IIIT-H

Web Mining
Lecture 23: Document Understanding by Log Mining

Manish Gupta

30th 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 22: Query Understanding by Log Mining

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

Announcements

Today’s Agenda

- Motivation
- Enriched models using log data
- Tackling sparsity
- Application examples

Today's Agenda

- **Motivation**
- Enriched models using log data
- Tackling sparsity
- Application examples

Modeling Documents

- Traditionally, a document is modeled as a bag of words
- Vector model
  - V = {v_1, …, v_n}, the set of terms
  - A document d = (w_1, …, w_n), where w_i is the importance of term vi in d
  - Importance can be measured by, for example, TFIDF
- TF(v, d) = # of times term v appears in d
- IDF(v) = log (N/# of documents in the corpus containing v)
- TFIDF(v, d) = TF(v, d) * IDF(v)
- A vector model tries to capture what the author of a document wants to express using the terms in the document

Web Documents and Links

- A Web page may be referred (pointed to) by other Web pages
  - A link to the target page
  - Anchor text: a short annotation on the intension of reference
- A page having many incoming links tends to be important (well explored by link-based ranking methods, e.g., PageRank)
- What does anchor text tell us? – what others on the Web think about the target page

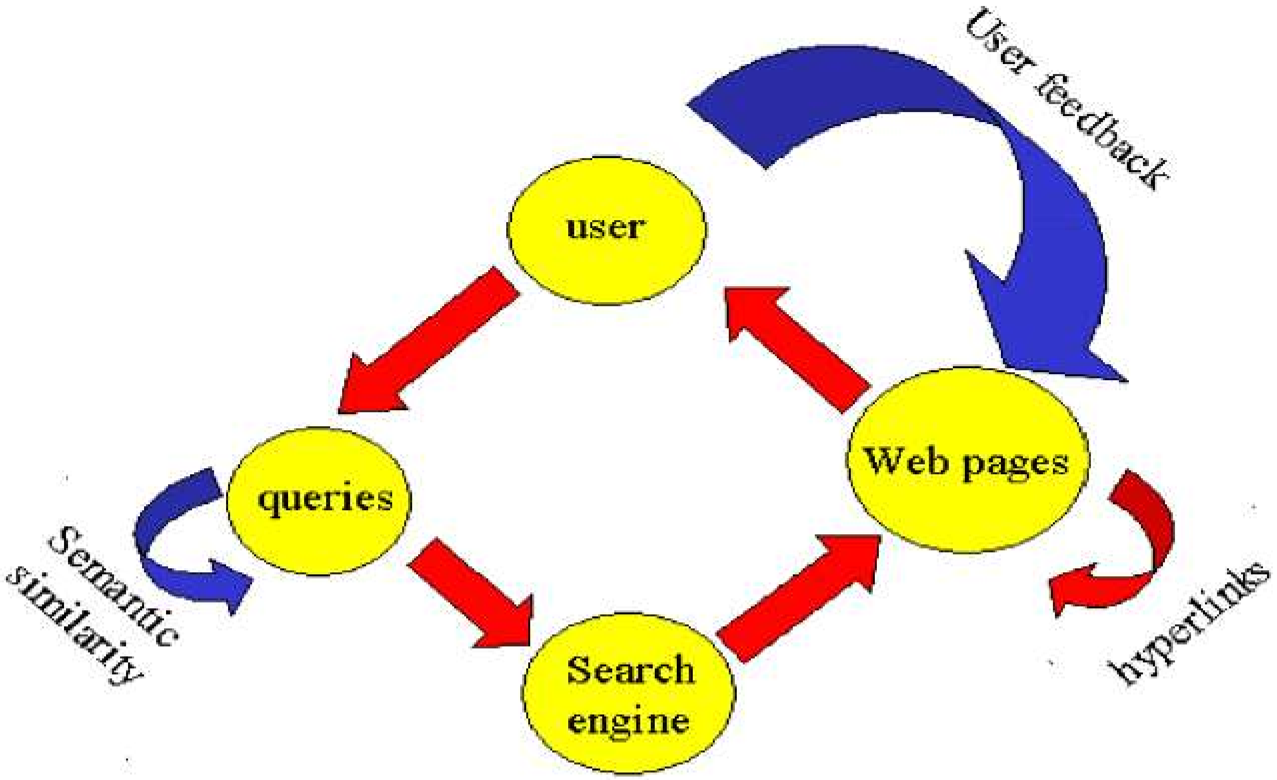
Anchor Text

- Anchor text may not be consistent with the vector model of the target page
  - Example: Anchor text “my publications” → DBLP bibliography server
- Anchor text can be used to complement the vector model of a target Web page
  - What the author writes + what some others read
- Anchor text tends to be bias and static
  - Old pages may receive more references, and thus more anchor text annotations
  - Once a link is made, more often than not, it will not be updated (at least in a long time)

Search Logs as User Annotations

- User queries → search results → user clickthroughs
  - If a user asks a query Q and clicks a page P, likely P is related to Q – Q can be used as an annotation of why a user wants to read P
- User clickthroughs can be used as dynamic, continuously updated, more accurate (after aggregation) annotation of Web pages

A Bigger Picture



- Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query Clustering for Boosting Web Page Ranking AWIC, 2004.

A User Study

- For a query Q
  - Query suggestion generated by frequently asked queries containing Q as a substring or following Q in sessions
  - Query destination – also show Web pages that most clicked by the users asking similar queries
- Two types of tasks
  - Known-item task: “find three tropical storms that have caused property damage and/of loss of life”
  - Exploratory task: “learn about VoIP technology and service providers, select the provider and telephone that best suits you”
- Query suggestion and destination improve user search experience substantially
  - Query destination works particularly well for exploratory tasks
- [R. W. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. SIGIR'07]

Why Can We Learn?

- In query destination, Web pages in the destination recommendation part are selected based on their "query annotation" instead of their content vector model
- Queries as annotation can improve the accuracy of matching user information needs and documents

Challenges

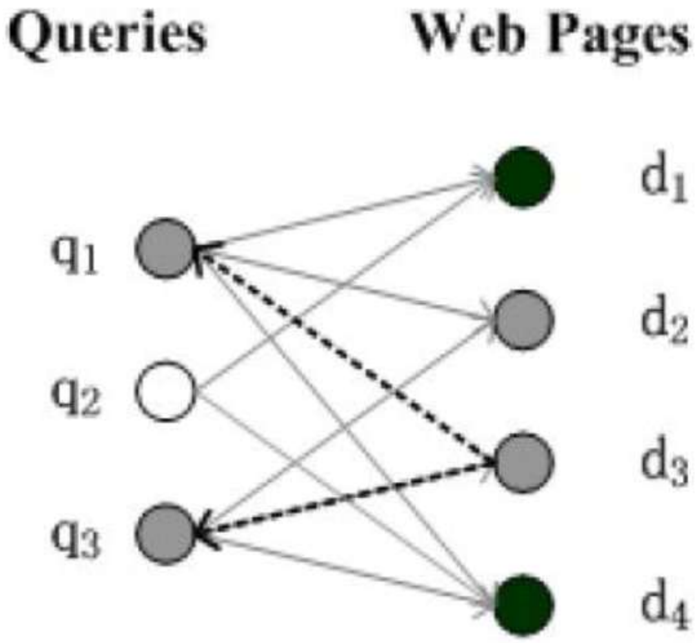
- How to model “query annotations”?
- Search log data is sparse, how to handle documents that have very few or even no clicks?
  - A small number of queries are frequently asked, many queries are rarely asked
  - A small number of Web pages are heavily clicked, many Web pages have very few or even no clicks
- How to use “query annotations”?

Today's Agenda

- Motivation
- **Enriched models using log data**
- Tackling sparsity
- Application examples

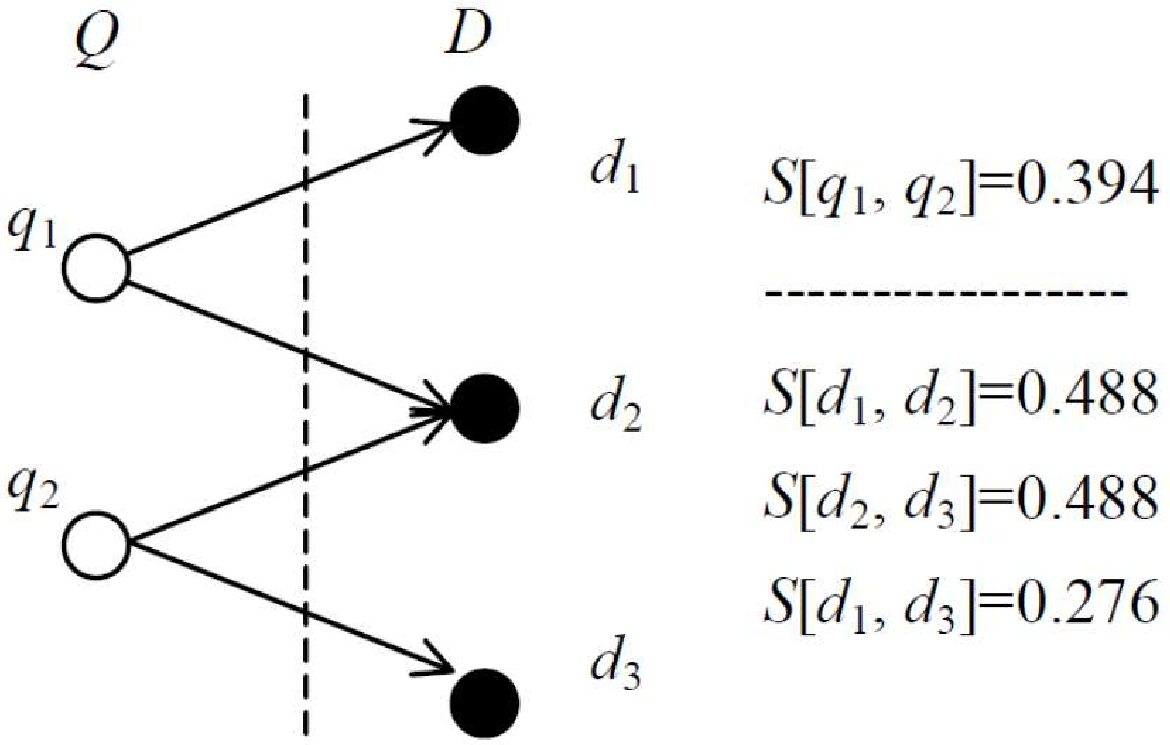
Using Queries as Features

- Queries can be used as features to model documents
- Two documents are similar if they are clicked in the same set of queries
- Using queries as “bridges”, similar documents d_2 and d_3 can be captured
- 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.



Two-Way Annotation

- Can we use documents as features of queries?
- 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.

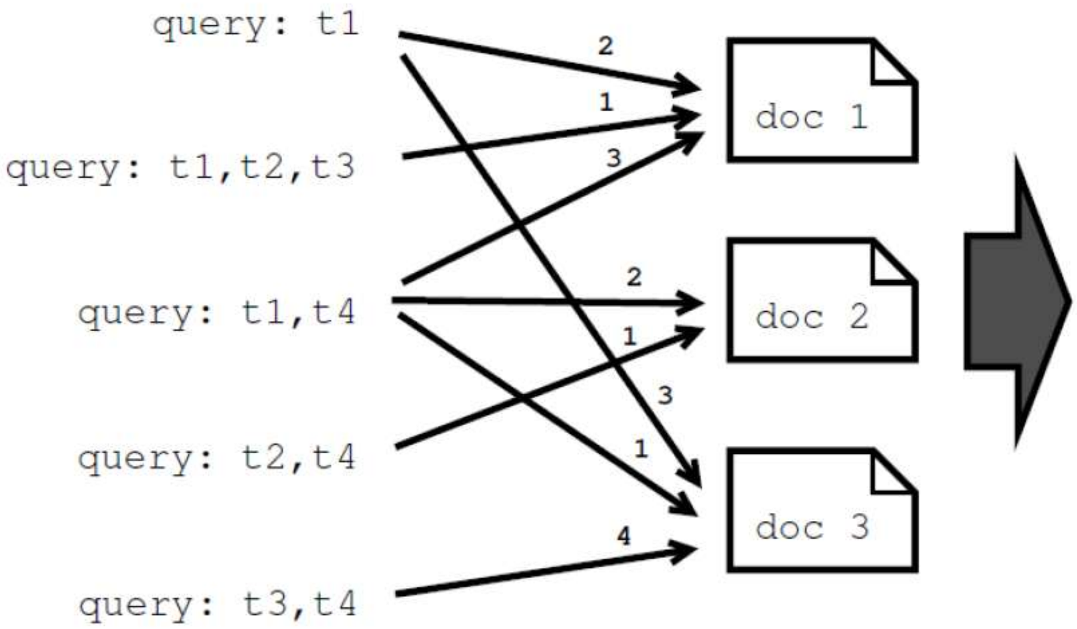


Query-Document Model

- Let V = {t_1, …, t_m} be the vocabulary of all queries in the access log L, where t_1, …, t_m are the terms in V
- Let Q(d) be the set of all queries in L from which users clicked at least one time on d
- Let the frequency of t in Q(d) be the total number of times that queries that contained t were used to visit d
  - $\vec{d} = \langle C_1, \ldots, C_m \rangle$
  - Where C_i = TFIDF(t_i, Q(d))
- [Barbara Poblete, Ricardo Baeza-Yates, Query-sets: using implicit feedback and query patterns to organize web documents. WWW'08.]

Example

- Barbara Poblete, Ricardo Baeza-Yates, Query-sets: using implicit feedback and query patterns to organize web documents. WWW'08



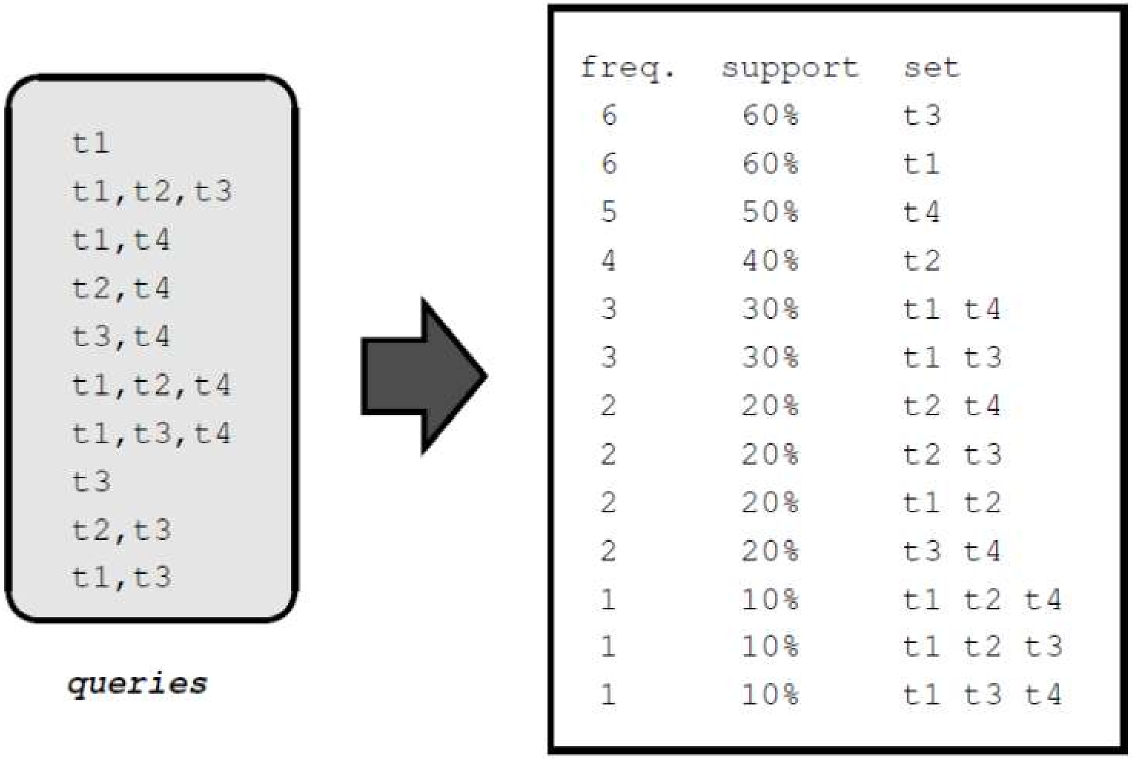
	t1	t2	t3	t4
doc 1:	<6,	1,	1,	3>
doc 2:	<2,	1,	0,	3>
doc 3:	<4,	0,	4,	5>

Query-Set Document Model

- Query-document model considers terms in queries independently even if some of them co-occur frequently
  - "Apple" and "Apple phone" carry very different meanings
- Query-set document model includes frequent term combinations as features for documents
- [Barbara Poblete, Ricardo Baeza-Yates, Query-sets: using implicit feedback and query patterns to organize web documents. WWW'08.]

Example

- Barbara Poblete, Ricardo Baeza-Yates, Query-sets: using implicit feedback and query patterns to organize web documents. WWW'08.



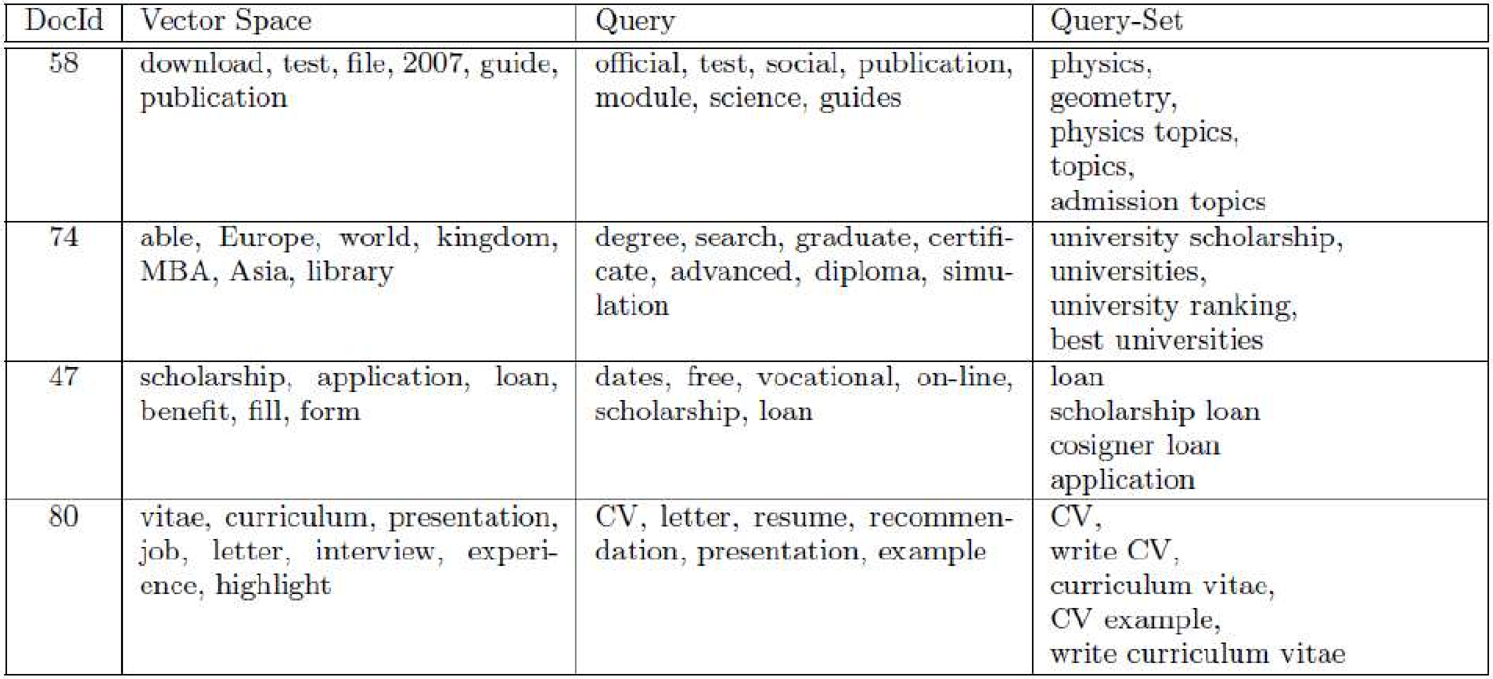
```
t1
t1,t2,t3
t1,t4
t2,t4
t3,t4
t1,t2,t4
t1,t3,t4
t3
t2,t3
t1,t3
```

queries

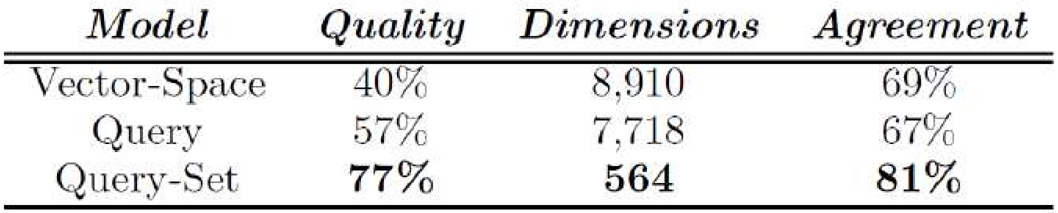
freq.	support	set
6	60%	t3
6	60%	t1
5	50%	t4
4	40%	t2
3	30%	t1 t4
3	30%	t1 t3
2	20%	t2 t4
2	20%	t2 t3
2	20%	t1 t2
2	20%	t3 t4
1	10%	t1 t2 t4
1	10%	t1 t2 t3
1	10%	t1 t3 t4

term sets

Case Study



DocId	Vector Space	Query	Query-Set
58	download, test, file, 2007, guide, publication	official, test, social, publication, module, science, guides	physics, geometry, physics topics, topics, admission topics
74	able, Europe, world, kingdom, MBA, Asia, library	degree, search, graduate, certificate, advanced, diploma, simulation	university scholarship, universities, university ranking, best universities
47	scholarship, application, loan, benefit, fill, form	dates, free, vocational, on-line, scholarship, loan	loan scholarship loan cosigner loan application
80	vitae, curriculum, presentation, job, letter, interview, experience, highlight	CV, letter, resume, recommendation, presentation, example	CV, write CV, curriculum vitae, CV example, write curriculum vitae

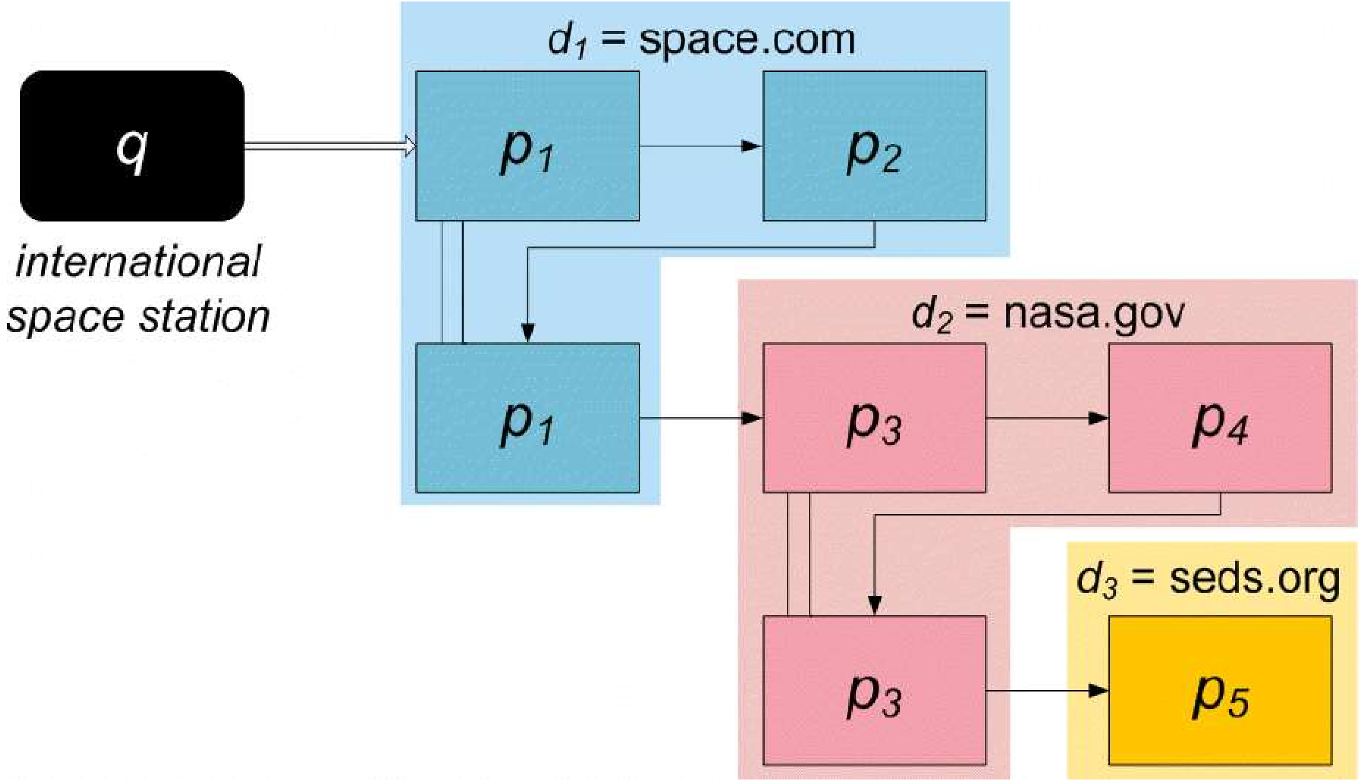


Model	*Quality*	*Dimensions*	*Agreement*
Vector-Space	40%	8,910	69%
Query	57%	7,718	67%
Query-Set	**77%**	**564**	**81%**

- Barbara Poblete, Ricardo Baeza-Yates, Query-sets: using implicit feedback and query patterns to organize web documents. WWW'08.

Browse Logs and Search Trails

- Browse logs may contain more information than search log
  - Search trails record other browsing activities in addition to queries
- Mikhail Bilenko, Ryen W. White. Mining the search trails of surfing crowds: identifying relevant websites from user activity WWW'08.



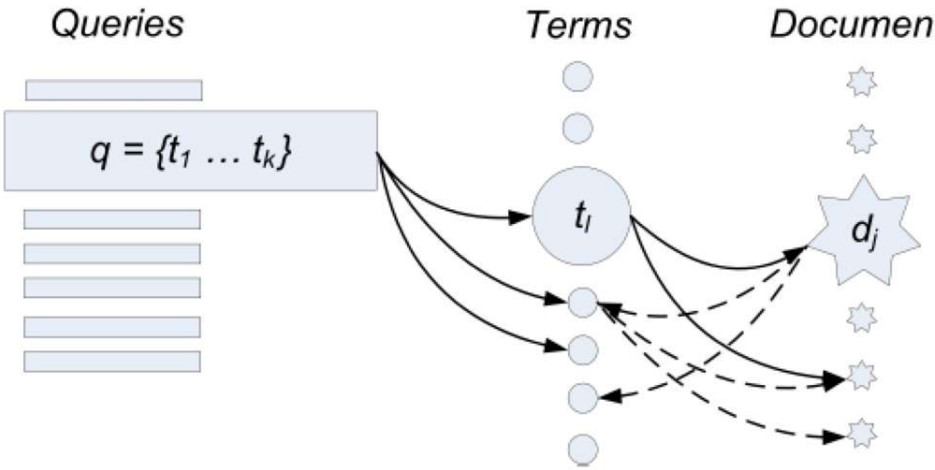
A Generative Model

- [Mikhail Bilenko, Ryen W. White. Mining the search trails of surfing crowds: identifying relevant websites from user activity WWW'08.]
- A set of search trails $D = \{q \rightarrow (d_1, \ldots, d_m)\}$, where $d_1, \ldots, d_m$ are documents
- Assuming every query q instantiates a multinomial distribution over its terms

$$\mathrm{Rel}_{\mathrm{P}}(d,\hat{q}) = p(d \mid \hat{q}) = \sum_{\hat{t} \in q} p(\hat{t} \mid \hat{q}) p(d \mid \hat{t})$$

A Random Walk Extension

- The probability of reaching a document starting from a given query is the likelihood of hitting the document node via the two-step random walk that originates at the query node and proceeds via the term nodes



- [Mikhail Bilenko, Ryen W. White. Mining the search trails of surfing crowds: identifying relevant websites from user activity WWW'08.]

Today’s Agenda

- Motivation
- Enriched models using log data
- **Tackling sparsity**
- Application examples

Tackling Query Sparsity

- Many queries are rarely asked
- Idea: clustering similar queries to identify groups of user information needs of significant sizes → reliable annotations on Web pages clicked
- A two phase algorithm
  - Preprocessing phase
  - Online searching phase
- [Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query Clustering for Boosting Web Page Ranking AWIC, 2004.]

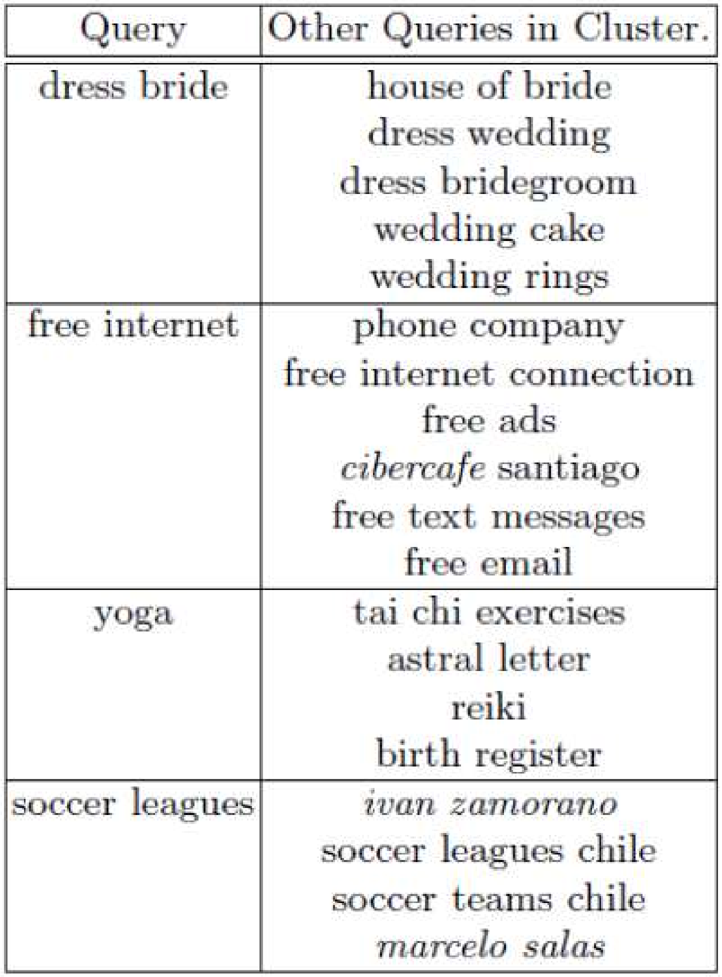
Preprocessing Phase

- At periodical and regular intervals
- Extract queries and clicked URLs from the Web log, and cluster them using the text of all the clicked URLs (by k-means)
- For each cluster C_i, compute and store
  - A list Q_i containing queries in the cluster
  - A list U_i containing the k-most popular URLs along with their popularity
- [Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query Clustering for Boosting Web Page Ranking AWIC, 2004.]

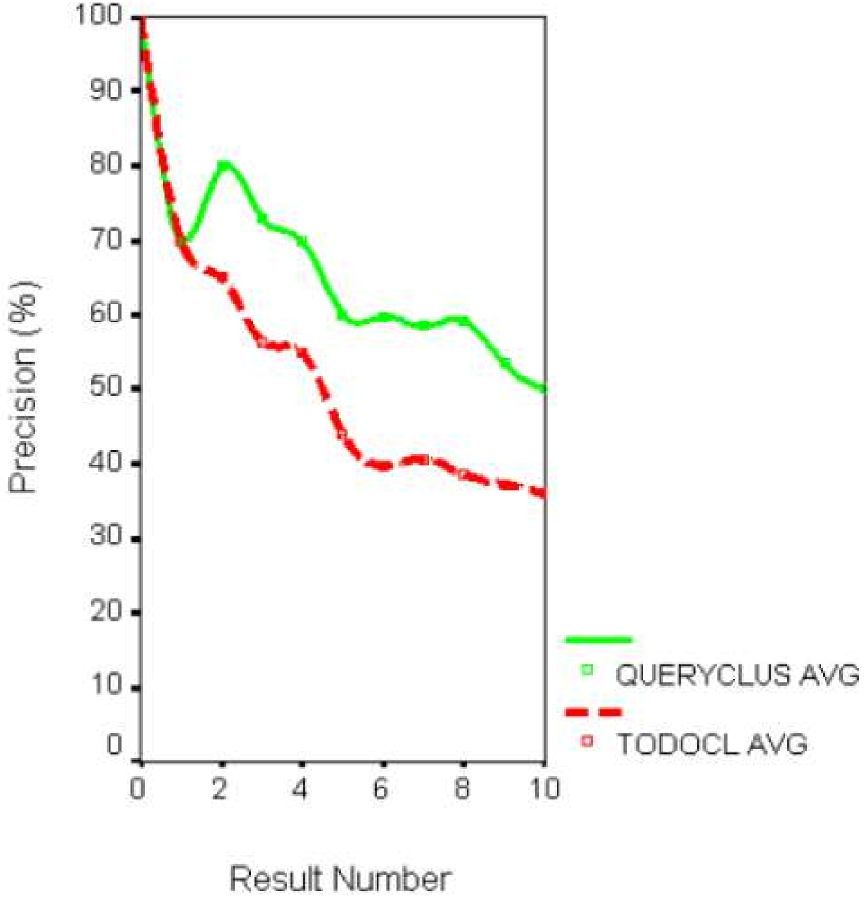
Online Searching Phase

- Input: a query q
- If q appears in the stored clusters, find the corresponding cluster C_i containing q, use U_i to boost the search engine ranking algorithm by
  - $NewRank(u) = \beta \times OrigRank(u) + (1 - \beta) \times Rank(u)$
- [Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query Clustering for Boosting Web Page Ranking AWIC, 2004.]

Examples & Effectiveness



Query	Other Queries in Cluster.
dress bride	house of bride dress wedding dress bridegroom wedding cake wedding rings
free internet	phone company free internet connection free ads *cibercafe* santiago free text messages free email
yoga	tai chi exercises astral letter reiki birth register
soccer leagues	*ivan zamorano* soccer leagues chile soccer teams chile *marcelo salas*

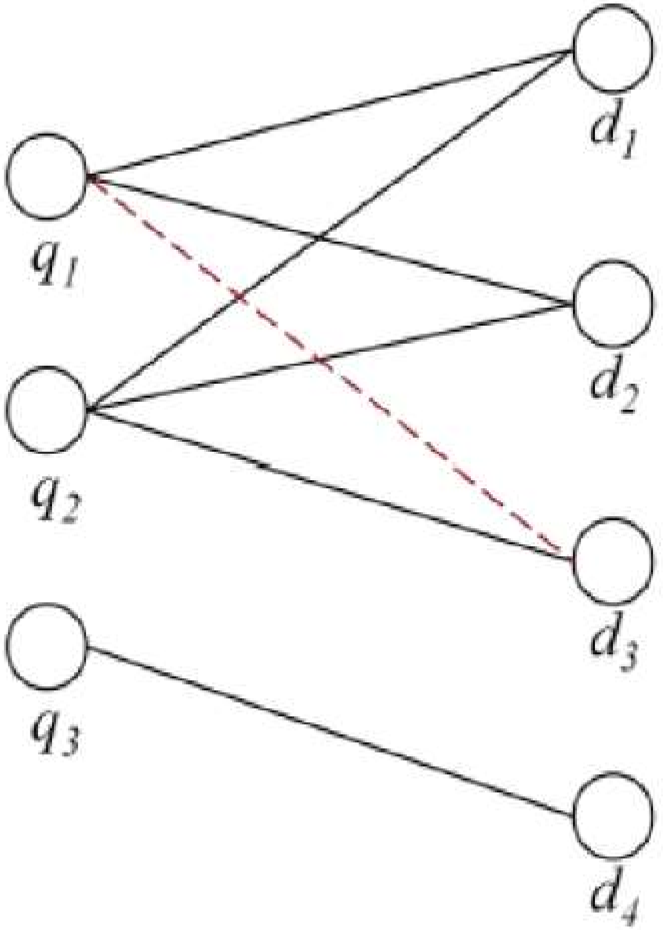


- Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query Clustering for Boosting Web Page Ranking AWIC, 2004.

Documents Not Clicked

- Many documents may have very few or even no clicks
  - 75% of a sample of 2.62 million Web pages do not have any click in a real case study
- Idea: use smoothing techniques
  - Random walk
  - Discounting
- [Gao, J., et al. Smoothing clickthrough data for web search ranking. SIGIR'09.]

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 expansion, document d_3 has a clickthrough stream of q_2 only; after a random walk expansion, the click-through stream is augmented with query q_1, which has a similar click pattern as q_2

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

Good-Turing Estimator

- Let N be the size of a sample text, n_r be the number of words which occur in the text exactly r times $N = \sum_r r n_r$
- Estimate P_{GT} for a probability of a word that occurred in the sample r times as $P_{GT} = \frac{r^*}{N}$, where $r^* = \frac{(r+1)n_{r+1}}{n_r}$
- Heuristic: not discounting high values of counts, i.e., for r > k (typically k = 5), r* = r

Discounting

- Applying Good-Turing estimate on raw clickthrough data does not work – all not-clicked words take the same free ride
  - Those features are meaningless
- Idea: discounting the clickthrough feature values
  - Details in [Gao, J., et al. Smoothing clickthrough data for web search ranking. SIGIR'09.]
- The discounting method works very well in the empirical studies

Today's Agenda

- Motivation
- Enriched models using log data
- Tackling sparsity
- **Application examples**

Using Logs and Query Annotations

- Generating keyword
- Learning document importance
- Organizing search results
- …

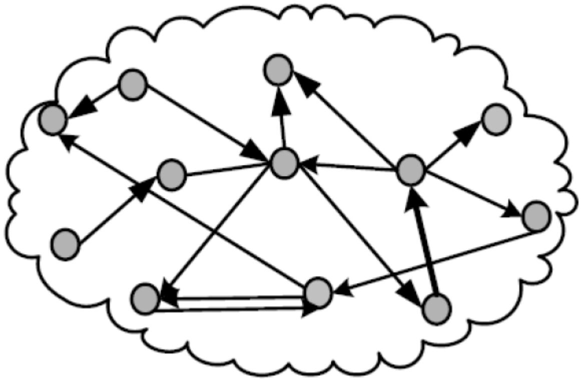
Keyword Generation

- What are the keywords that describe the concept “shoes”?
  - Starting point: shoes.com is about shoes
  - A user asked “running shoes” and clicked shoes.com “running shoes” is about shoes
  - The user also clicked runningshoes.com runningshoes.com is also about shoes
  - Queries “reebok shoes” and “rebok shoes” led to clicks on runningshoes.com those keywords are also about shoes
- Given a concept (e.g., shoes), a set of elements representing the concept (e.g., a set of URLs), and the relationship between the documents and the queries, find a set of keywords capturing the concept best
- [Ariel Fuxman, Panayiotis Tsaparas, Kannan Achan, Rakesh Agrawal Using the wisdom of the crowds for keyword generation WWW'08.]

A Semi-Supervised Version

- Input
  - A set of labeled objects about a concept (e.g., URLs)
  - A set of unlabeled objects (the remaining URLs and the queries in the log)
  - A set of constraints between labeled and unlabeled objects (the click log)
- Task: label some of the unlabeled elements in a meaningful way
- Idea: use Markov random fields to model the query click graph
- [Ariel Fuxman, Panayiotis Tsaparas, Kannan Achan, Rakesh Agrawal Using the wisdom of the crowds for keyword generation WWW'08.]

Modeling User Browsing Behavior



- User browsing graph
  - Vertices representing pages
  - Directed edges representing transitions between pages in browsing history
  - Lengths of staying time are included
- Using the continuous-time Markov process
  - The stationary probability distribution of the process is the importance of a page
- [Yuting Liu, Bin Gao, Tie-Yan Liu, Ying Zhang, Zhiming Ma, Shuyuan He, Hang Li. BrowseRank: letting web users vote for page importance. SIGIR'08.]

ClickRank

- A session is modeled as a logical sequence of hops through the Web graph according to the user's retrieval intension
  - Temporal attributes (e.g., dwell time) reflects user's interest on a page
- For a session s, the local ClickRank defines a random variable associated with all pages on the Web graph reflecting how important a page is to the user's retrieval intension in this session
- [Zhu, G, Mishne, G. Mining rich session context to improve web search. KDD'09.]

Organizing Search Results

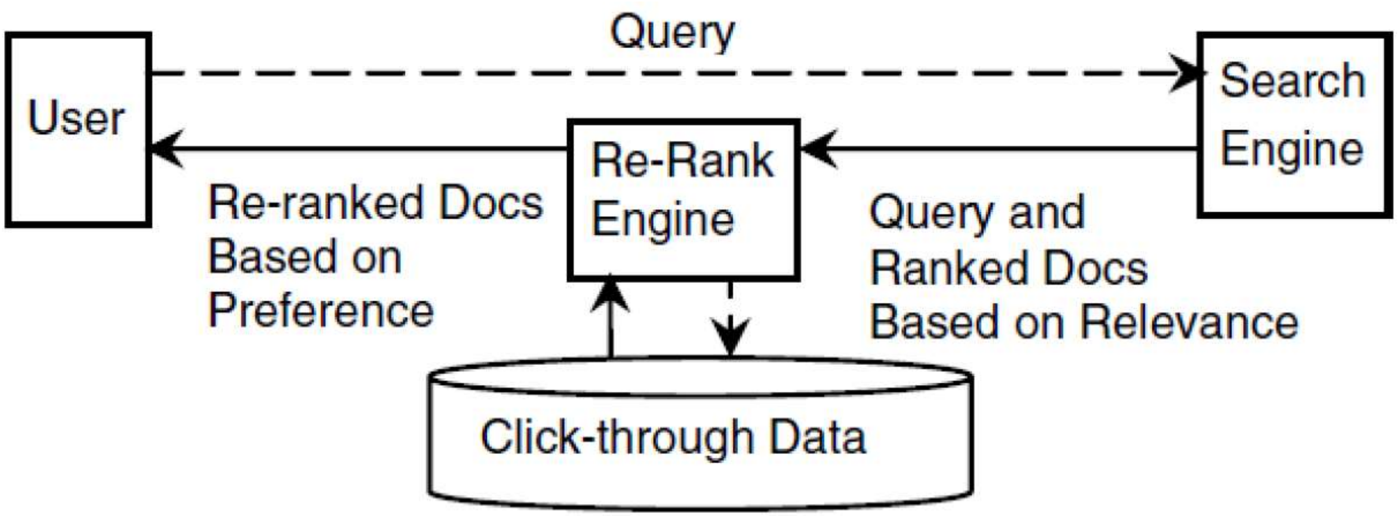
- Query “jaguar” is ambiguous: car, animal, software, or a sport team?
  - Instead of presenting a mixed list of results, a user may prefer clusters of results according to the senses
- Challenges in clustering results
  - Clusters may not necessarily correspond to the interesting aspects of a topic from the user’s perspective
  - Cluster labels may not be informative
- [X. Wang and C. Zhai. Learn from web search logs to organize search results. SIGIR'07.]

Clustering Using Search Logs

- What kinds of pages viewed by users in the results of a query?
  - Finding aspects interesting to users by mining user clickthrough data
- Generate meaningful cluster labels using query words entered by users
- [X. Wang and C. Zhai. Learn from web search logs to organize search results. SIGIR'07.]

Using Search Logs in Re-Ranking

- Search engine → candidate answers and baseline ranking
- Click-through data → learning user preference for re-ranking


Query
User
Search Engine
Re-Rank Engine
Re-ranked Docs Based on Preference
Query and Ranked Docs Based on Relevance
Click-through Data

Min Zhao, Hang Li, Adwait Ratnaparkhi, Hsiao-Wuen Hon, and Jue Wang. Adapting document ranking to users' preferences using click-through data, AIRS’06.

More Examples …

- Considering clickthrough data in page summarization
- Category maintenance
- …
- [J.-T. Sun, D. Shen, H.-J. Zeng, Q. Yang, Y. Lu, and Z. Chen, Web-page summarization using clickthrough data, SIGIR '05.]
- [A. Cid, C. Hurtado, and M. Mendoza, Automatic maintenance of web directories using click-through data, in ICDEW '06.]

Challenges

- From document modeling to document cluster/site modeling
  - Several Web pages are often visited together
- Modeling temporal characteristics of search activities
  - Detecting bursts of new interests
- Many applications can be improved by using search/browse log data

Take-away Messages

- Search logs and browse logs can be used to improve document search
  - Verified by user studies
- Enriched models of documents considering log data
  - Central idea: using query terms and segments as features
- Tackling sparsity of log data
  - Clustering similar queries
  - Smoothing
- Many applications: generating keywords, computing importance of documents, organizing search results, considering clickthrough data in page summarization, and category maintenance

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.

Preview of Lecture 24: Query-Document Matching by Log Mining

- Learning user preferences from logs
- Modeling and predicting clicks

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!