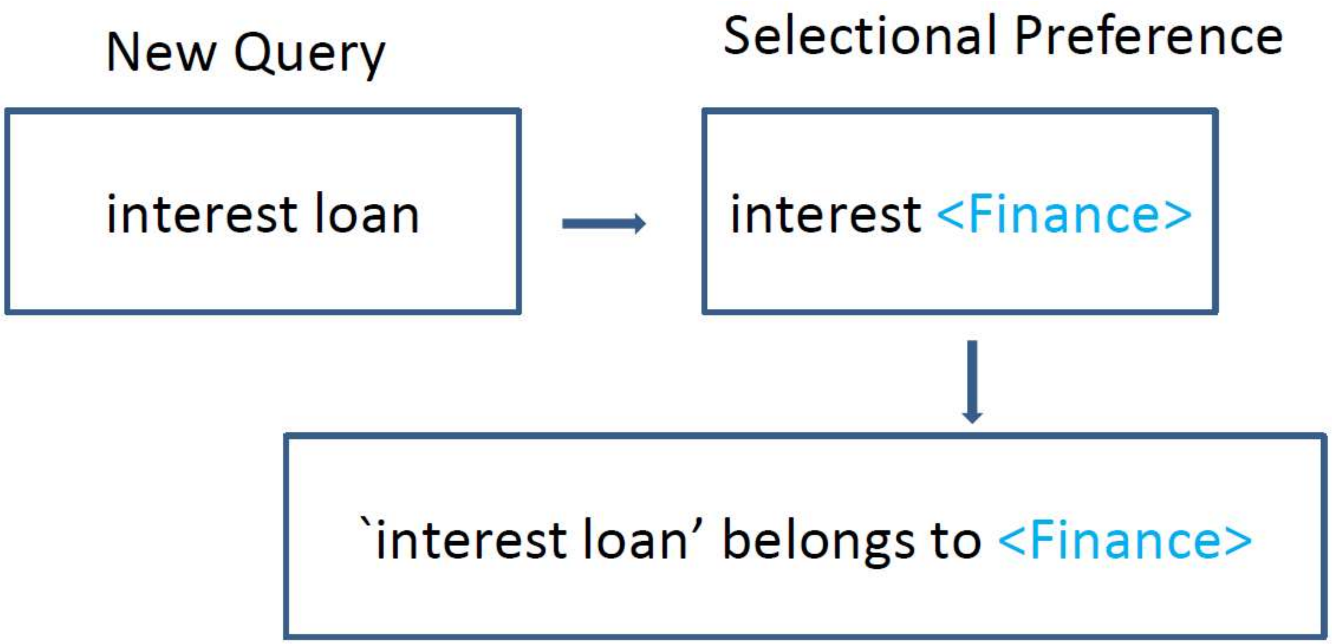
interest <Finance>

Selectional Preference: Step 3

- Score each preference using Resnik's formula
- $S(u|x) = \max_{u} P(u|x) \log \frac{P(u|x)}{P(u)}$
  - *x* denotes lexical unit and *u* denotes category of the other lexical unit

Selectional Preference: Step 4

- Use mined selectional preferences to assign categories to unseen queries



Query Topic Classification Using Click-through Data [Fuxman07]

- View click-through bipartite as undirected graph
- Define random walk model
- Probability on edge represents transition probability (calculated using click-through counts)
- Probability of node represents probability of belonging to class
- Propagate class labels on graph

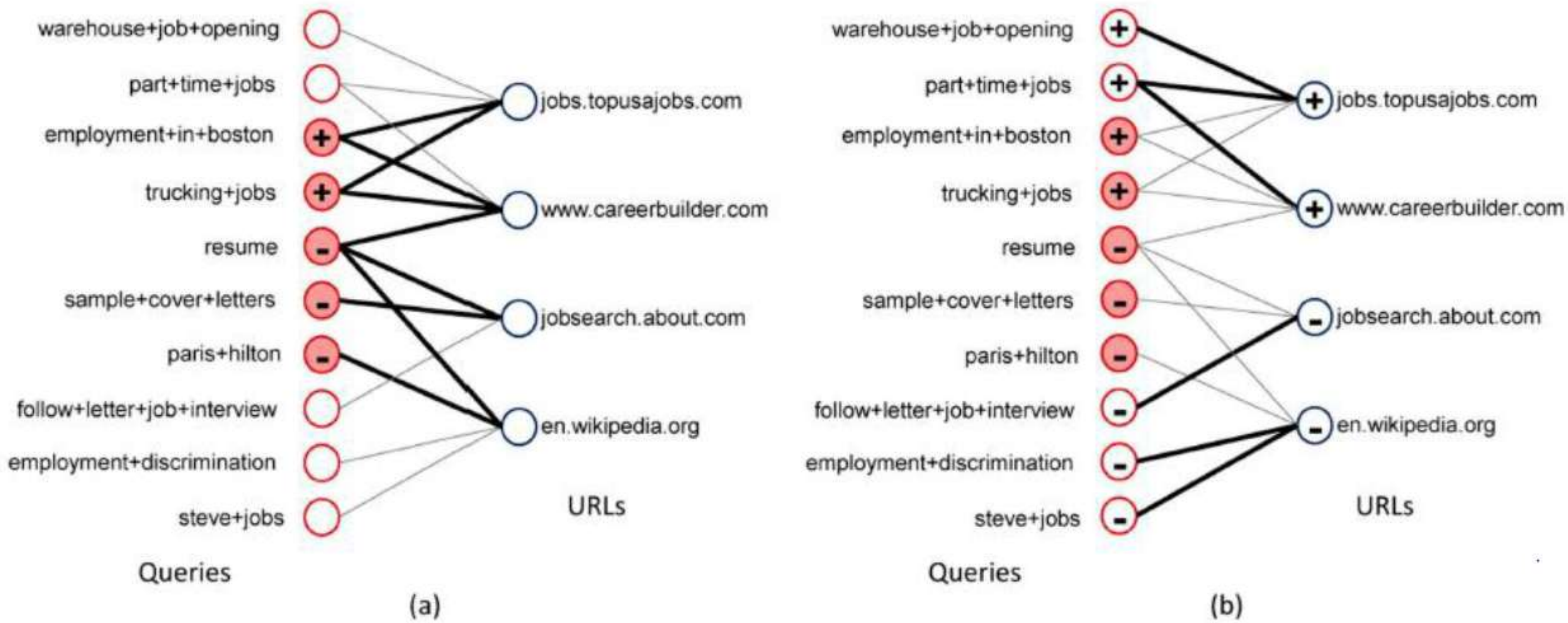
Random Walk Algorithm

- Add ‘null’ node to the click through bipartite
  - Each node may walk to null node with probability
- Iteration between two processes
  - Estimate probability of query node
    - $P\left(l_q = c\right) = (1-\alpha)\sum_{(q,u)} P(q \rightarrow u) P(l_u = c)$
  - Estimate probability of URL node
    - $P(l_u = c) = (1-\alpha)\sum_{(q,u)} P(u \rightarrow q) P(l_q = c)$
- It is guaranteed to converge
- Analogy to electrical network.

Query Topic Classification Using Click-through Data [Li08]

- Given a set of labeled queries (representing same topic)
- Train classifier based on content of queries
- Propagate class labels through click-though bipartite
- Iteratively combining content-based classification and click-based classification
- Accuracy: F score = 0.74 to 0.88

Propagation through Click-Through Bipartite



Labeled seeds

After propagation

Content-based Classifier

- Maximum Entropy Classifier
  - $P_\lambda(y|x) = \frac{\exp(\sum_i \lambda_i \phi_i(x,y))}{\sum_y \exp(\sum_i \lambda_i \phi_i(x,y))}$
- *x* denotes query, *y* denotes query topic class, $\phi(x, y)$ denotes feature, λ denotes parameter
- Using n-grams of query or snippets of query as features

Click-based Classifier

- Let *W* be $m \times n$ matrix where $W[i, j]$ is click count on URL *j* for query *i*
- Let *F* be $m \times 2$ matrix where $W[i, j]$ is non-negative, real number indicate likelihood that query *i* belongs to class *y*
- Random walk converges to
  - $F^* = (1 - \alpha)(1 - \alpha A)^{-1} F^0$
  - Where $A = D^{\frac{1}{2}} W W^T D^{-\frac{1}{2}}$, *D* is diagonal matrix in which element $d_{i,i}$ equals sum of elements in row i of WW^T

Combining Classifiers

- Step 1: initialize F* by labeled seeds, initialize λ as random
- Step 2: repeat
  - Train λ^* of content-based classifier using classification results by current F*
  - Train F* of click-based classifier use classification results by current λ^*
- until convergence

Summary of Query Classification

- Classify queries based on tasks
  - Using click distribution and anchor text distribution
- Classify queries based on topics
  - Using query log, exact match, selectional preference, etc
  - Using click-through data and random walk

Take-away Messages

- Search & browse logs
- Log mining applications
  - Query understanding, document understanding, user understanding, query-document matching,
- Four data structures
  - Query histogram, click-through bipartite, click patterns, session patterns
- We discussed a few query statistics
- Query Classification
  - Classify queries based on tasks
  - Classify queries based on topics

Further Reading

- Daxin Jiang, Jian Pei, Hang Li. Web Search/Browse Log Mining: Challenges, Methods, and Applications. Tutorial at WWW 2010
- Daxin Jiang, Jian Pei, Hang Li. Mining Search and Browse Logs for Web Search: A Survey. ACM Transactions on Computational Logic, Vol. V, No. N, February 2013, Pages 1–42.
- Maristella Agosti, Franco Crivellari, Giorgio Maria Di Nunzio. Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. Data Min Knowl Disc (2012) 24:663–696
- Fabrizio Silvestri. Mining Query Logs: Turning Search Usage Data into Knowledge. Foundations and Trends in Information Retrieval. Vol. 4, Nos. 1–2 (2010) 1–174
- Marius Pasca. Tutorial. Web Search Queries as a Corpus. ACL 2011
- Ricardo Baeza-Yates, Fabrizio Silvestri. Query Log Mining.
- Qiaozhu Mei, Kenneth Church. Entropy of Search Logs. WSDM 2008.

Preview of Lecture 22: Query Understanding by Log Mining

- Query Expansion, Refinement, and Suggestion
- Temporal and Spatial Aspects of Queries
- Text Mining from Query Logs

Disclaimers

- This course represents opinions of the instructor only. It does not reflect views of Microsoft or any other entity (except of authors from whom the slides have been borrowed).
- Algorithms, techniques, features, etc mentioned here might or might not be in use by Microsoft or any other company.
- Lot of material covered in this course is borrowed from slides across many universities and conference tutorials. These are gratefully acknowledged.

Thanks!