Web Search and Browse Log Mining: Challenges, Methods, and Applications

Daxin Jiang (姜大昕)

Lead Researcher, WSM, MSRA

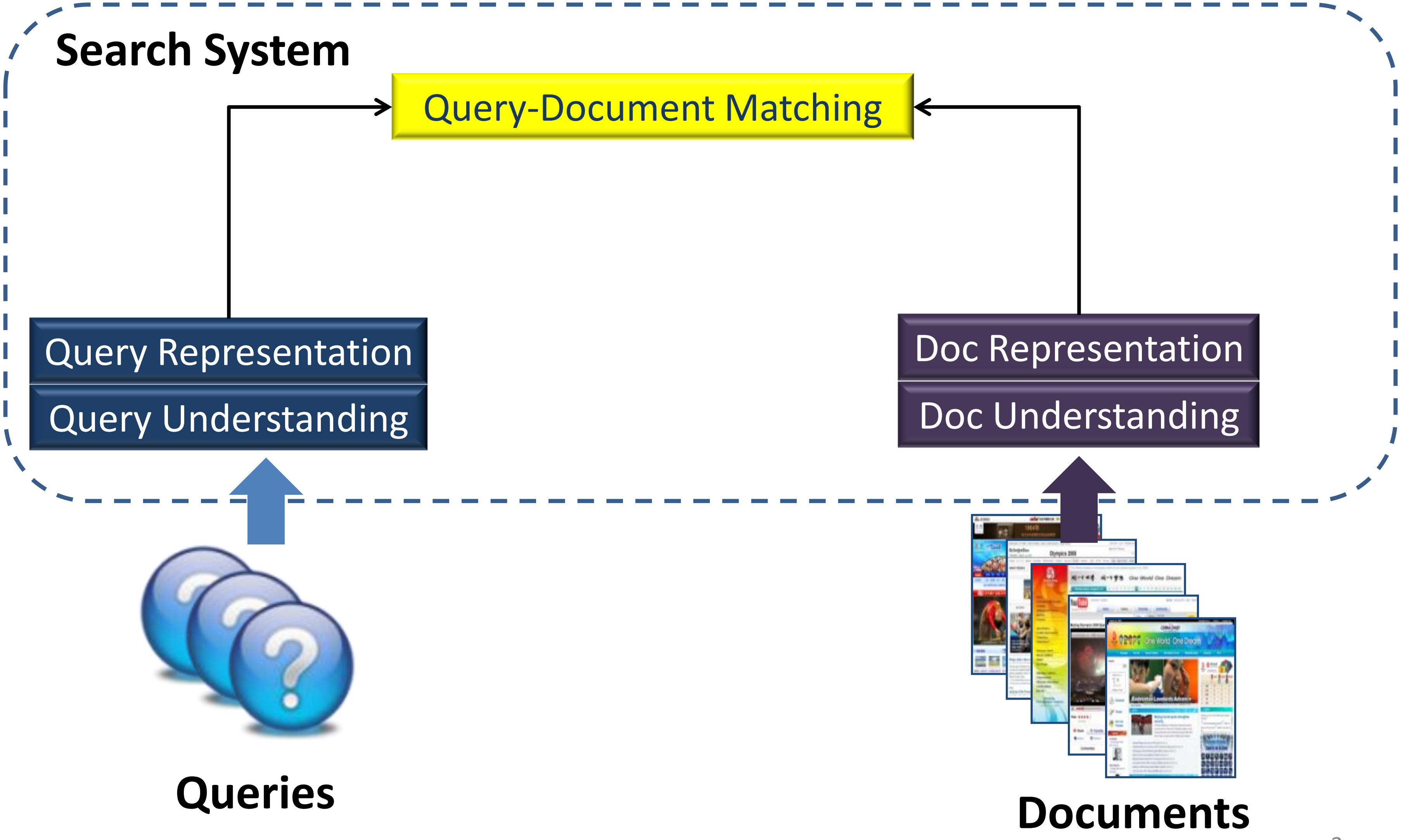
Email: djiang@microsoft.com

http://research.microsoft.com/en-us/people/djiang/

About this Lecture

- Based on a tutorial presented at WWW’10, SIGIR’10, KDD’10, SIGIR’11
- Co-authored with Dr. Hang Li and Prof. Jian Pei
- This lecture only covers the first half
  - The second half will be your homework ☺
- The full slide deck can be obtained from my homepage http://research.microsoft.com/en-us/people/djiang/

Traditional Search System



Search System

Query-Document Matching

Query Representation

Query Understanding

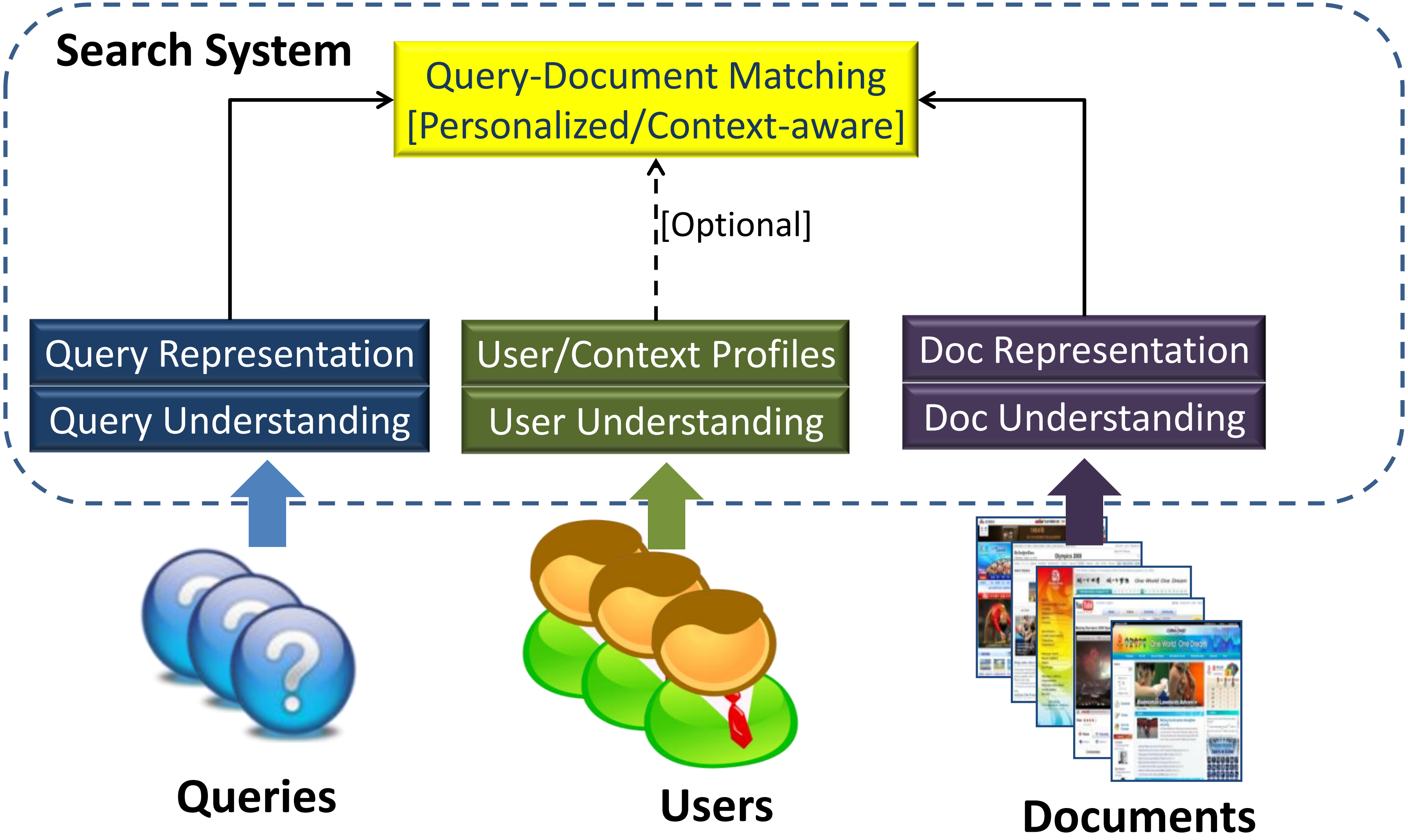
Doc Representation

Doc Understanding

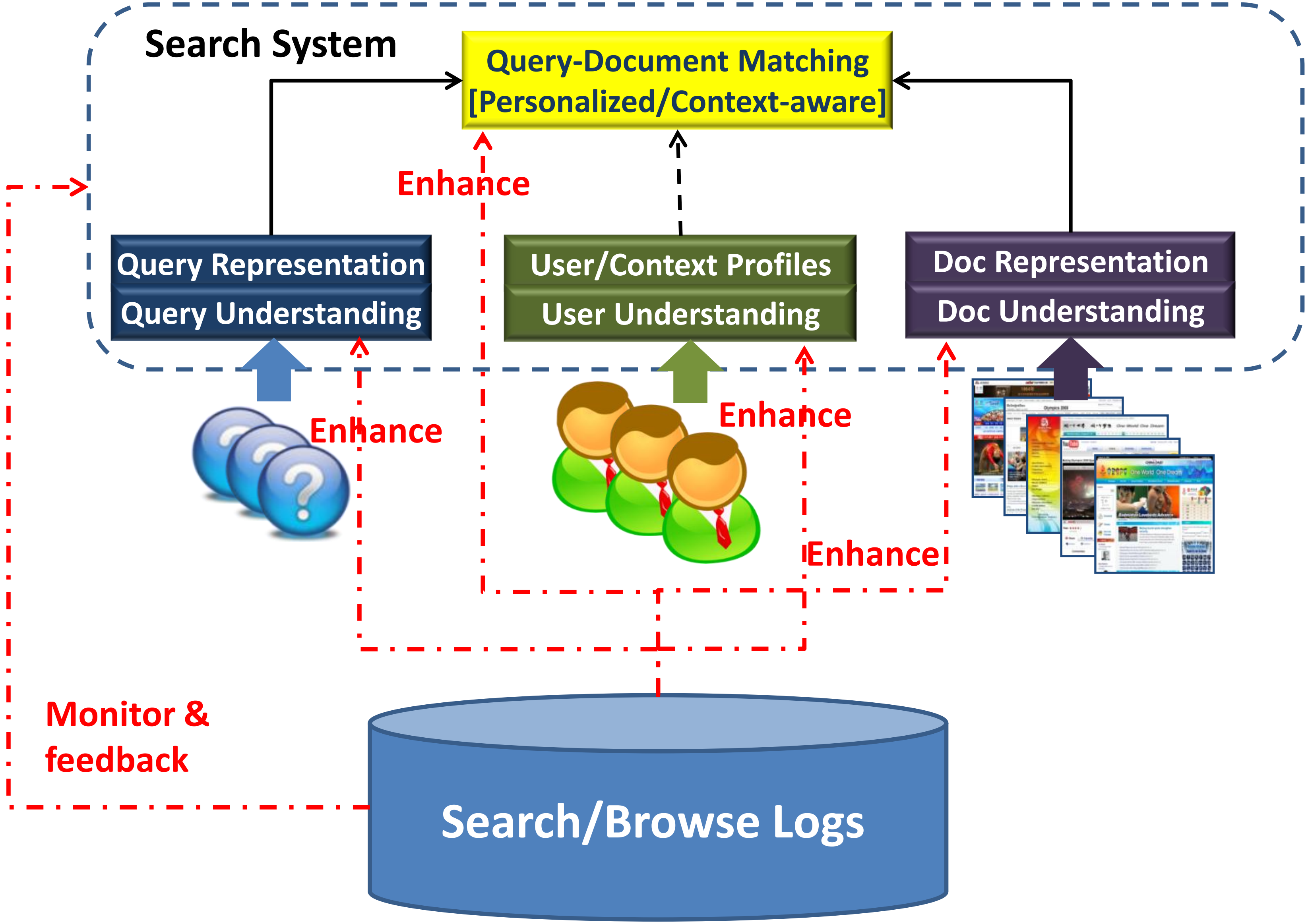
Queries

Documents

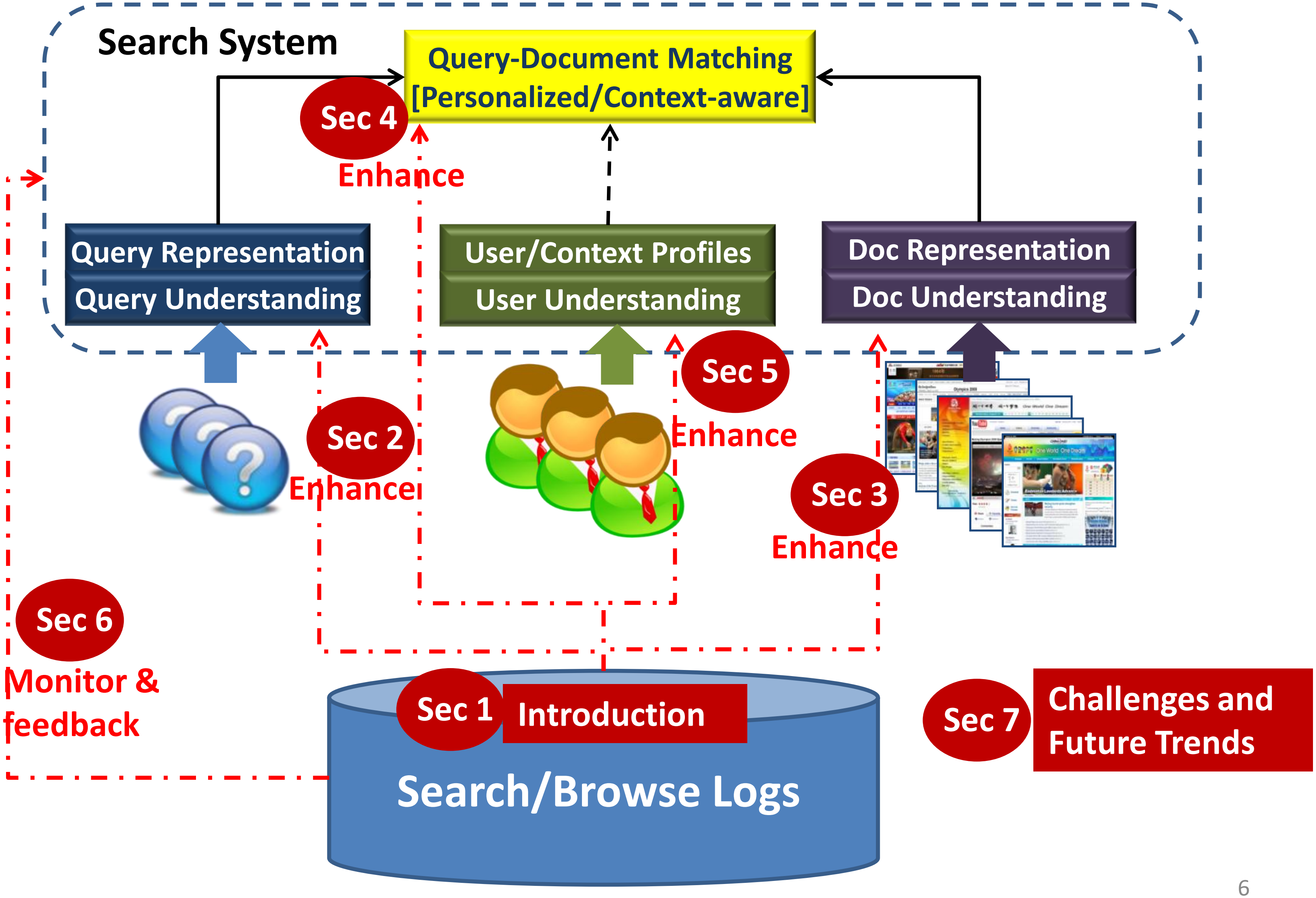
Personalized/Context-Aware Search System



Search and Browse Log Mining



Road Map



Search System

Query-Document Matching [Personalized/Context-aware]

Query Representation

Query Understanding

User/Context Profiles

User Understanding

Doc Representation

Doc Understanding

Sec 1 Introduction

Search/Browse Logs

Sec 2 Enhance

Sec 3 Enhance

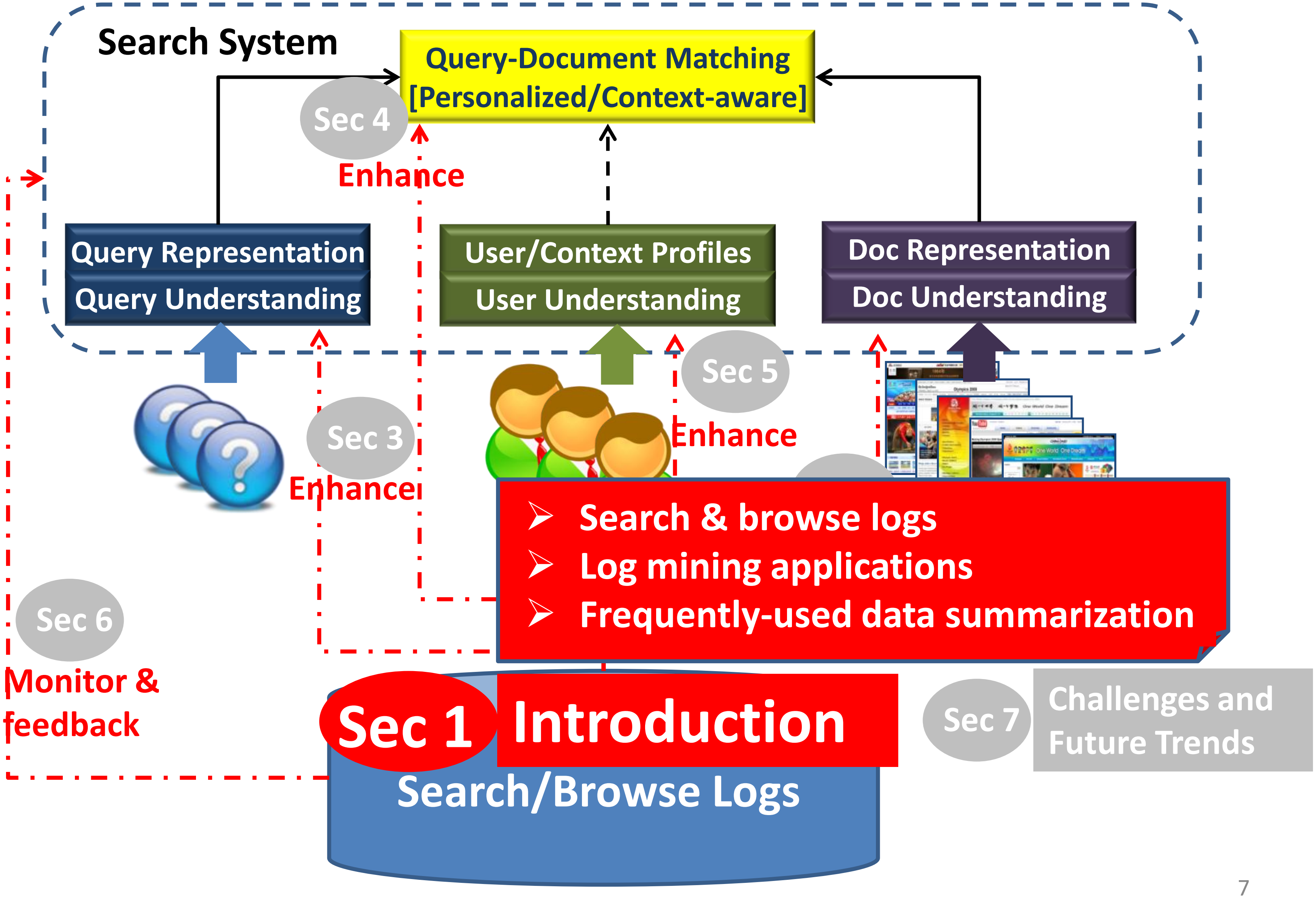
Sec 4 Enhance

Sec 5 Enhance

Sec 6 Monitor & feedback

Sec 7 Challenges and Future Trends

Road Map



Search System

Query-Document Matching [Personalized/Context-aware]

Query Representation

Query Understanding

User/Context Profiles

User Understanding

Doc Representation

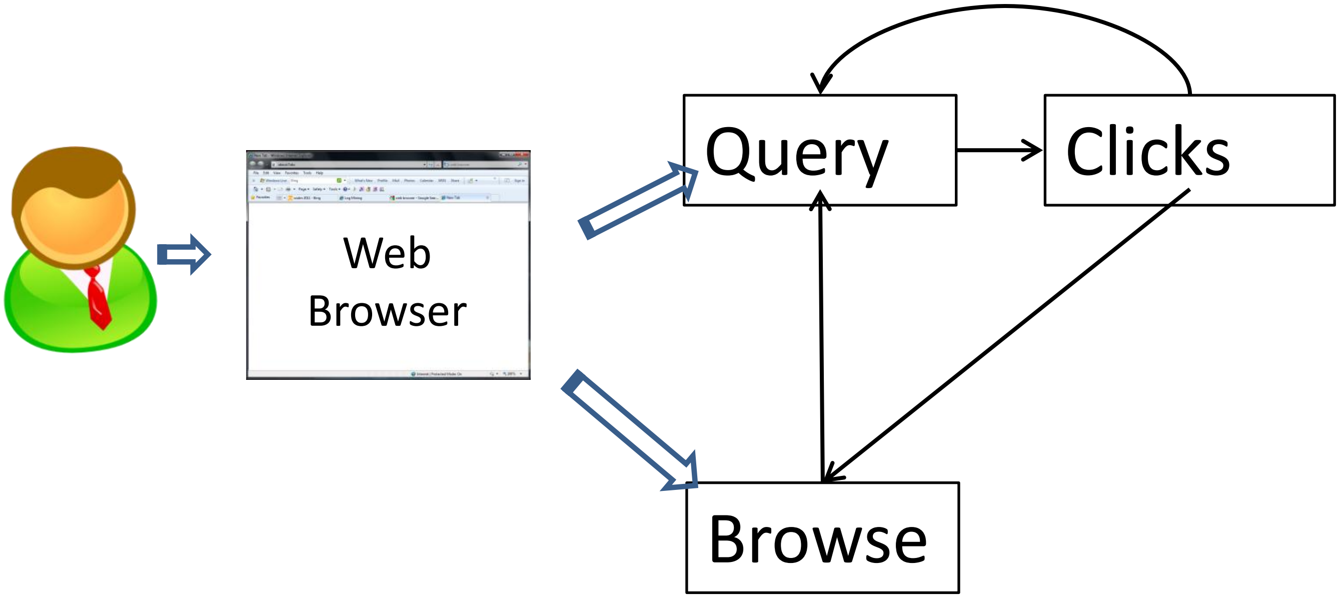
Doc Understanding

- Search & browse logs
- Log mining applications
- Frequently-used data summarization

Sec 1 Introduction

Search/Browse Logs

User Behavior in Web Browser



Web Browser

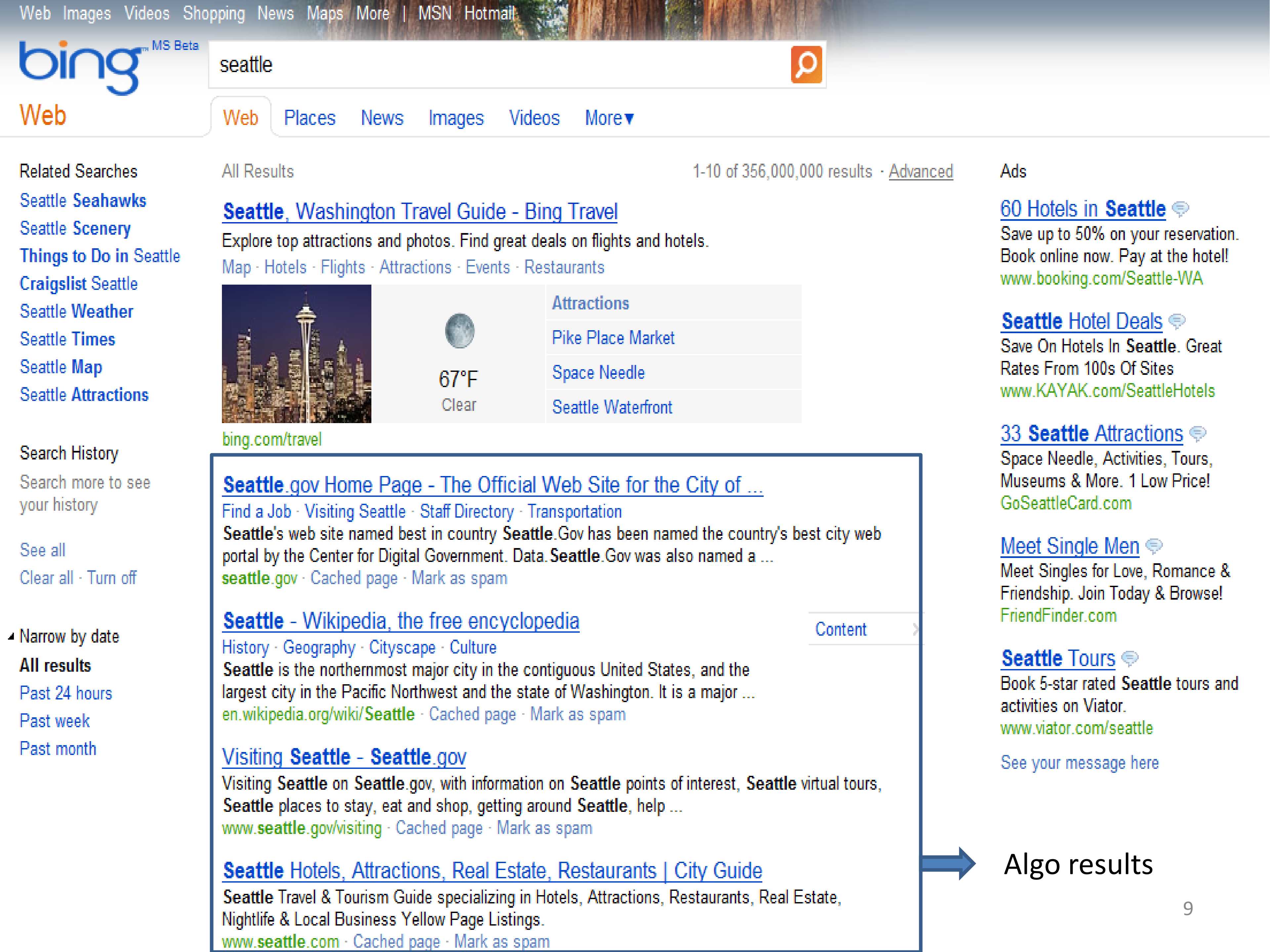
Query

Clicks

Browse

Search Results

- Algorithmic results; or “algo results” for short
  - Sometimes referred to as the “ten blue links”
- Advertisement results; or “ad results” for short
  - Sometimes referred to as the “sponsor links”



Web · Images · Videos · Shopping · News · Maps · More | MSN · Hotmail
bing MS Beta
seattle
Web
Web · Places · News · Images · Videos · More▾

Related Searches

- Seattle Seahawks
- Seattle Scenery
- Things to Do in Seattle
- Craigslist Seattle
- Seattle Weather
- Seattle Times
- Seattle Map
- Seattle Attractions

Search History

Search more to see your history

See all

Clear all · Turn off

Narrow by date

All results

- Past 24 hours
- Past week
- Past month

All Results 1-10 of 356,000,000 results · Advanced

Seattle, Washington Travel Guide - Bing Travel

Explore top attractions and photos. Find great deals on flights and hotels.

Map · Hotels · Flights · Attractions · Events · Restaurants

67°F
Clear

Attractions
- Pike Place Market
- Space Needle
- Seattle Waterfront

bing.com/travel

Seattle.gov Home Page - The Official Web Site for the City of ...

Find a Job · Visiting Seattle · Staff Directory · Transportation

Seattle's web site named best in country **Seattle**.Gov has been named the country's best city web portal by the Center for Digital Government. Data.**Seattle**.Gov was also named a ...

seattle.gov · Cached page · Mark as spam

Seattle - Wikipedia, the free encyclopedia

Content ›

History · Geography · Cityscape · Culture

Seattle is the northernmost major city in the contiguous United States, and the largest city in the Pacific Northwest and the state of Washington. It is a major ...

en.wikipedia.org/wiki/**Seattle** · Cached page · Mark as spam

Visiting **Seattle** - **Seattle**.gov

Visiting **Seattle** on **Seattle**.gov, with information on **Seattle** points of interest, **Seattle** virtual tours, **Seattle** places to stay, eat and shop, getting around **Seattle**, help ...

www.**seattle**.gov/visiting · Cached page · Mark as spam

Seattle Hotels, Attractions, Real Estate, Restaurants | City Guide

Seattle Travel & Tourism Guide specializing in Hotels, Attractions, Restaurants, Real Estate, Nightlife & Local Business Yellow Page Listings.

www.**seattle**.com · Cached page · Mark as spam

Ads

60 Hotels in **Seattle**
Save up to 50% on your reservation. Book online now. Pay at the hotel!
www.booking.com/Seattle-WA

Seattle Hotel Deals
Save On Hotels In **Seattle**. Great Rates From 100s Of Sites
www.KAYAK.com/SeattleHotels

33 **Seattle** Attractions
Space Needle, Activities, Tours, Museums & More. 1 Low Price!
GoSeattleCard.com

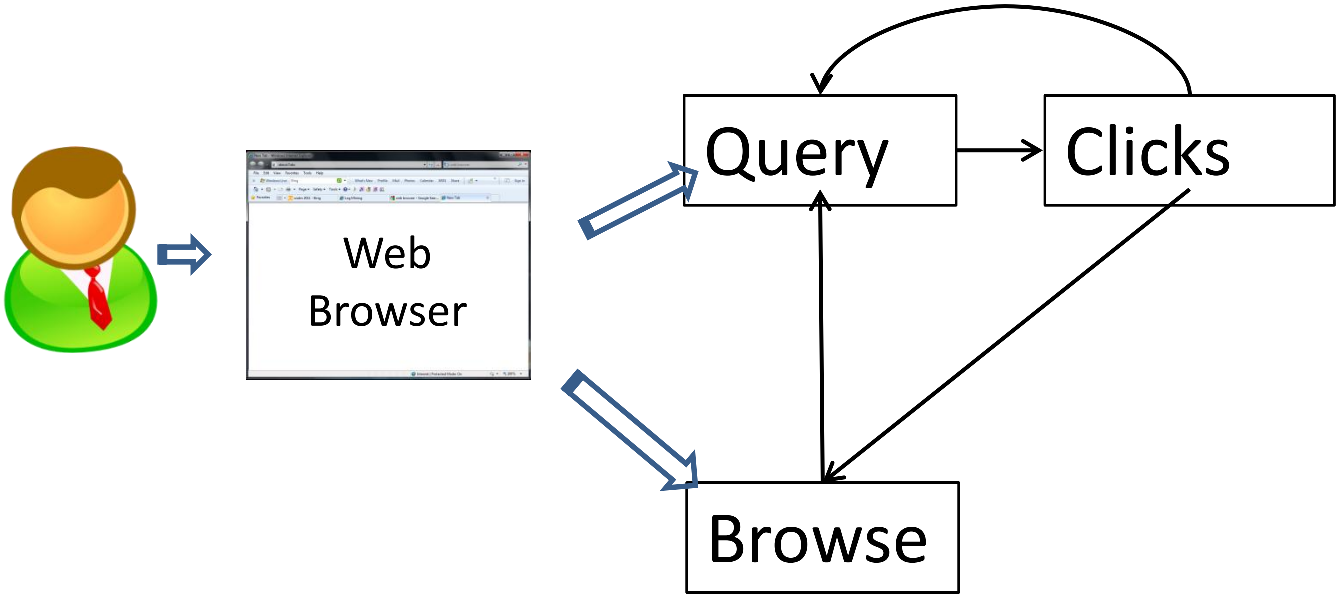
Meet Single Men
Meet Singles for Love, Romance & Friendship. Join Today & Browse!
FriendFinder.com

Seattle Tours
Book 5-star rated **Seattle** tours and activities on Viator.
www.viator.com/seattle

See your message here

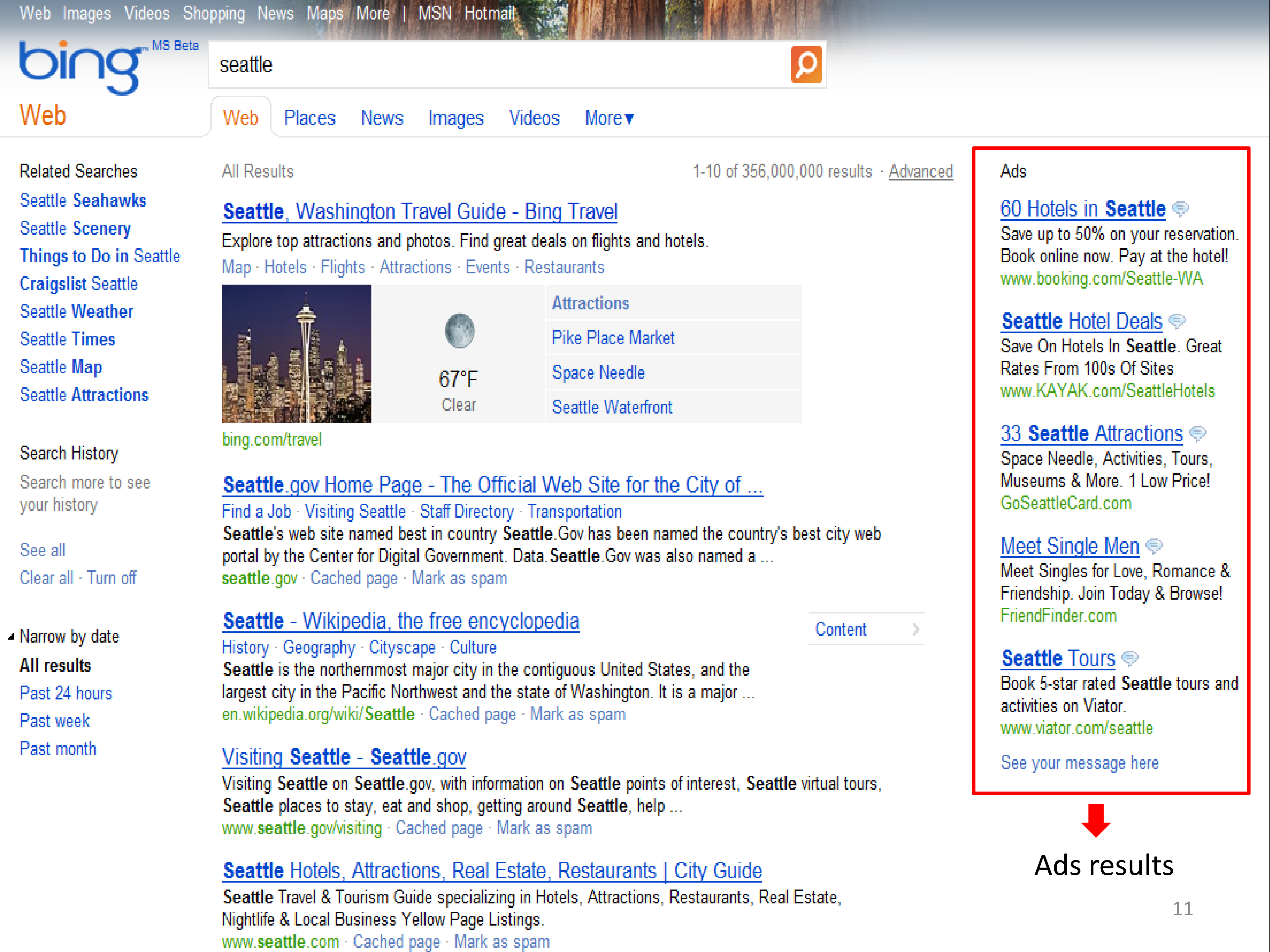
Algo results

User Behavior in Web Browser



Search Results

- Algorithmic results; or “algo results” for short
  - Sometimes referred to as the “ten blue links”
- Advertisement results; or “ad results” for short
  - Sometimes referred to as the “sponsor links”


Web Images Videos Shopping News Maps More | MSN Hotmail
bing MS Beta
Web
Web Places News Images Videos More▾

seattle

Related Searches

- Seattle Seahawks
- Seattle Scenery
- Things to Do in Seattle
- Craigslist Seattle
- Seattle Weather
- Seattle Times
- Seattle Map
- Seattle Attractions

Search History

Search more to see your history

See all

Clear all · Turn off

Narrow by date

- All results
- Past 24 hours
- Past week
- Past month

All Results   1-10 of 356,000,000 results · Advanced

Seattle, Washington Travel Guide - Bing Travel
Explore top attractions and photos. Find great deals on flights and hotels.
Map · Hotels · Flights · Attractions · Events · Restaurants
67°F
Clear
Attractions
Pike Place Market
Space Needle
Seattle Waterfront
bing.com/travel

Seattle.gov Home Page - The Official Web Site for the City of ...
Find a Job · Visiting Seattle · Staff Directory · Transportation
Seattle's web site named best in country Seattle.Gov has been named the country's best city web portal by the Center for Digital Government. Data.Seattle.Gov was also named a ...
seattle.gov · Cached page · Mark as spam

Seattle - Wikipedia, the free encyclopedia   Content ›
History · Geography · Cityscape · Culture
Seattle is the northernmost major city in the contiguous United States, and the largest city in the Pacific Northwest and the state of Washington. It is a major ...
en.wikipedia.org/wiki/Seattle · Cached page · Mark as spam

Visiting Seattle - Seattle.gov
Visiting Seattle on Seattle.gov, with information on Seattle points of interest, Seattle virtual tours, Seattle places to stay, eat and shop, getting around Seattle, help ...
www.seattle.gov/visiting · Cached page · Mark as spam

Seattle Hotels, Attractions, Real Estate, Restaurants | City Guide
Seattle Travel & Tourism Guide specializing in Hotels, Attractions, Restaurants, Real Estate, Nightlife & Local Business Yellow Page Listings.
www.seattle.com · Cached page · Mark as spam

Ads

60 Hotels in Seattle
Save up to 50% on your reservation. Book online now. Pay at the hotel!
www.booking.com/Seattle-WA

Seattle Hotel Deals
Save On Hotels In Seattle. Great Rates From 100s Of Sites
www.KAYAK.com/SeattleHotels

33 Seattle Attractions
Space Needle, Activities, Tours, Museums & More. 1 Low Price!
GoSeattleCard.com

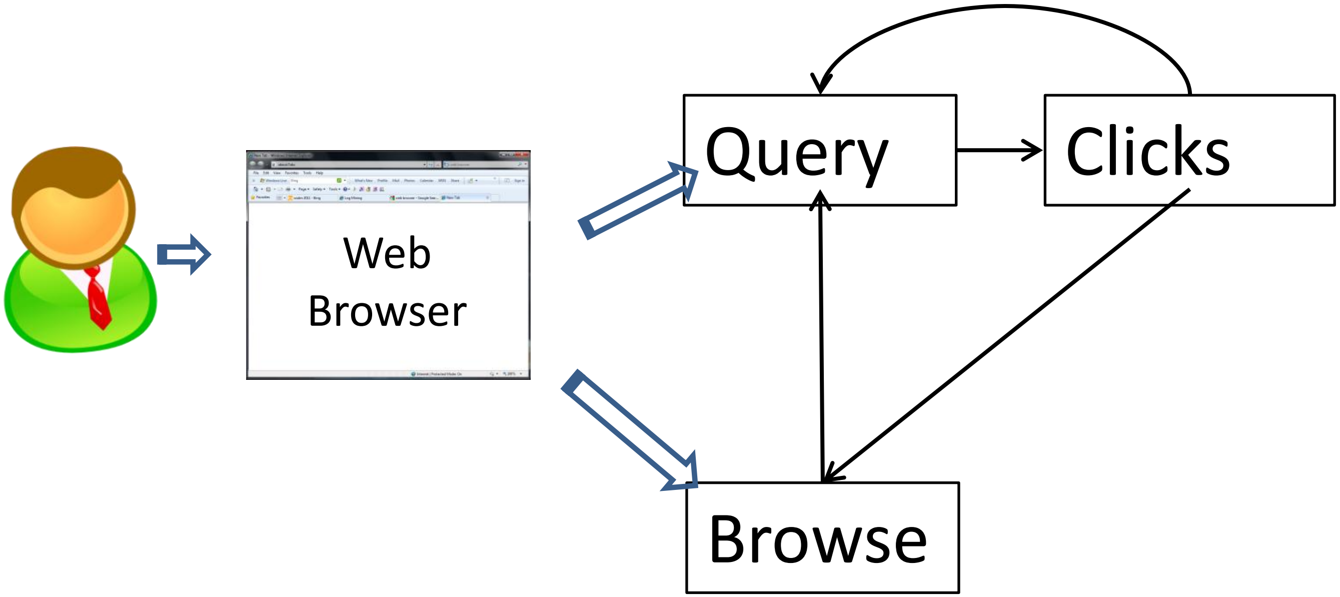
Meet Single Men
Meet Singles for Love, Romance & Friendship. Join Today & Browse!
FriendFinder.com

Seattle Tours
Book 5-star rated Seattle tours and activities on Viator.
www.viator.com/seattle

See your message here

Ads results

User Behavior in Web Browser



Web Browser

Query

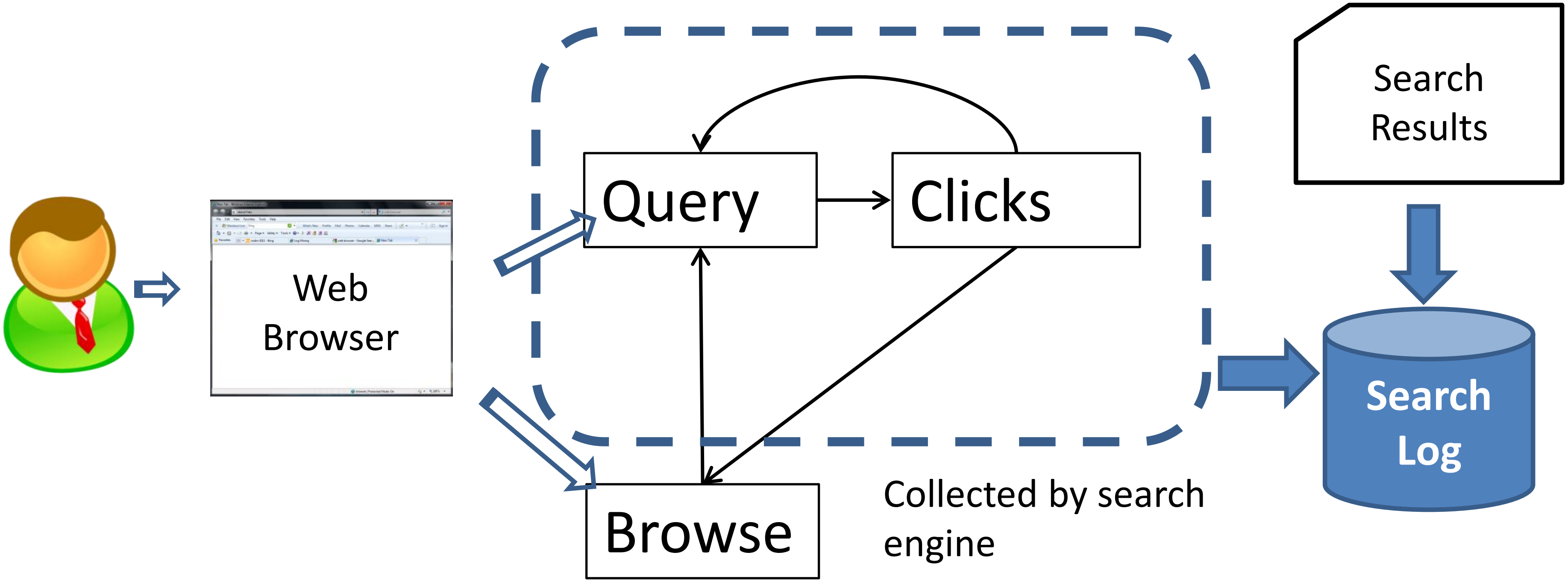
Clicks

Browse

Search Results

- Algorithmic results; or “algo results” for short
  - Sometimes referred to as the “ten blue links”
- Advertisement results; or “ad results” for short
  - Sometimes referred to as the “sponsor links”

Search Logs

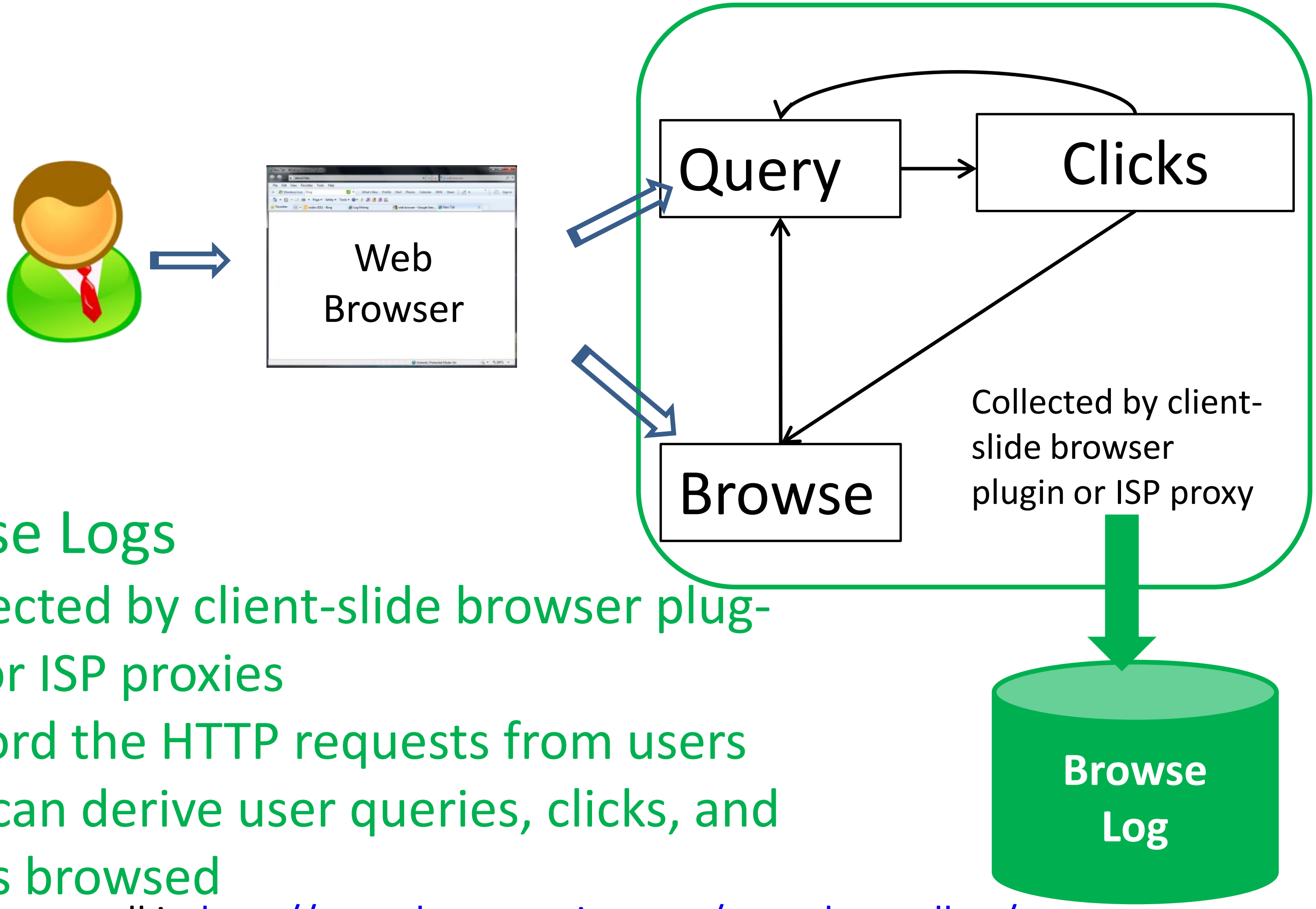


Collected by search engine

Search Logs

- Collected by search engine
- Recording user queries, clicks, as well as search results provided by search engines

Browse Logs



Collected by client-slide browser plugin or ISP proxy

Browse Logs

- Collected by client-slide browser plug-ins or ISP proxies
- Record the HTTP requests from users
- We can derive user queries, clicks, and URLs browsed

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.

Major Information in Search Logs

- Recorded by search engine severs
- Four categories of information
  - User info: user ID & IP
  - Query info: terms in query, time stamp, location, search device, etc.
  - Click info: URL, time stamp, etc.
  - Search results
    - Algo results, Ad 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
  - Collecting data only when user permission is granted
  - User can choose to opt-out at any time

Search Logs VS. Browse Logs

	Search Logs	Browse Logs
Common	User ID & IP, queries, clicks	
Diff 1	Collected by search engines	Collected by browser plug-ins or ISP proxies
Diff 2	Contains search results, position and type of clicks	No search results info
Diff 3	No browse info	Contains browse info

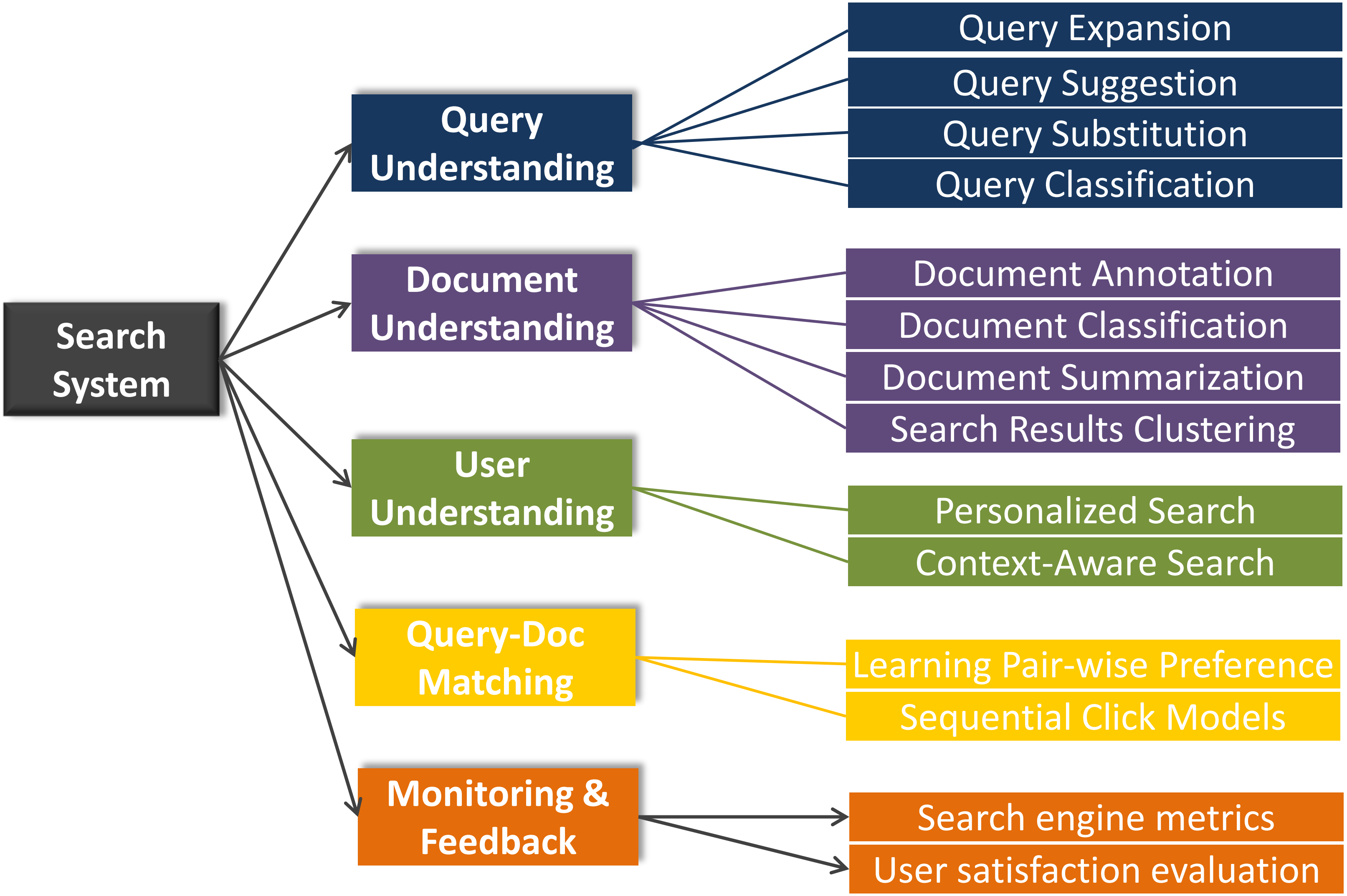
Log Mining Applications

- Categorization by efficiency versus effectiveness [Silvestri09]
  - Enhancing efficiency of search systems
  - Enhancing effectiveness of search systems

In this tutorial, we only focus on the effectiveness aspect

Fabrizio Silvestri and Ricardo Baeza-Yates. Query Log Mining. WWW’09 tutorial.

Log Mining Application Examples



Search System

Query Understanding

- Query Expansion
- Query Suggestion
- Query Substitution
- Query Classification

Document Understanding

- Document Annotation
- Document Classification
- Document Summarization
- Search Results Clustering

User Understanding

- Personalized Search
- Context-Aware Search

Query-Doc Matching

- Learning Pair-wise Preference
- Sequential Click Models

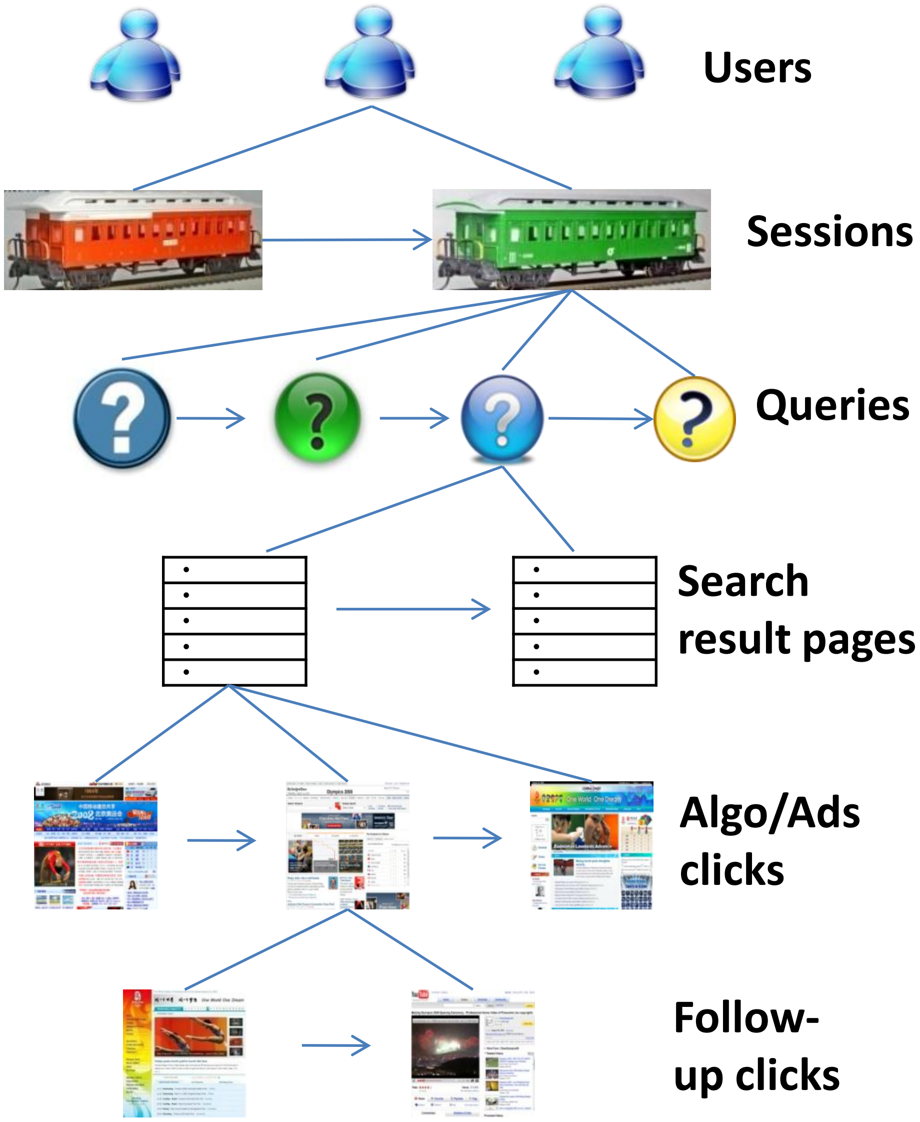
Monitoring & Feedback

- Search engine metrics
- User satisfaction evaluation

Summarizing Raw Log Data

- Raw log data are stored in the format of plain text: unstructured data
  - Huge amounts, very detailed
- Can we summarize the textual logs using some effective data structures to facilitate various log mining applications?
- Challenges: complex objects, diverse and complicated applications

Complex Objects



Users

Sessions

Queries

Search result pages

Algo/Ads clicks

Follow-up clicks

- Various types of data objects in log data
- Complex relationship among data objects
  - Hierarchical relationship
  - Sequential relationship

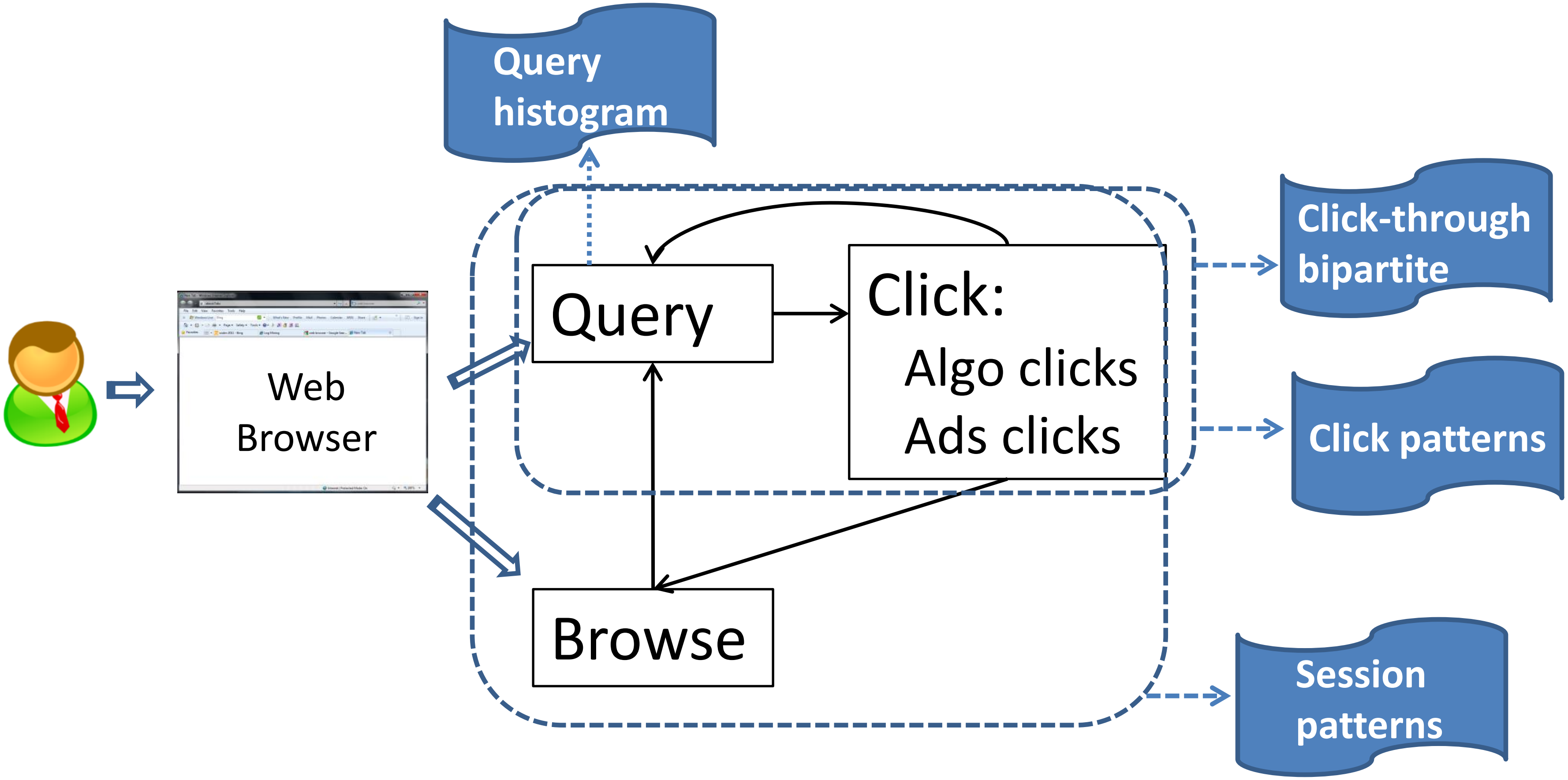
How to describe various objects as well as their relationships?

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 frequently clicked URLs?
  - Given a URL u, what are the top-K queries often leading to a click on u?
- Document understanding
- User understanding

How to provide effective summarization to support various applications?

Popular Data Summarization in Log Mining



Query histogram

Web Browser

Query

Click:

Algo clicks

Ads clicks

Browse

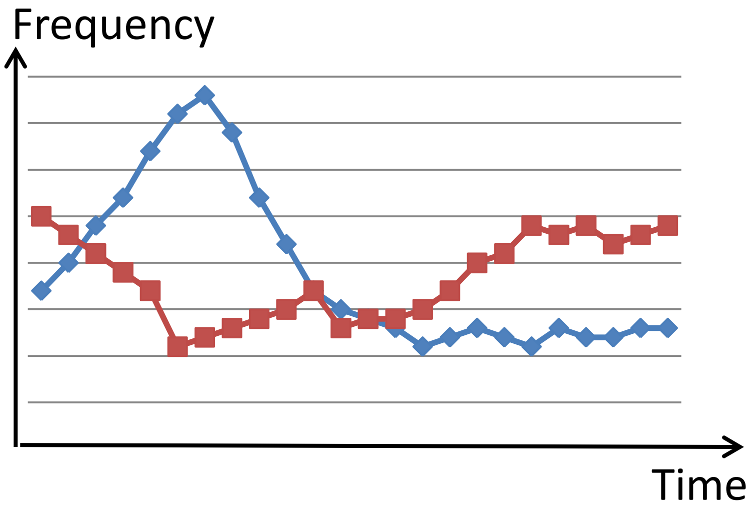
Click-through bipartite

Click patterns

Session patterns

Query Histogram

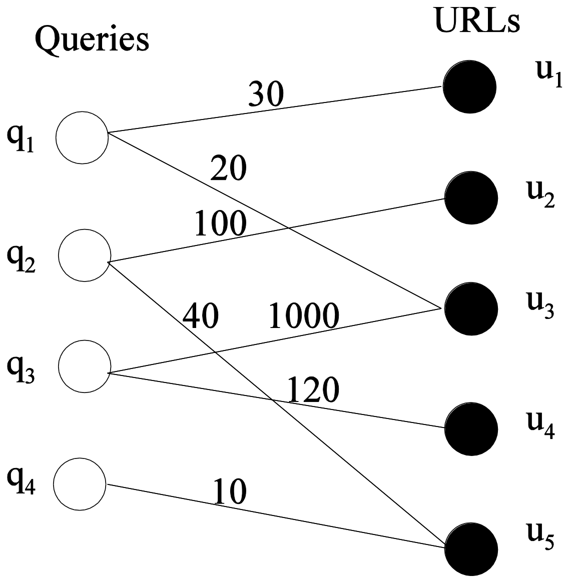
Query String	Count
facebook	3,157 K
google	1,796 K
youtube	1,162 K
myspace	702 K
facebook com	665 K
yahoo	658 K
yahoo mail	486 K
yahoo com	486 K
ebay	486 K
facebook login	445 K



Example applications:

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

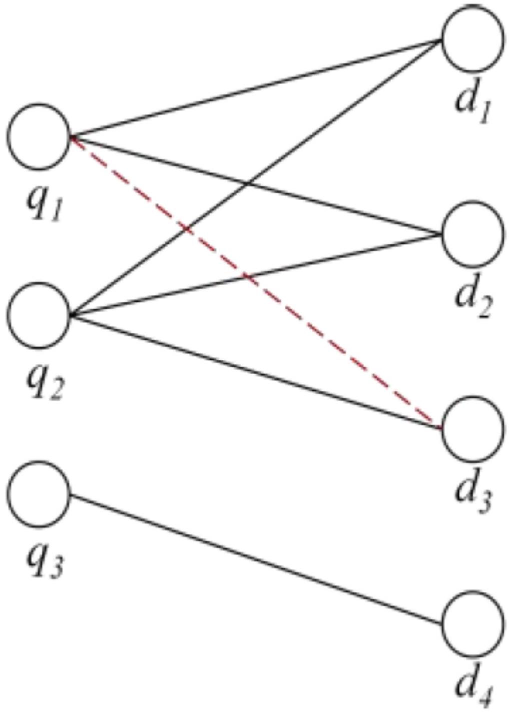
Click-through Bipartite



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 d3 is connected with q2 only; after a random walk expansion, d3 is also connected with q1, which has similar neighbors as q2

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

Click Pattern

Query

×	Doc 1
	Doc 2
	…
×	…
	…
	…
	…
	…
	…
	Doc N

Pattern 1 (count)

	Doc 1
×	Doc 2
	…
	…
	…
	…
	…
	…
	…
×	Doc N

Pattern 2 (count)

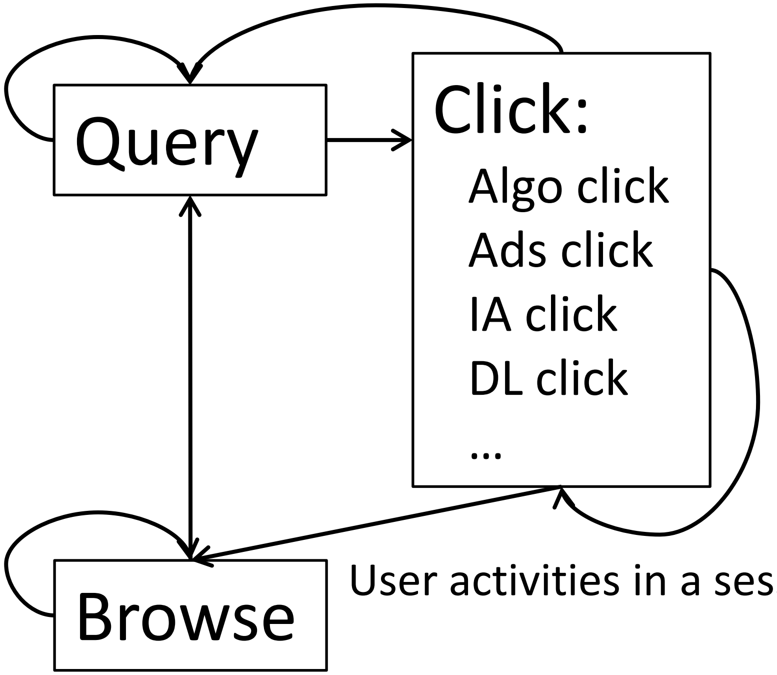
…

×	Doc 1
×	Doc 2
	…
×	…
	…
	…
	…
	…
	…
	Doc N

Pattern n (count)

- 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)

Session Patterns



Query

Click:

Algo click

Ads click

IA click

DL click

…

Browse

User activities in a session

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

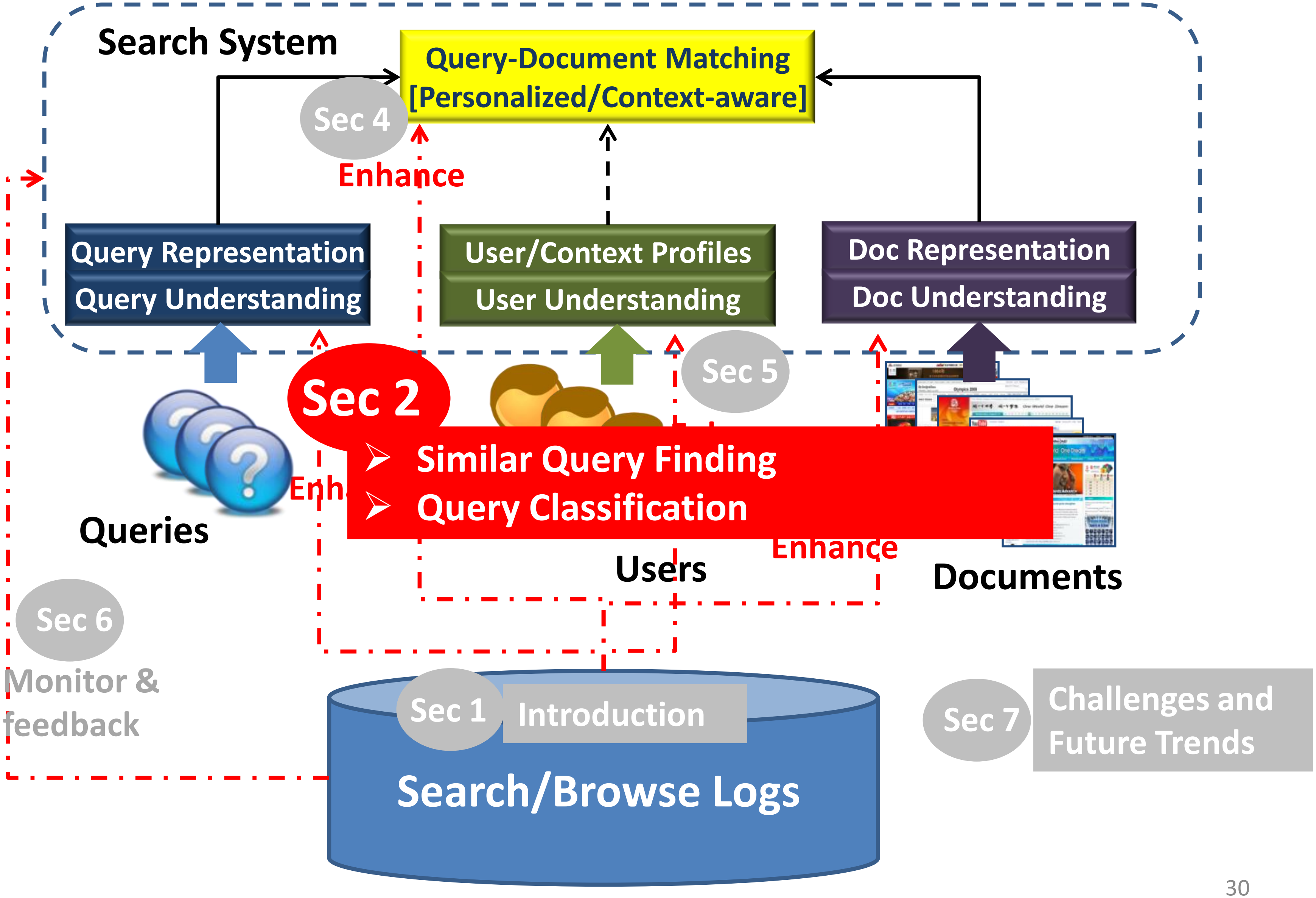
- Example applications
  - Doc (re-)ranking
  - Query suggestion
  - Site recommendation
  - User satisfaction prediction

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

Summary of Introduction

- Search & browse logs
  - Search logs: collected by search engine servers; store queries, clicks, and search results
  - Browse logs: collected by client-side browser plug-ins or ISP proxy servers; store queries, clicks, and browse information
- Log mining applications
  - Query understanding, document understanding, user understanding, query-document matching, monitoring & feedback
- Frequently-used data summarization
  - Query histogram, click-through bipartite, click patterns, session patterns

Road Map



Search System

Query-Document Matching [Personalized/Context-aware]

Sec 4

Enhance

Query Representation

Query Understanding

User/Context Profiles

User Understanding

Doc Representation

Doc Understanding

Sec 5

Sec 2

- ➢ Similar Query Finding
- ➢ Query Classification

Queries

Enh

Users

Enhance

Documents

Sec 6

Monitor & feedback

Sec 1

Introduction

Search/Browse Logs

Sec 7

Challenges and Future Trends

Query Understanding Using Log Data

- Query understanding: receive queries and represent them in certain forms
  - In Traditional IR: usually represented by terms
  - In Web search: queries are often short, ambiguous, and error-prone
- Using log data to enhance query representation
  - Similar query finding : represent queries by groups
    - Refined queries (e.g., spelling error correction)
    - Related queries (e.g., more specific and general queries)
  - Query classification: represent queries by meta-data
    - User goals (informational, navigational, transactional)
    - Topics (e.g., ODP taxonomy)
    - Time sensitivity
    - Location sensitivity
    - More dimensions…

Applications of Similar Query Finding

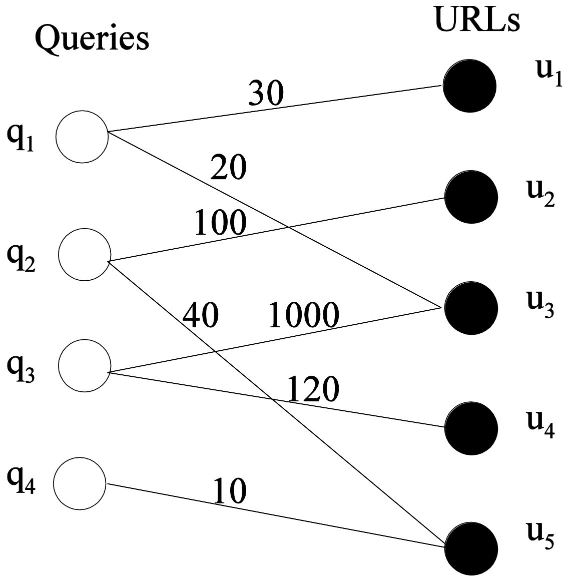
- Query expansion
  - Rewrite a query to increase search recall
    - Example: ‘ny times’ -> `ny times new york’
- Query substitution
  - Rewrite a query into its “standard” form
    - Spelling error correction: e.g., ‘machin learning’ -> ‘machine learning’
    - Acronym expansion: e.g., ‘msr’->’microsoft research’
    - Word merging/splitting : e.g., ‘on line book store’-> ‘online bookstore’
- Query suggestion
  - Provide recommendations to users
    - Specialization: e.g., ‘harry potter’ by ‘harry potter books’
    - Generalization: e.g., ‘seattle employment rate’ by ‘employment rate’
    - Association: e.g., ‘walmart’ by ‘sears’

Using log data for similar query finding

- Using log data for similar query finding
  - Using click-through data
  - Using session data

Methods Using Click-Through Data

- Build a click-through bipartite from log data
- Measure the similarity of queries
  - Overlap of clicked document [Beeferman00], [Wen01], [Cao08]
    - Example: both "MSRA" and "microsoft research asia" lead to clicks on http://research.microsoft.com/en-us/labs/asia



- Cluster queries
  - Agglomerative hierarchical method [Beeferman00], DBScan [Wen01], K-means [Yates04]

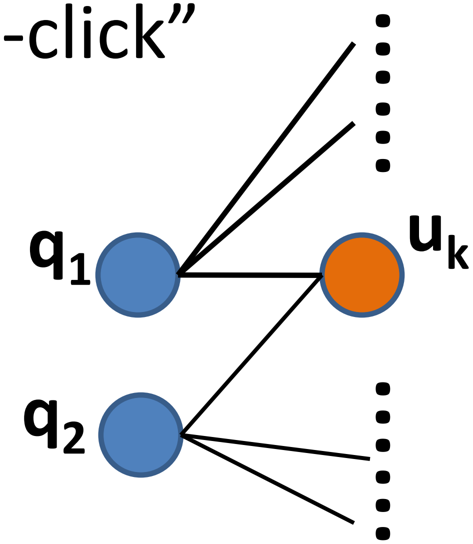
Challenges of Clustering a Click-Through Bipartite

- A click-through bipartite can be huge
  - Millions or billions of unique queries
- Data set is of extremely high dimensionality
  - Millions or billions of unique URLs
- The number of clusters is unknown
- Search logs increase dynamically

Cao, H., et al. Context-aware query suggestion by mining click-through and session data. KDD'08.

Observations from Real Data

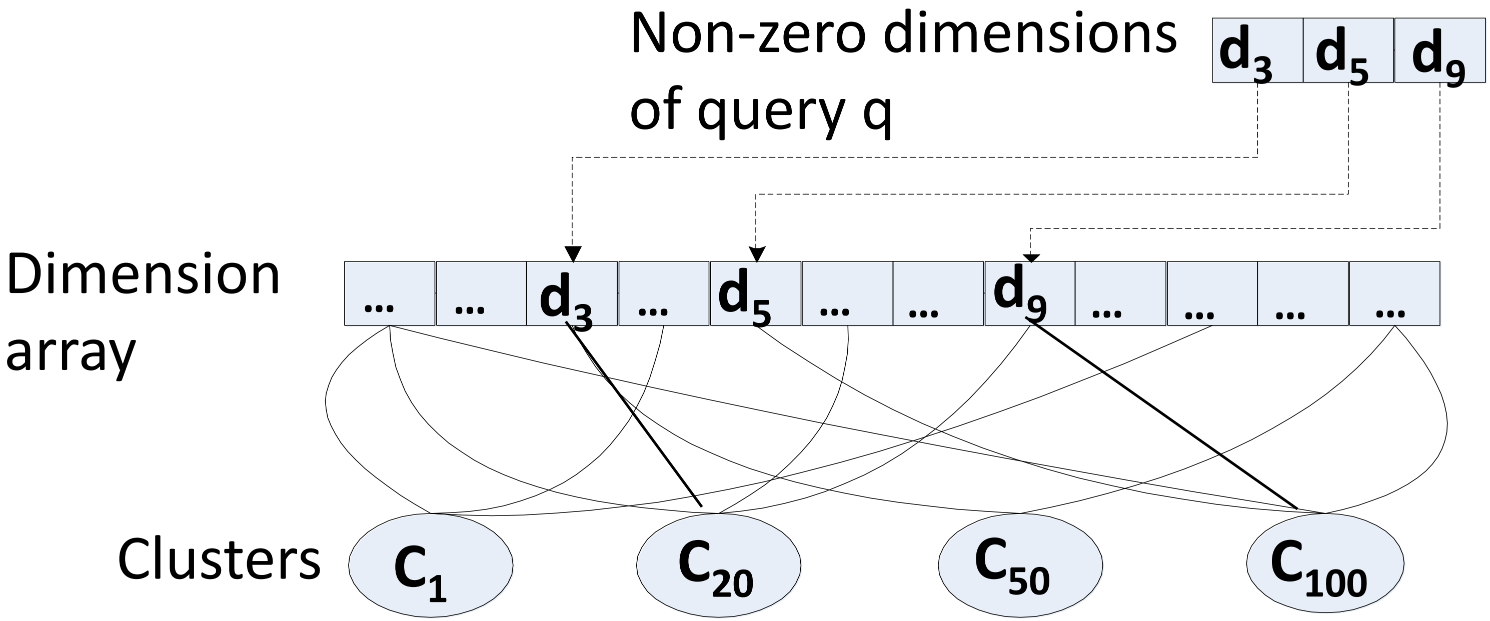
- Average degrees of query and URL nodes are low
  - E.g., average degree of query nodes is 3.1; and average degree of URL nodes is only 3.7.
- On average, a query has only a few "co-click" queries
  - Average number of co-click queries is upper-bounded by $3.1 \cdot (3.7 - 1) = 8.37 < 9$
  - Only need to consider a small number of co-click queries



"Co-click" queries

Cao, H., et al. Context-aware query suggestion by mining click-through and session data. KDD'08.

Query Stream Clustering Algorithm



Non-zero dimensions of query q

Dimension array

Clusters

- A BIRCH-like algorithm
- Major difference: dimension array instead of cluster feature tree
  - Each element corresponds to one URL
  - d_i — C_j if $\exists q_k \in C_j$ such that q_k is connected to URL u_i
- Only one scan of the data set (details in paper)

Cao, H., et al. Context-aware query suggestion by mining click-through and session data. KDD'08.

Example of Query Clusters

Example Cluster 1	bothell wa
	city of bothell
	bothell washington
	city of bothell wa
	city bothell washington
	city of bothell washington
Example Cluster 2	catcountry
	cat country
	cat country radio
	catcountryradio.com
	cat country radio station

Cao, H., et al. Context-aware query suggestion by mining click-through and session data. KDD'08.

Using log data for similar query finding

- Using log data for similar query finding
  - Using click-through data
  - Using session data

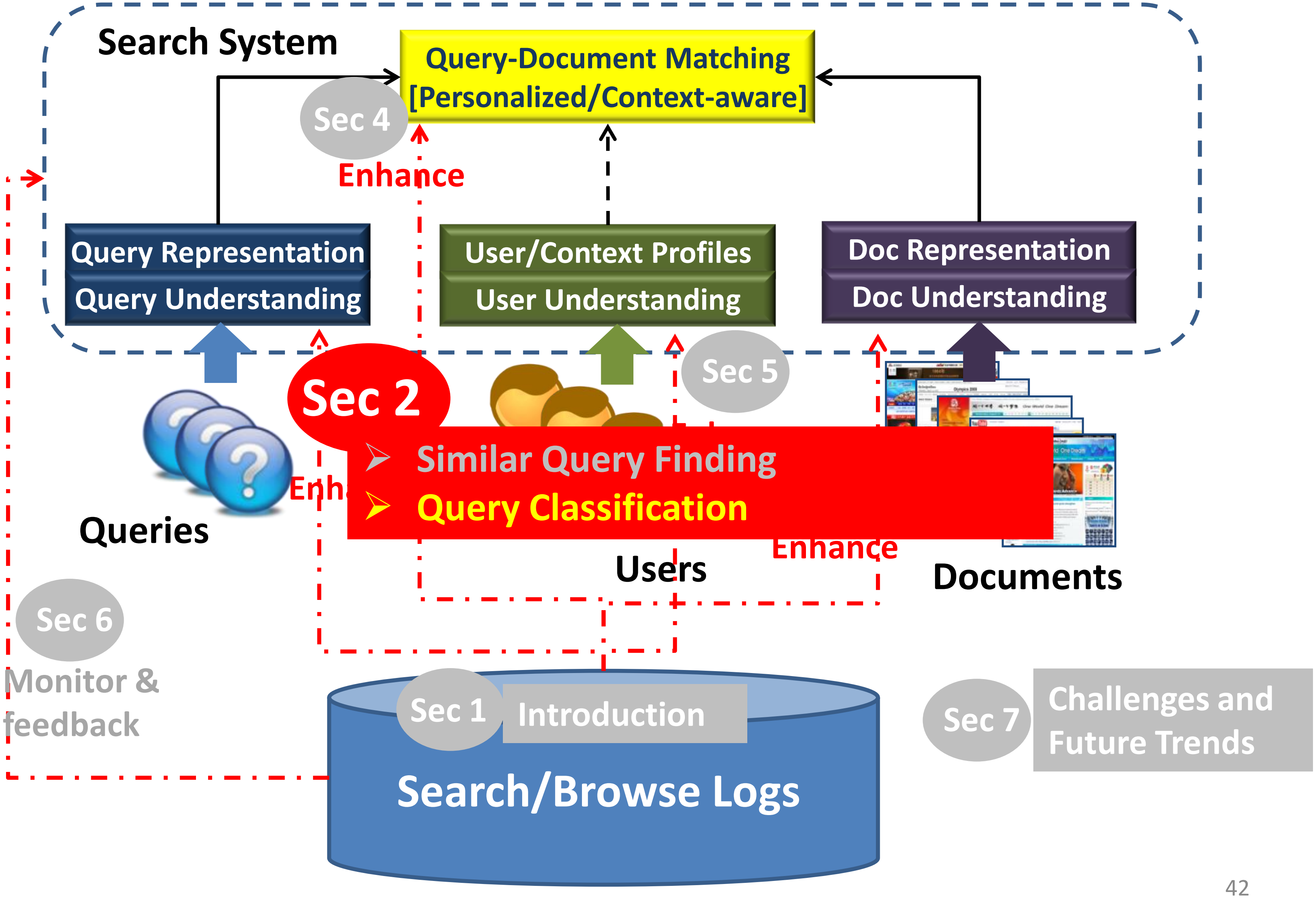
Methods Using Session Data

- Extract sessions from log data
  - E.g., setting up a session boundary if the interval between two adjacent queries exceeds a threshold
- Count co-occurrence or adjacency in sessions
  - If two queries are often adjacent or co-occurring in the same session, they are similar to each other [Jensen06][Huang03][Jones06]
    - Example: “walmart” and “sears” often appear in the same session
- Measure correlation between queries
  - Mutual information, weighted mutual information [Jensen06]
  - Jaccard similarity, dependency, cosine similarity [Huang03]
  - Log likelihood ratio [Jones06]

Why Measuring Query Correlation

“淘宝” & “北京天气”	淘宝	北京天气
100	1,000,000	1,000,000
“中科院软件所” & “中科院计算所”	中科院软件所	中科院计算所
50	1,000	1,000

Road Map



Search System

Query-Document Matching
[Personalized/Context-aware]

Sec 4

Enhance

Query Representation

Query Understanding

User/Context Profiles

User Understanding

Doc Representation

Doc Understanding

Sec 5

Sec 2

- Similar Query Finding
- Query Classification

Enh

Queries

Users

Enhance

Documents

Sec 6

Monitor & feedback

Sec 1 Introduction

Search/Browse Logs

Sec 7 Challenges and Future Trends

Query Classification

- Enrich query representation by various meta-data
- Queries can be classified on multiple dimensions
  - User goals (navigational, informational, transactional)
  - Topics (ODP categories, auto-created concepts)
  - Time-sensitiveness (e.g., ‘WWW conference’)
  - Location-sensitiveness (e.g., ‘pizza’)
  - More dimensions…
- Using log data for query classification

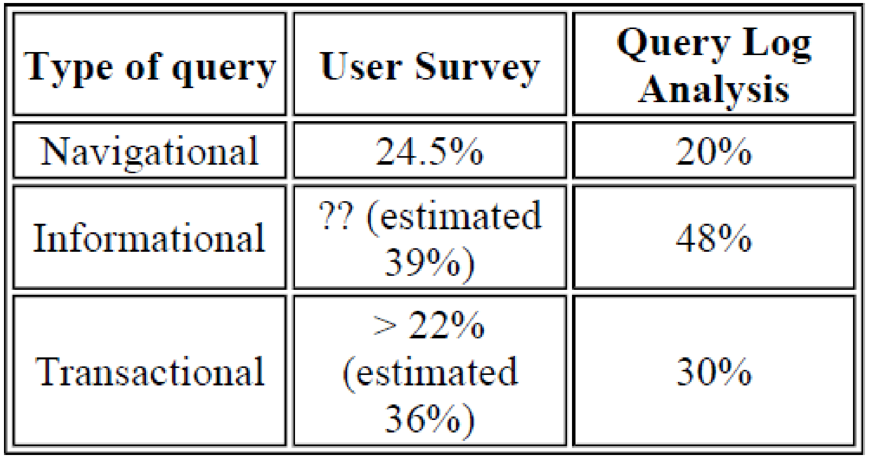
Using Log Data for Query Classification

- Using ***click patterns*** for classifying navigational/informational queries [Lee05]
- Using ***click-through bipartite*** for classifying query topics [Fuxman07][Li08]

User Goals

- According to Broder [Broder02]
  - Navigational. The immediate intent is to reach a particular site.
  - Informational. The intent is to acquire some information assumed to be present on one or more web pages.
  - Transactional. The intent is to perform some web-mediated activity.

- Approaches to automatically classifying user goals
  - Using Web pages
    - Kang and Kim [kang03]
  - Using log data and anchors
    - Lee et al. [lee05]



Type of query	User Survey	Query Log Analysis
Navigational	24.5%	20%
Informational	?? (estimated 39%)	48%
Transactional	> 22% (estimated 36%)	30%

Broder, A. 2002. A Taxonomy of Web Search. SIGIR Forum. 36, 2, 3-10.

Classifying User Goals Using Log Data and Anchor Data

- Only two categories considered, i.e., navigational and informational
- Results
  - Using anchor text data alone: ~75% accuracy
  - Using click-through data alone: ~80% accuracy
  - Combining anchor text and click-through: ~90% accuracy

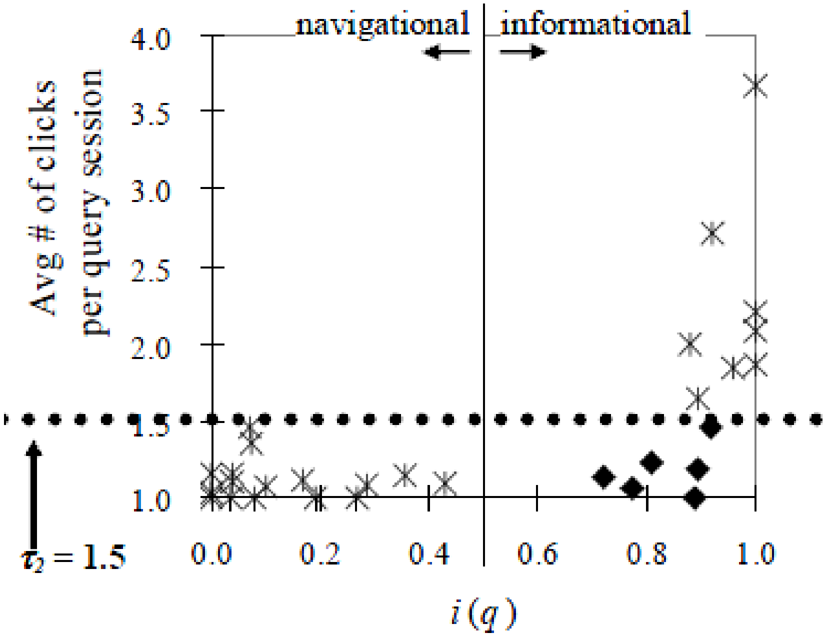
Lee, U. et al. Automatic Identification of User Goals in Web Search. WWW'05.

Using Click-Patterns for Classifying User Goals

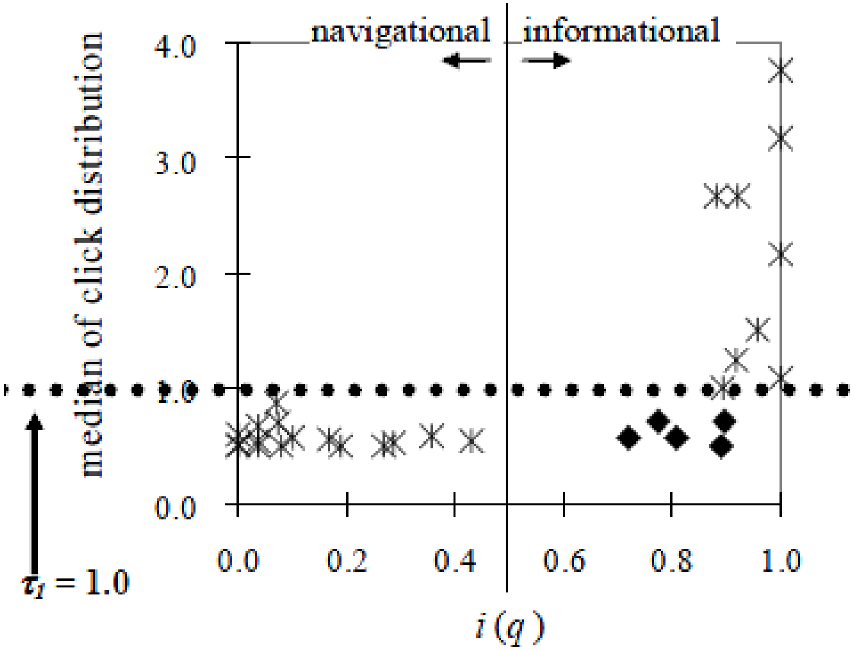
- According to Broder [Broder02]: for navigational queries, the user goal is to reach a particular site
- Heuristics to detect navigational queries
  - Distribution of clicked documents is skewed
  - Number of clicks per query is small

Lee, U. et al. Automatic Identification of User Goals in Web Search. WWW'05.

Effectiveness of Heuristics



Number of clicks per query (accuracy: 80%)



Median of click distribution (accuracy: 83.3%)

- Left figure: if the number of clicks per query is used, accuracy: 80%
- Right figure: if the median of click distribution is used, accuracy: 83.3%

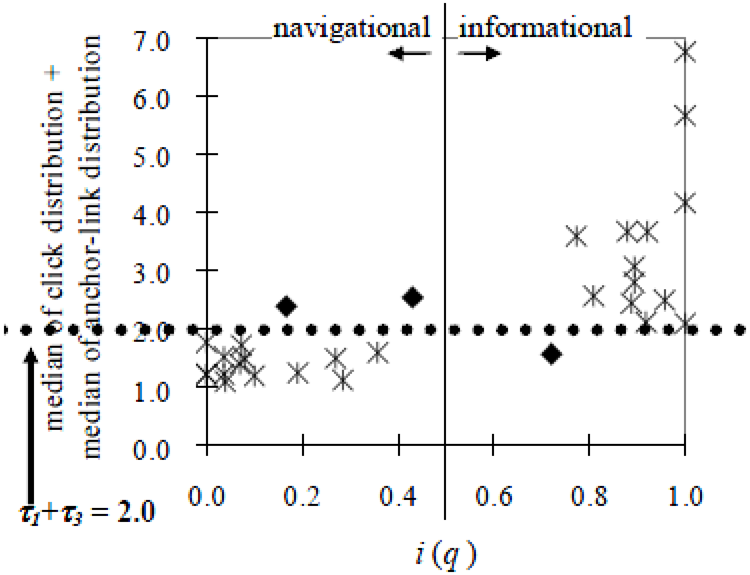
Lee, U. et al. Automatic Identification of User Goals in Web Search. WWW'05.

Combining Click-Through Features with Anchor Text Features

- Linear combination

$$f = w_1 \cdot f_1 + w_2 \cdot f_2 + \cdots + w_n \cdot f_n$$

- A simple combination shows a better accuracy



- Combines two features
- Equal weights
- Accuracy reaches 90%

$$f = (\mathit{median\ of\ click\ distribution}) \\ + (\mathit{median\ of\ anchor\ text\ distribut}$$

Lee, U. et al. Automatic Identification of User Goals in Web Search. WWW'05.

Using Log Data for Query Classification

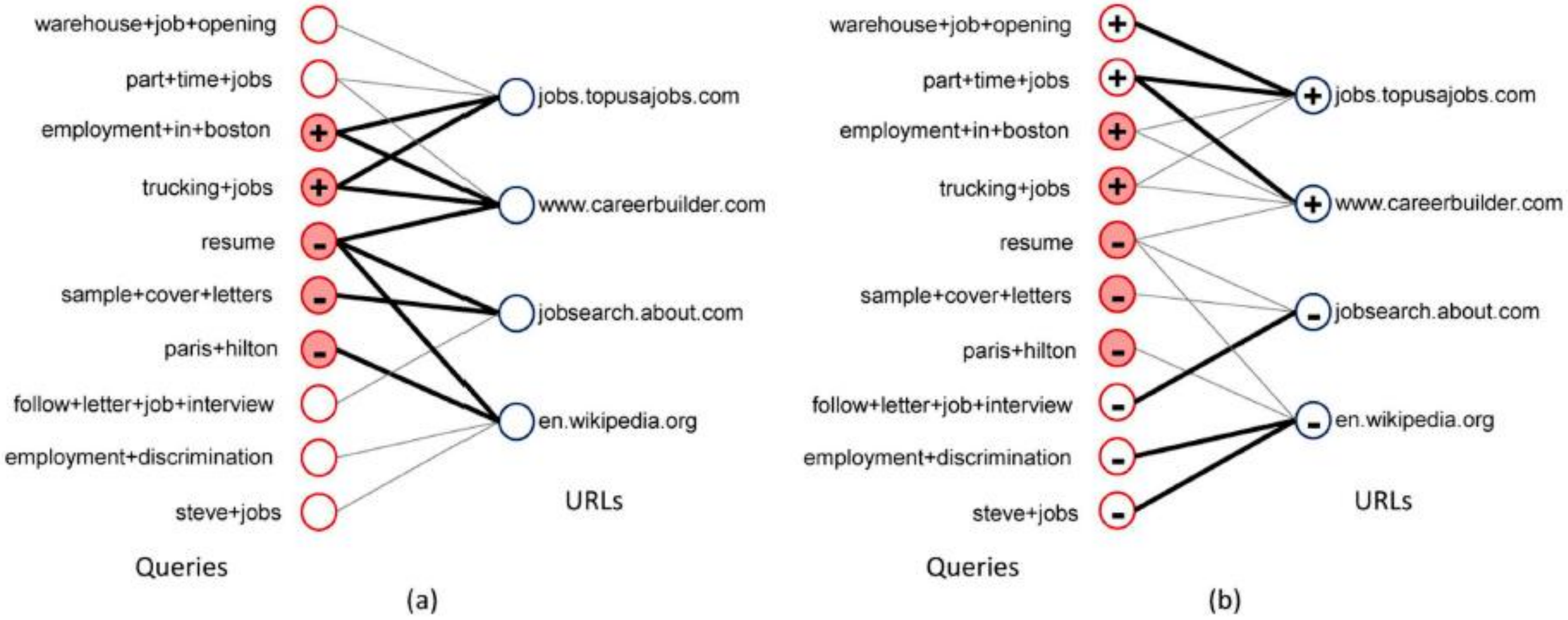
- Using ***click patterns*** for classifying navigational/informational queries [Lee05]
- Using ***click-through bipartite*** for classifying query topics [Fuxman07][Li08]

Classifying Query Topics

- Approaches to automatic classification
  - Traditional methods
    - Directly apply some text classification techniques on query terms
    - Exploit the Web pages returned by a search engine to enrich queries (e.g., [shen05])
  - Using log data
    - Using query histograms (e.g., [Beitzel07])
    - Using click-through bipartite (e.g., [Fuxman07] [Li08])

Using Click-Through Bipartite

- Basic idea: propagating the class labels along the edges of the click-through bipartite



Labeled seeds

After propagation

Figure from Li, X. et al. Learning query intent from regularized click graphs. SIGIR'08

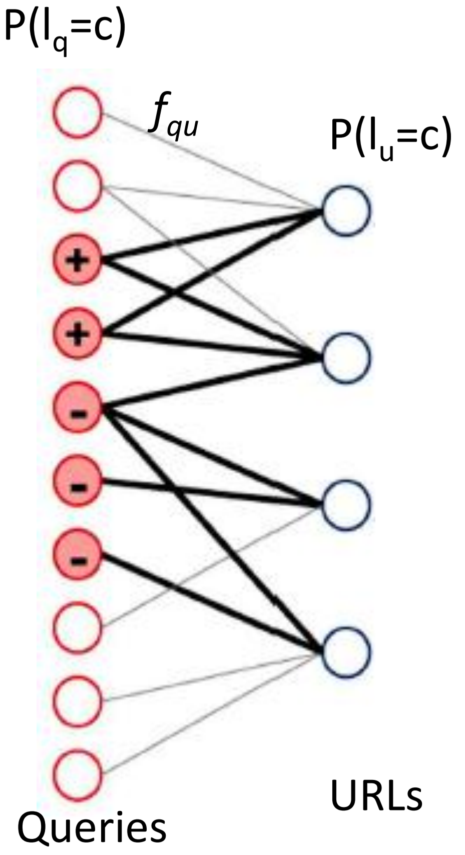
A Random Walk Algorithm [Fuxman08]

- Add a “null” node to the click-through bipartite
  - Each node may walk to the “null” node with probability α
  - The purpose is to penalize long paths
- Iterate between two steps
  - Estimate the probability $P(l_q=c)$

$$P(l_q = c) = (1-\alpha)\sum_{u:(q,u)\in E} w_{qu} P(l_u = c),$$
$$\text{where } w_{qu} = \frac{f_{qu}}{\sum_{u:(q,u)\in E} f_{qu}}$$

  - Estimate the probability $P(l_u=c)$

$$P(l_u = c) = (1-\alpha)\sum_{q:(q,u)\in E} w_{uq} P(l_q = c),$$
$$\text{where } w_{uq} = \frac{f_{qu}}{\sum_{q:(q,u)\in E} f_{qu}}$$

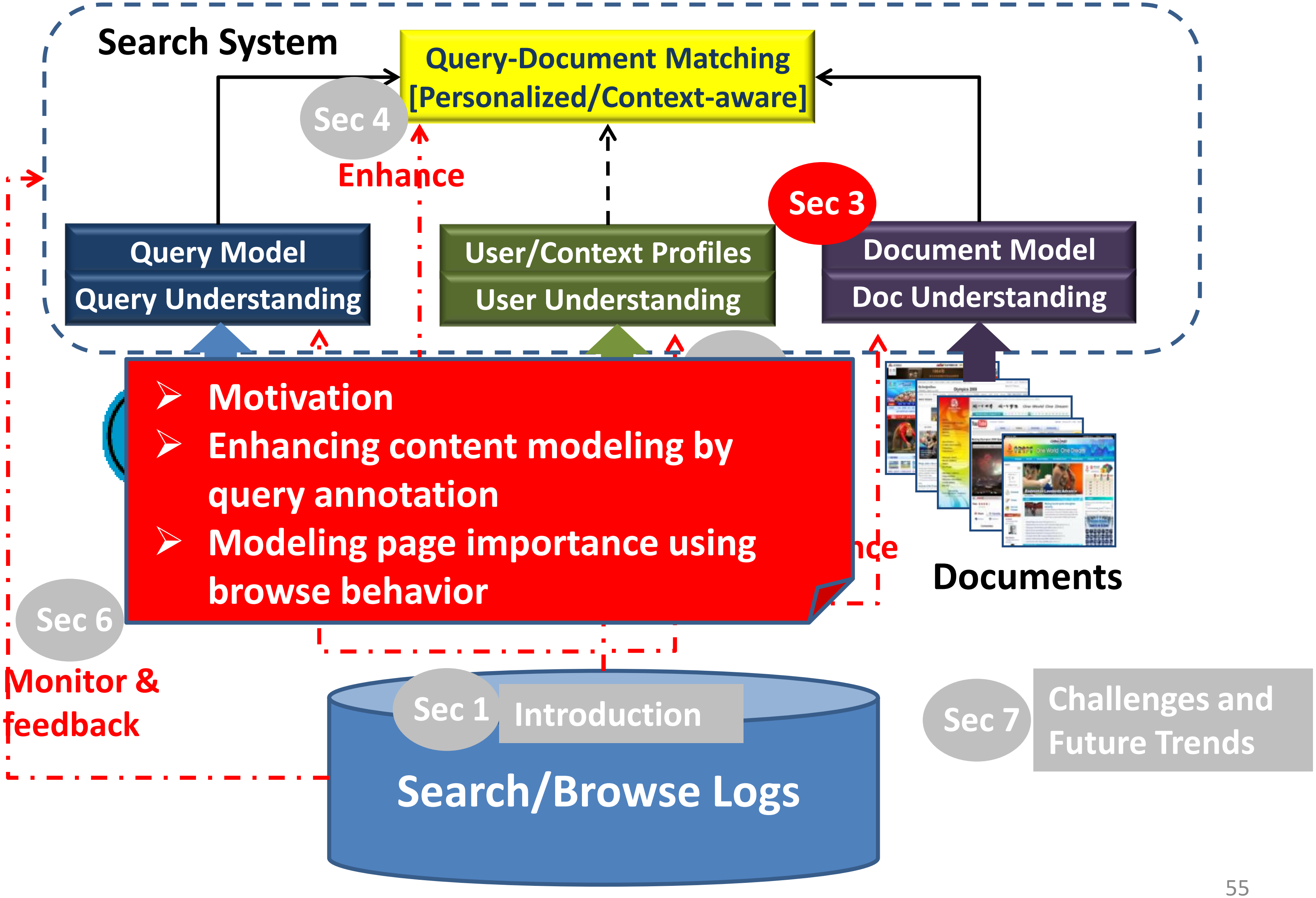


Fuxman, A. et al. Using the wisdom of the crowds for keyword generation WWW'08

Summary for Query Understanding

- Using log data to enhance query representation
- Similar query finding
  - Using click-through data and session data
- Query classification
  - Examples: using click patterns to classify navi/info queries, using click-through bipartite to classify query topics

Road Map



Search System

Query-Document Matching [Personalized/Context-aware]

Sec 4

Enhance

Sec 3

Query Model

Query Understanding

User/Context Profiles

User Understanding

Document Model

Doc Understanding

nce

Documents

Sec 6

Monitor & feedback

Sec 1

Introduction

Search/Browse Logs

Sec 7

Challenges and Future Trends

- ➢ Motivation
- ➢ Enhancing content modeling by query annotation
- ➢ Modeling page importance using browse behavior

Modeling Documents

- In traditional IR, a document is modeled as a bag of words
- Vector space model [Salton1975]
  - $V = \{v_1, \ldots, v_n\}$, the set of terms
  - A document $d = (w_1, \ldots, w_n)$, where w_i is the importance of term v_i with respect to d
  - Importance can be measured by, for example, TFIDF
    - TF(v, d) = # of times term v appears in d
    - IDF(v) = log (# documents in corpus / # of documents containing v)
    - TFIDF(v, d) = TF(v, d) * IDF(v)
- A vector space model tries to capture what the author of a document wants to express using the terms in the document

G. Salton, A. Wong, and C. S. Yang (1975), "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, nr. 11, pages 613–620.

Web Pages and Links

- Web pages contain hyperlinks
  - Anchor text
    - A short annotation on the intension of link
    - Reflect what other page authors think about the target page
    - Modeling the content of pages
  - Link structure
    - Modeling the importance of pages
    - A page having many incoming links tends to be important (well explored by link-based ranking methods, e.g., PageRank)

Using Search and Browse Logs for Document Understanding

- In search logs, we can observe user clicks
  - If a user asks a query Q and clicks on a page P, likely P is related to Q
  - Q can be used as an annotation of P
  - Reflect what the page readers think about P
- In browse logs, user browsing trails can be counted as votes for popular pages

Document Understanding by Text, Hyperlinks and Log Data

Tasks	Text	Hyperlinks	Log data
Modeling content	Bag of words	Anchor text	Query annotations
Modeling importance		Authorities and hubs indicated by the link structure of the Web	Users voting for Web page importance by browsing web pages

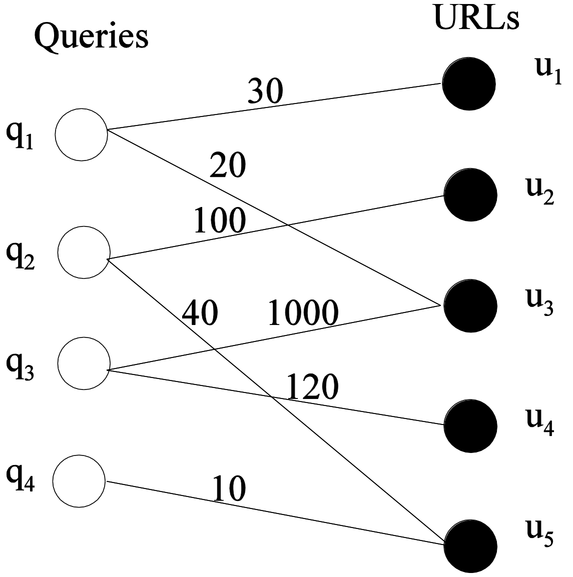
Web users' view on the page

Other page authors' view on the page

Reflect what the page authors want to express

Using Queries as Web Page Annotations

- For each page u_j, let Q_j be the set of queries which are connected with u_j
- Q_j can be considered as the "meta-data" to annotate u_j
- Weight the terms in Q_j in a method similar to TFIDF
  - TF: let $Q_{jt}=\{q|t\in q, q\in Q_j\}$ $TF(t,u_j)= \sum_{q_i\in Q_{jt}} c_{ij}$
  - IDF(t): log (# of queries/# of queries containing t)



G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan. Optimizing web search using web click-through data. CIKM '04.

Applications of Query Annotations

- Web page retrieval
  - [Xue04] Xue G.-R. et al. Optimizing web search using web click-through data. CIKM '04.
  - [Zhao06] Zhao, M. et al. Adapting document ranking to users' preferences using click-through data. AIRS'2006.
  - [Gao09] Gao, J., et al. Smoothing clickthrough data for web search ranking. SIGIR'09.
- Web pages clustering
  - [Poblete08] Poblete, B. and Baeza-Yates, R. Query-sets: using implicit feedback and query patterns to organize web documents. WWW'08
  - [Wang07] Wang X. and Zhai, C. Learn from web search logs to organize search results. SIGIR'07
- Web page summarization
  - [Sun05] Sun, J.-T. et al. Web-page summarization using clickthrough data, SIGIR '05.
- Web directories maintenance
  - [Cid06] Cid, A. et al. Automatic maintenance of web directories using click-through data, in ICDEW '06.

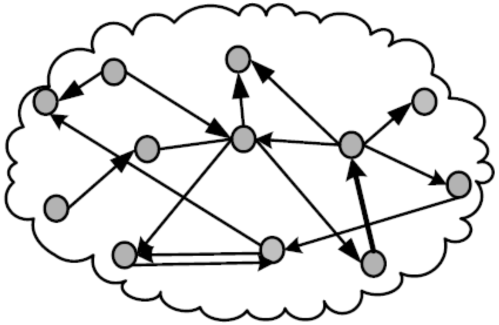
Challenges

- Search log data is sparse, how to handle documents that have very few or even no clicks?
  - Random walk on the click-through bipartite
    - E.g., [Craswell07], [Xue04]
  - Smoothing techniques
    - E.g., [Gao09]
  - Search trails [White07], [Bilenko08]
    - Suppose a user raises a query q, clicks on a search result d_1, and further clicks on a series of hyperlinks in d_1 to reach pages d_2, d_3,…,d_n
    - The sequence "q->d_1->d_2->…->d_n" forms a search trail
    - q can be considered as an annotation to all pages d_1, d_2, …, d_n

Modeling Importance of Web Pages

- An important task in document understanding is to evaluate the importance of web pages
- PageRank
  - A link from one page to another is regarded as an endorsement of the linking page
  - The more links pointed to a page, the more likely the page is important
  - The importance of pages can be propagated in the graph
- HITS
  - A hub page links to many pages
  - An authority page is pointed by many pages
  - Good hubs tend to link to good authorities, and vice versa
- No user feedback is considered

Using User Browsing Behavior



- User browsing graph
  - Vertices representing pages
  - Directed edges representing transitions between pages in browsing history
    - a->b: users browse page b after browsing page a
  - Lengths of user staying time are included
- Using the continuous-time Markov process
  - The stationary probability distribution of the process indicates the importance of web pages
  - Named as “browse rank”

Liu, Y. et al. BrowseRank: letting web users vote for page importance. SIGIR'08.

Example: Spam Fighting

- BrowseRank can push many spam websites to the tail buckets and the number of spam Websites in the top buckets in BrowseRank is smaller than that in PageRank
- Users stay longer on meaningful pages than on spam pages

Liu, Y. et al. BrowseRank: letting web users vote for page importance. SIGIR'08.

ClickRank

- Also leverages users’ browsing information
- Page importance depends on two factors
  - User staying time on pages: users tend to stay on meaningful pages for a non-trivial period of time
  - The visiting order in sessions: the earlier, the more important

Zhu, G, Mishne, G. Mining rich session context to improve web search. KDD'09.

Summary

- Search logs and browse logs can be used to improve document understanding in two aspects
  - Modeling the content of web pages by query annotations
  - Modeling the importance of web pages by users' browsing trails