MLA’08 Nanjing
Nov. 7, 2008

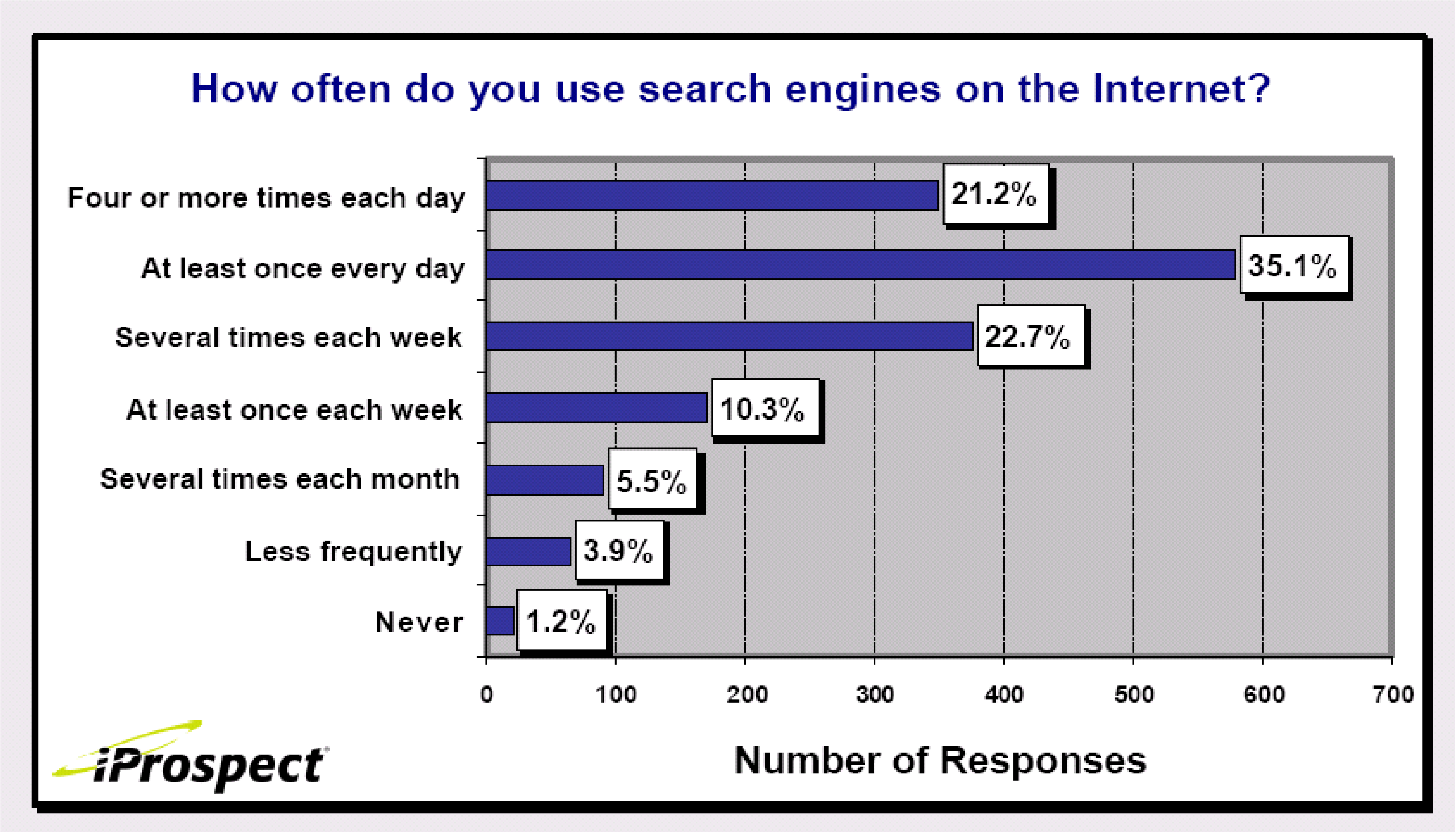
Machine Learning Meets Web Search

Hang Li
Microsoft Research Asia

Web Search is Part of Our Life

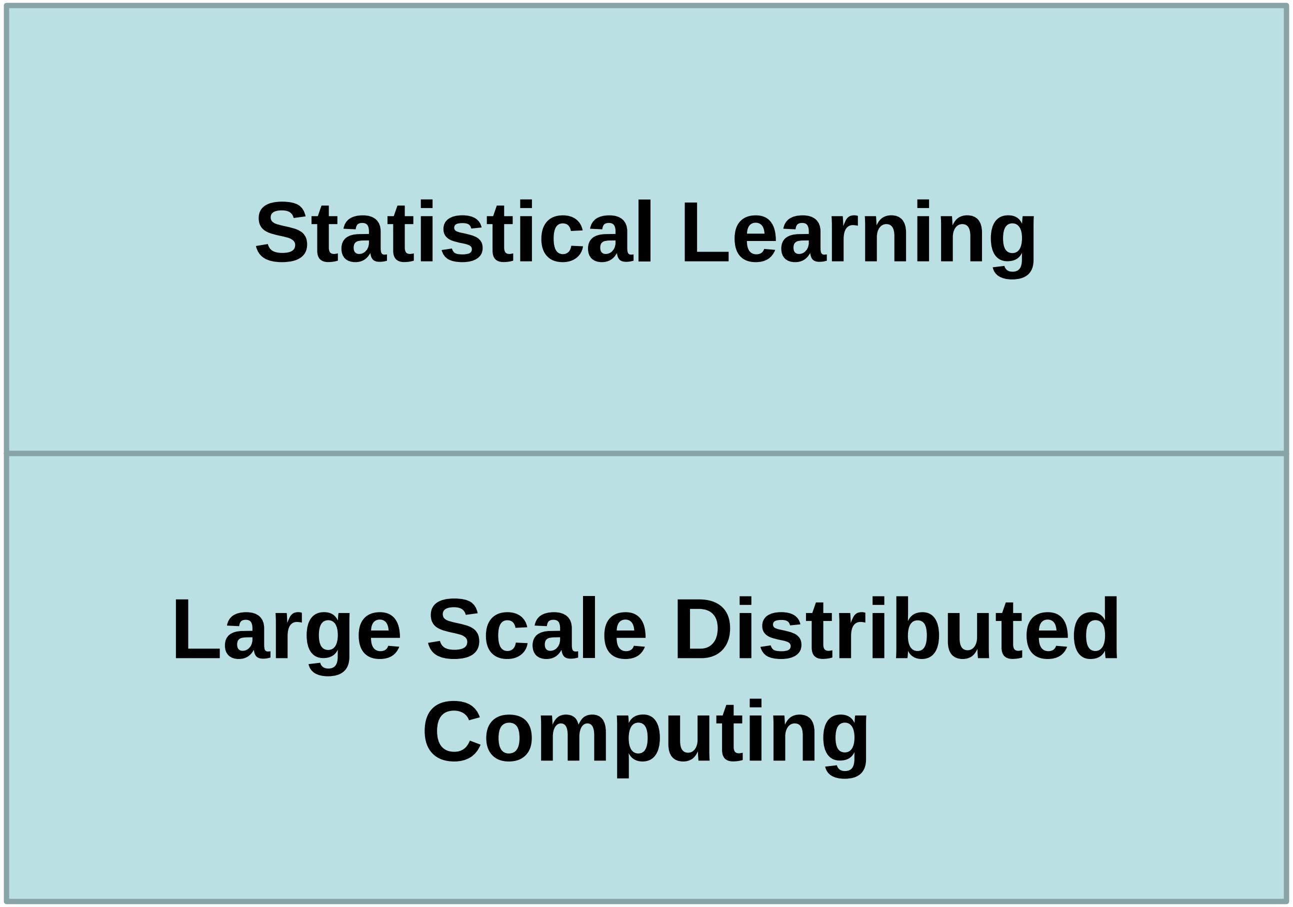
Web Users Heavily Rely on Search Engines

http://www.iprospect.com/premiumPDFs/iProspectSurveyComplete.pdf



Physically Search System is Data Center

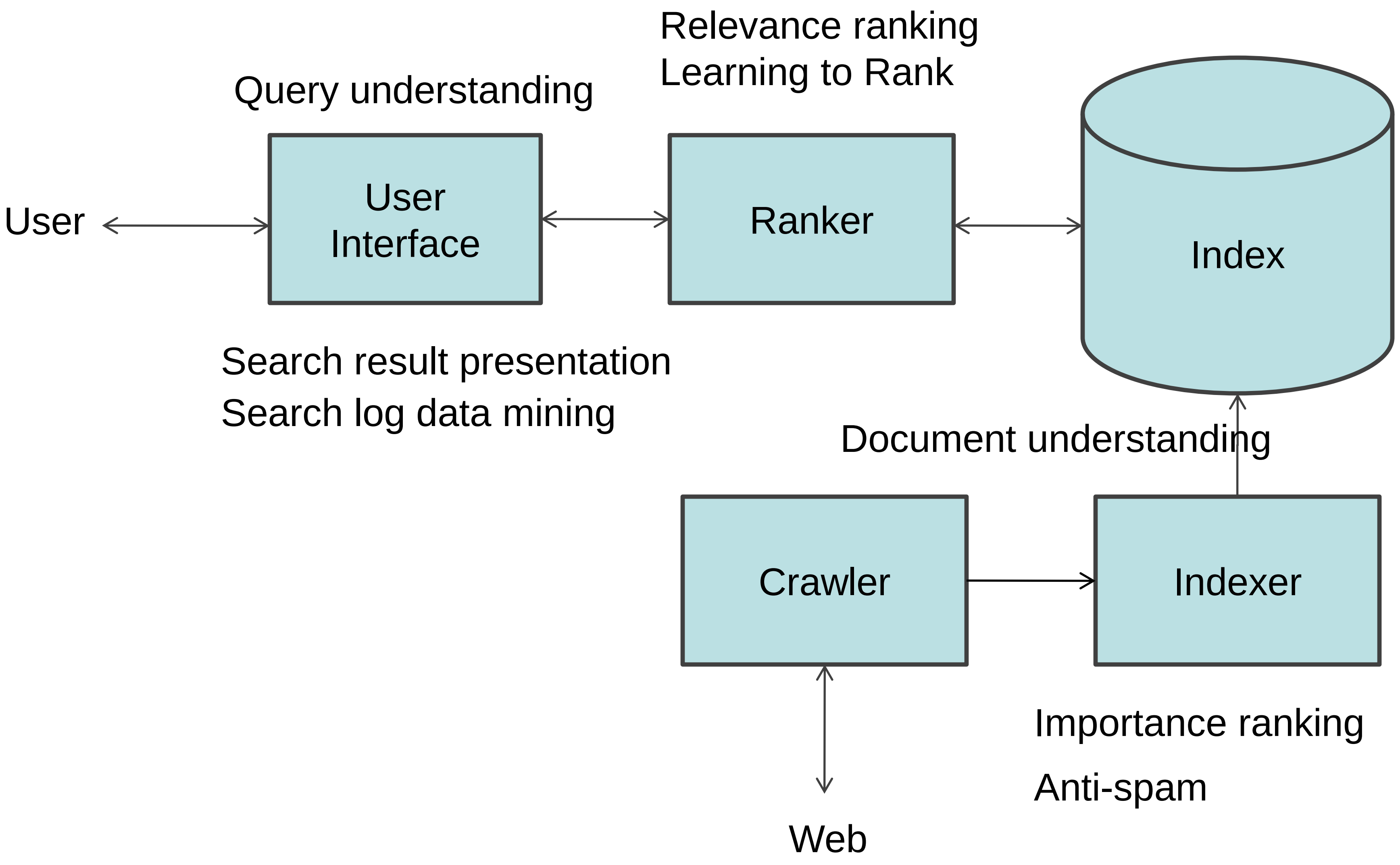
Advanced Web Search Technologies Are Used…



Statistical Learning

Large Scale Distributed Computing

Overview on Web Search System



Relevance ranking
Learning to Rank

Query understanding

User

User Interface

Ranker

Index

Search result presentation
Search log data mining

Document understanding

Crawler

Indexer

Importance ranking

Anti-spam

Web

Component Technologies for Web Search

- Relevance Ranking
- Importance Ranking
- Document Understanding
- Query Understanding
- Crawling
- Indexing
- Search Result Presentation
- Anti-Spam
- Learning to Rank
- Search Log Data Mining
- Evaluation and User Study

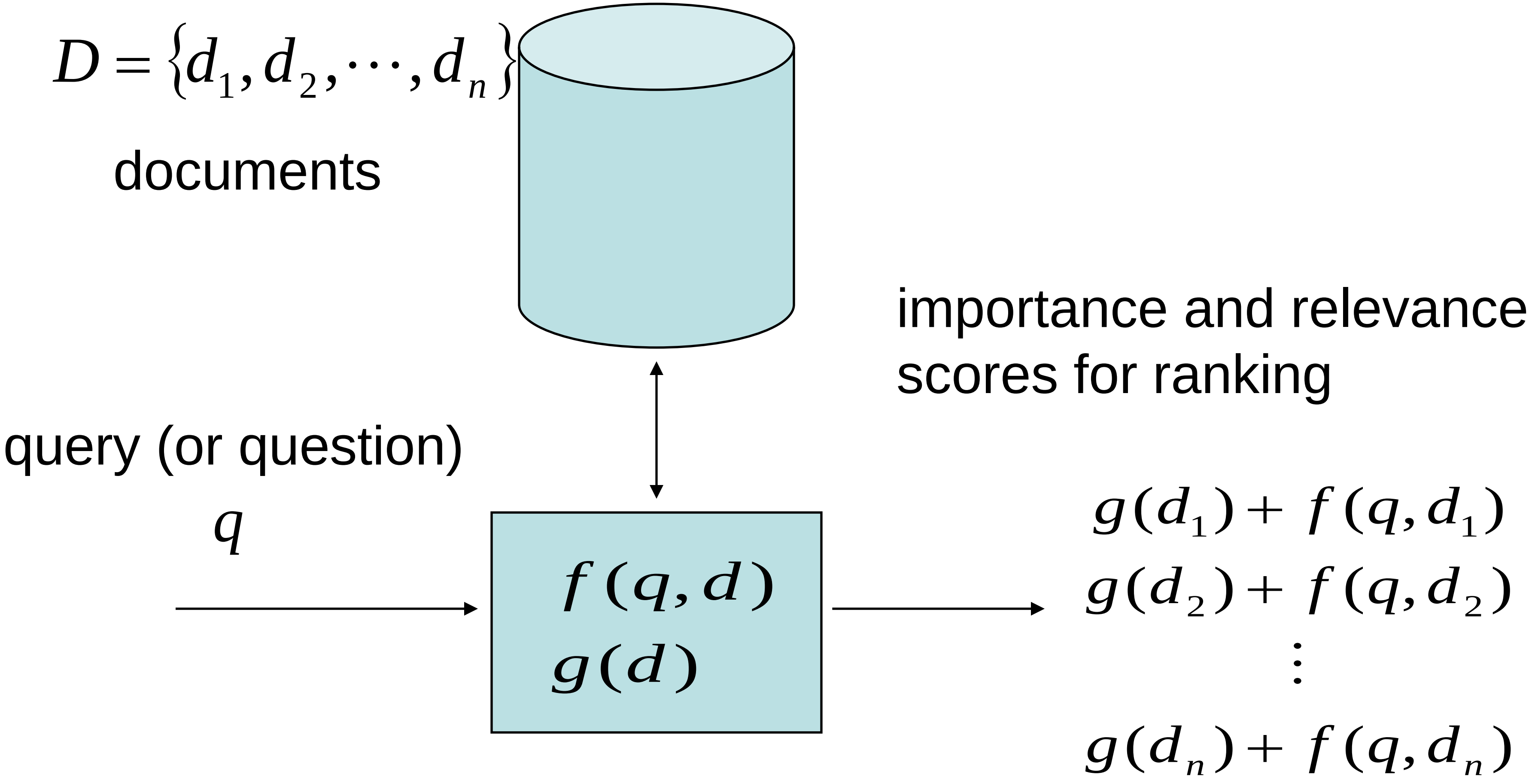
Statistical Learning Plays Key Role!

Talk Outline

- Statistical Learning is Important for Web Search
- Statistical Learning in Web Search
  - Importance Ranking: BrowseRank
  - Anti-Spam: Temporal Classifier
  - Query Understanding: CRF for Query Reformulation
  - Web Page Understanding: HyperText Topic Model
  - Result Presentation: Context Aware Query Suggestion
  - Learning to Rank: Listwise Approach and Global Ranking

Importance Ranking: BrowseRank (SIGIR 2008 Best Student Paper)

General Model for Importance Ranking



$D = \{d_1, d_2, \cdots, d_n\}$

documents

importance and relevance scores for ranking

query (or question)

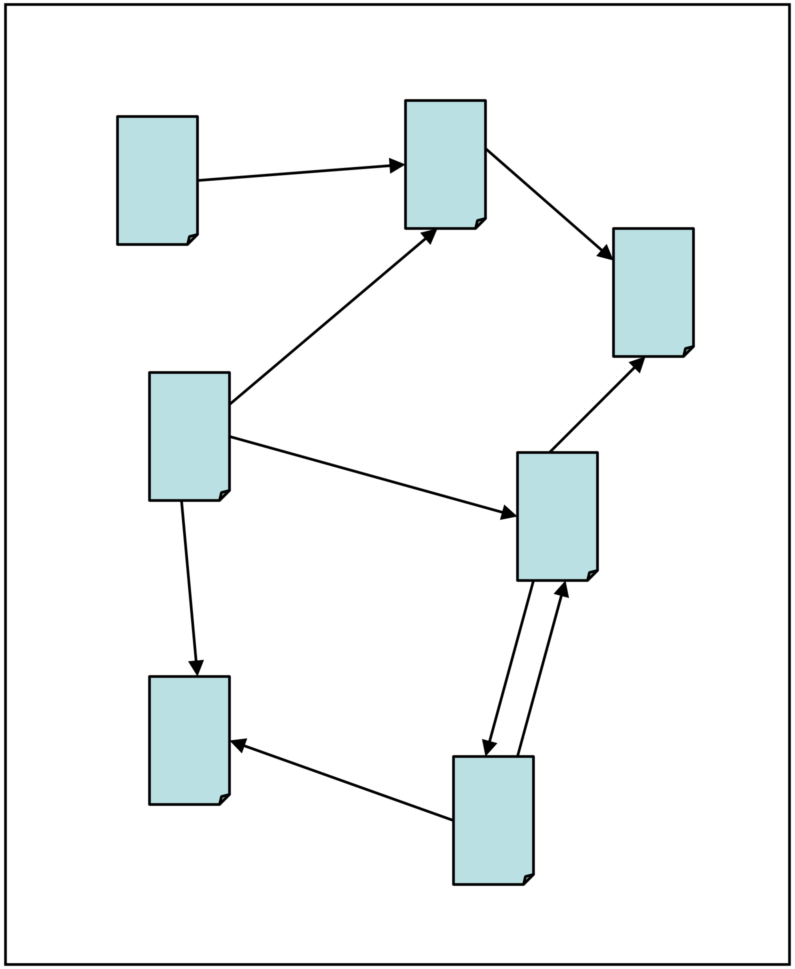
q

$f(q, d)$

$g(d)$

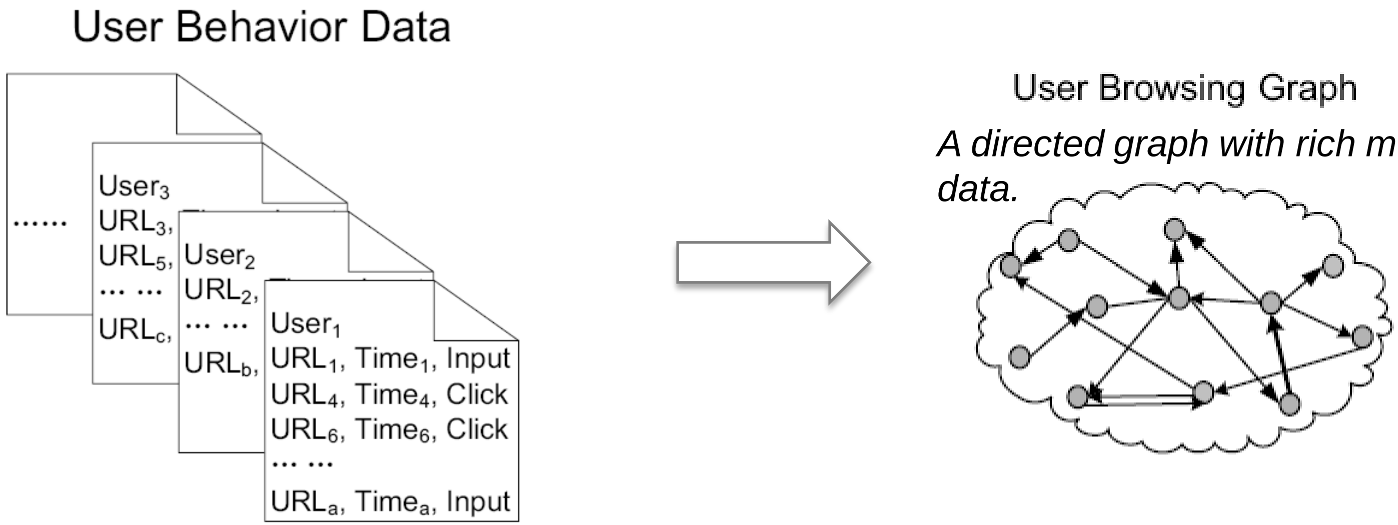
$$g(d_1) + f(q, d_1)$$
$$g(d_2) + f(q, d_2)$$
$$\vdots$$
$$g(d_n) + f(q, d_n)$$

Page Rank



$$P(d_i) = \alpha \sum_{d_j \in M(d_i)} \frac{P(d_j)}{L(d_j)} + (1-\alpha)\frac{1}{n}$$

Building User Browsing Graph



A directed graph with rich meta data.

Vertex: Web page

Edge: Transition

Edge weight w_{ij}: Number of transitions

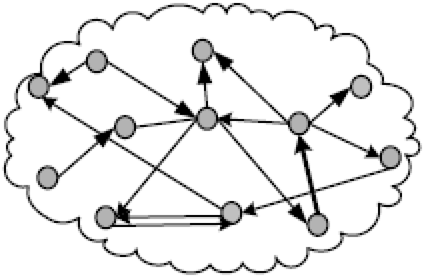
Staying time T_i: Time spend on page *i*

Vertex weight C_i: Number of visit for page *i*

Reset probability σ_i Normalized frequencies as first page of session

BrowseRank: Continuous-time Markov Model

User Browsing Graph



⇒ $P(t)$ ⇒ $\boxed{\pi = \pi P(t)}$

Hard

⇓

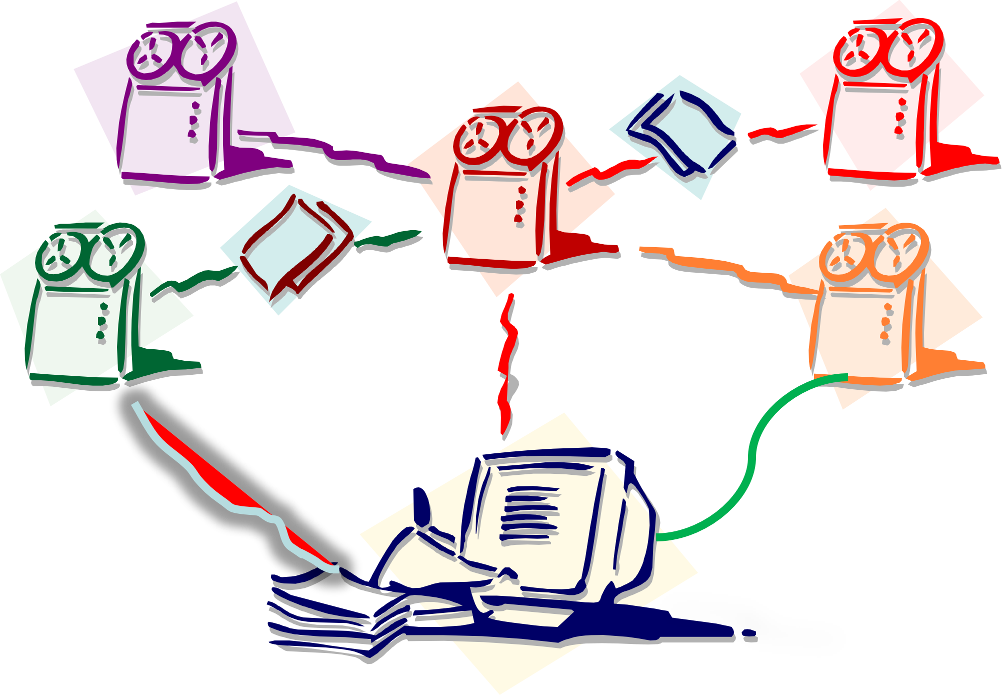
Calculating π = Estimating staying time distribution of each state + Computing the stationary distribution $\tilde{\pi}$ of a discrete-time Markov chain (*called embedded Markov chain*)

Anti-Spam:
Temporal Classifier
(ICDM 2006)

Anti-Spam

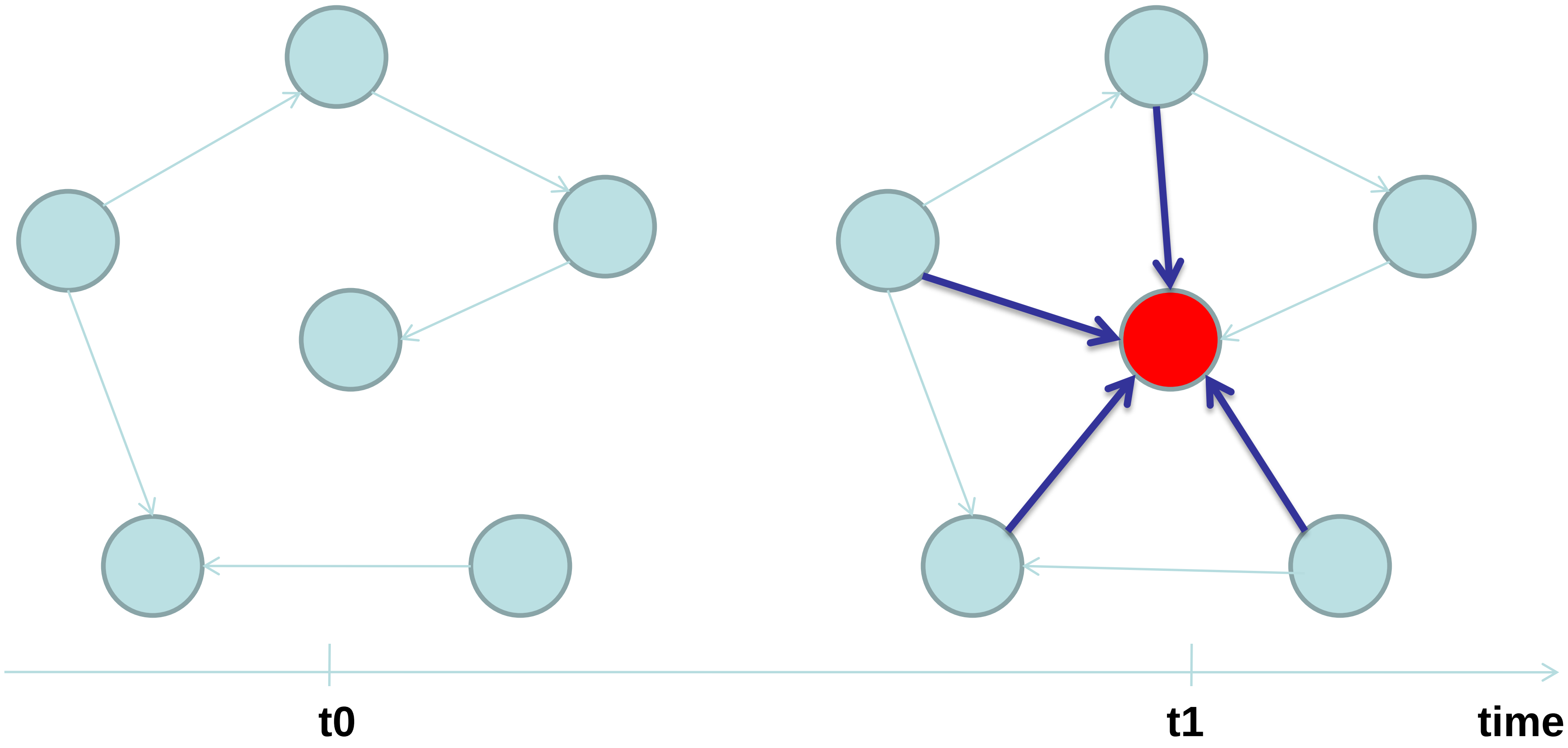
- Spam
  - Manipulate relevance and importance
  - Not ethical, if to be ranked higher beyond real value
  - “Cheating” search engines
- Spam Type
  - Content Spam
  - Link Spam
  - Cloaking

Link Spam



- Link from Blog, Forum, etc
- Link Exchange
- Link Farm

Inlink May Increase Drastically at Spam Page



Detection Using Temporal Information

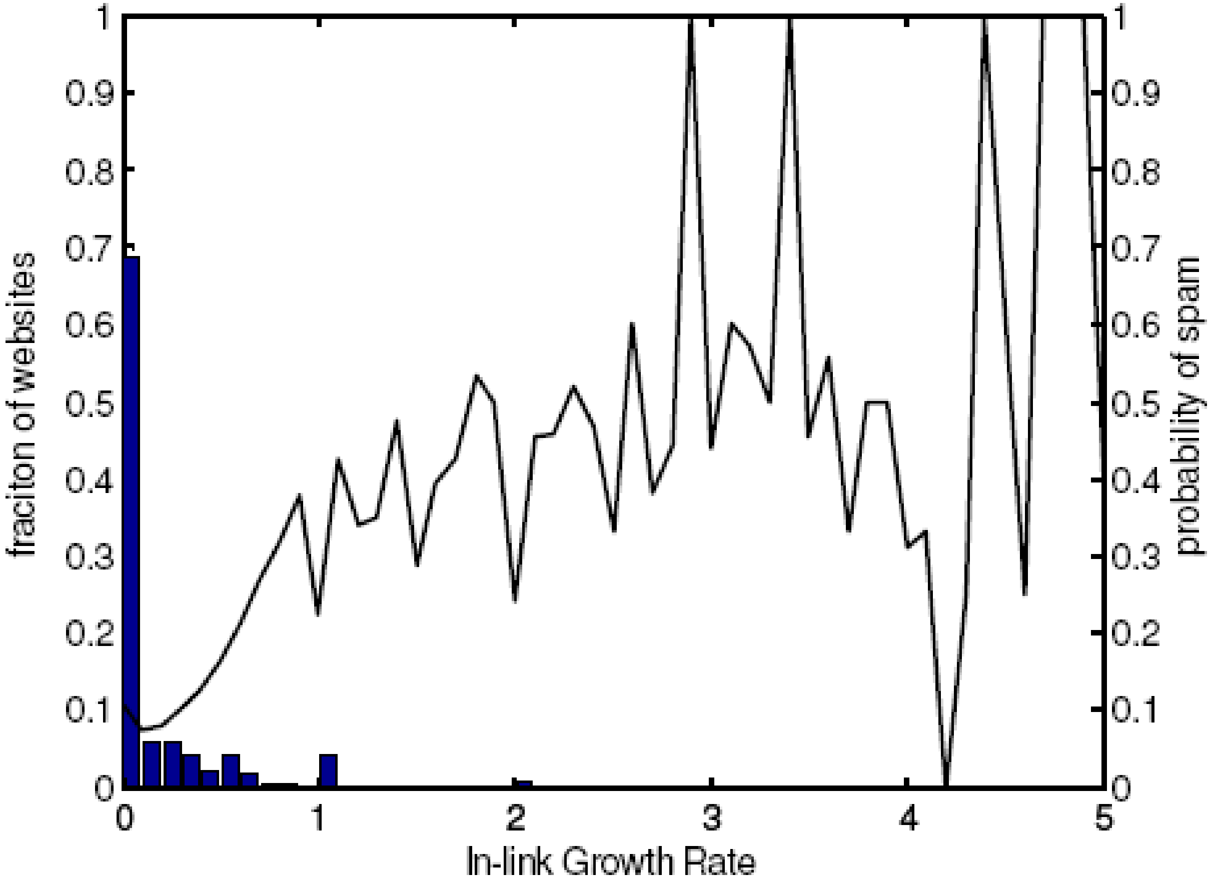


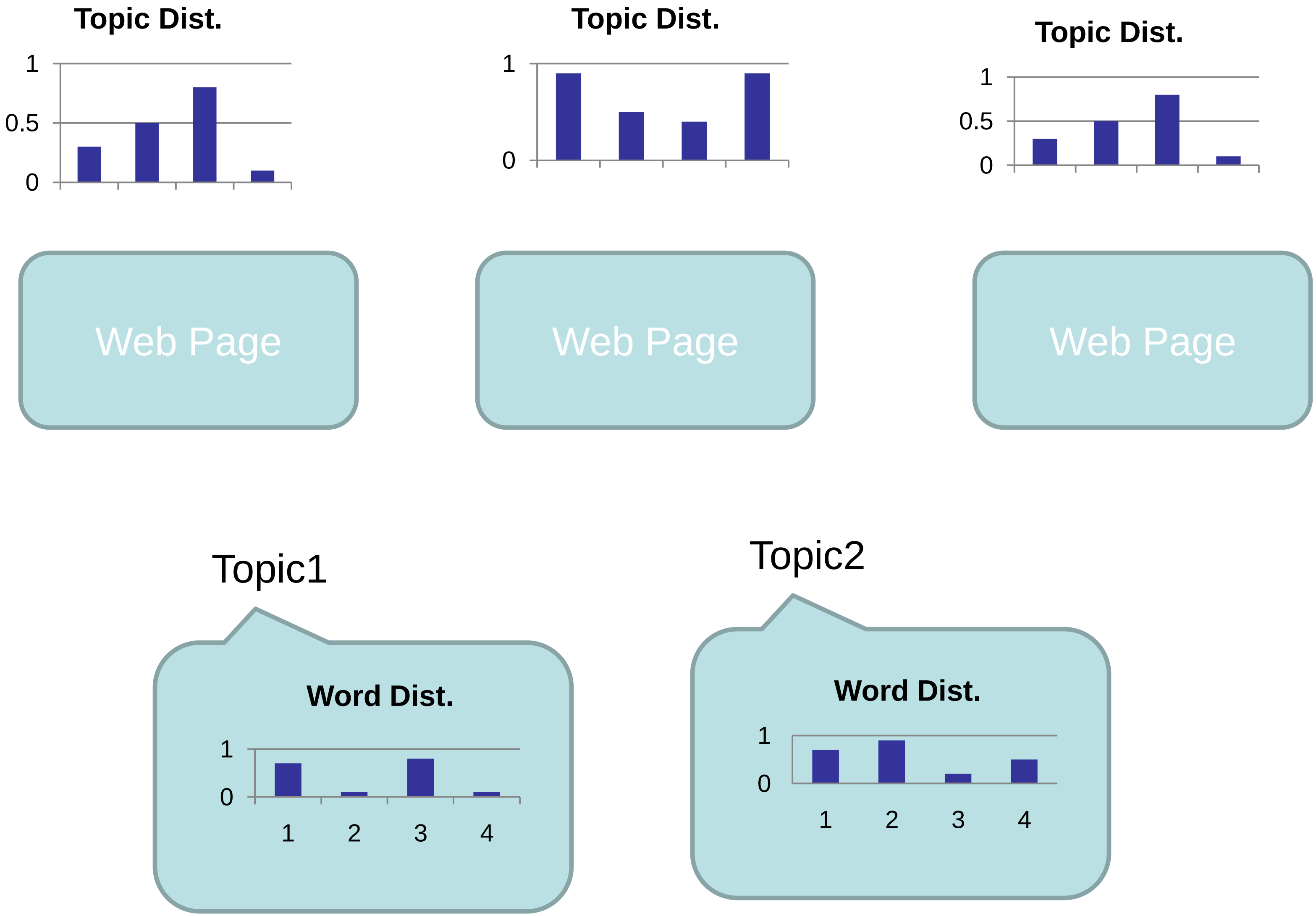
Figure 1. Probability of spam versus IGR.

Web Page Understanding: Topic Model for Hypertext (EMNLP 2008)

Web Page Understanding

- Web Information Extraction
  - Block Analysis
  - Metadata Extraction (Title, Date, etc)
  - Text Information Extraction
  - Wrapper Generation
- Web Page Classification
  - Based on Semantics
  - Based on Type (Homepage, Spec, etc)

LDA (Latent Dirichlet Allocation)



Topic Dist.

Topic Dist.

Topic Dist.

Web Page

Web Page

Web Page

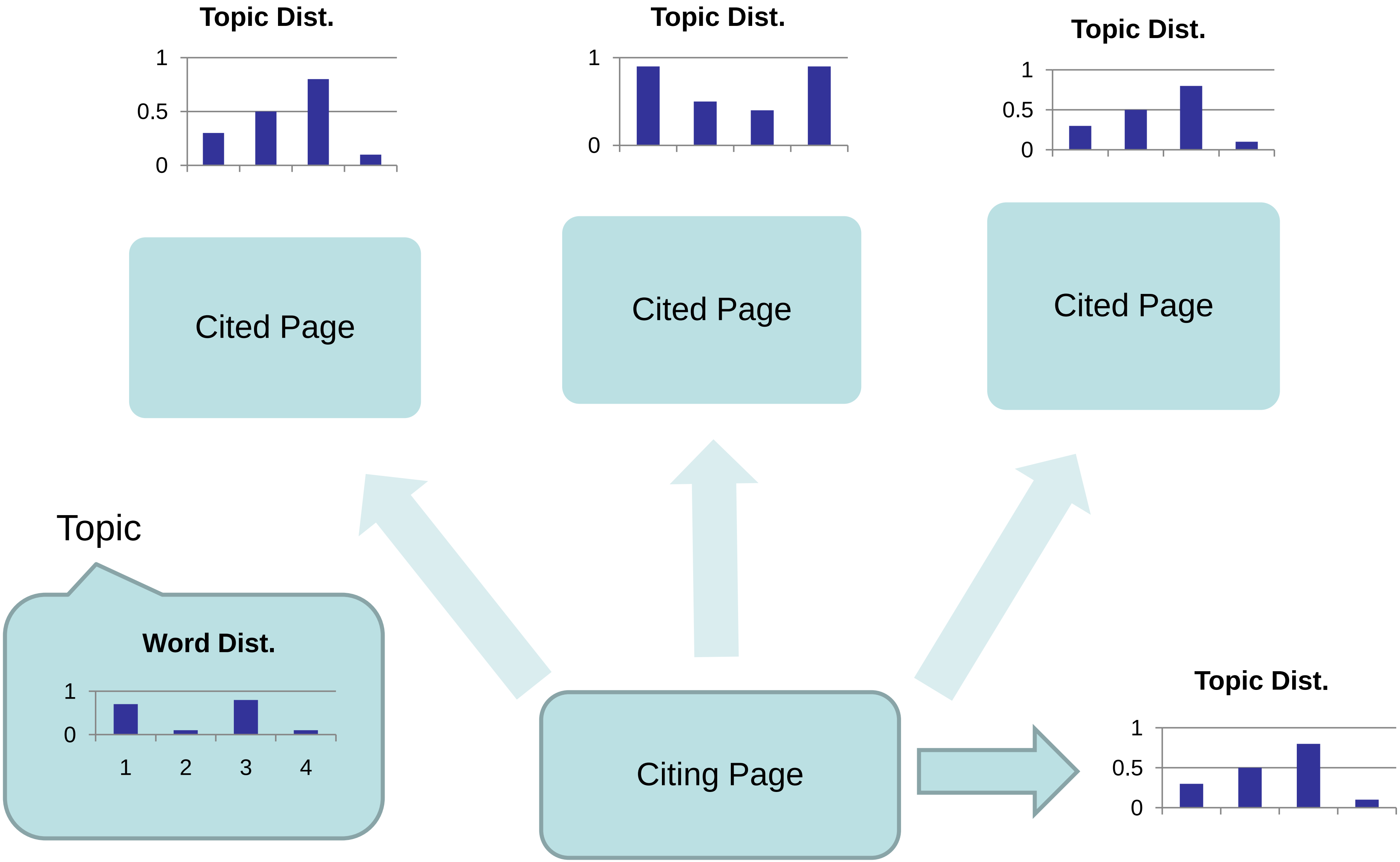
Topic1

Word Dist.

Topic2

Word Dist.

HTM: Topic Model for Hyper Text



Topic Dist.

Topic Dist.

Topic Dist.

Cited Page

Cited Page

Cited Page

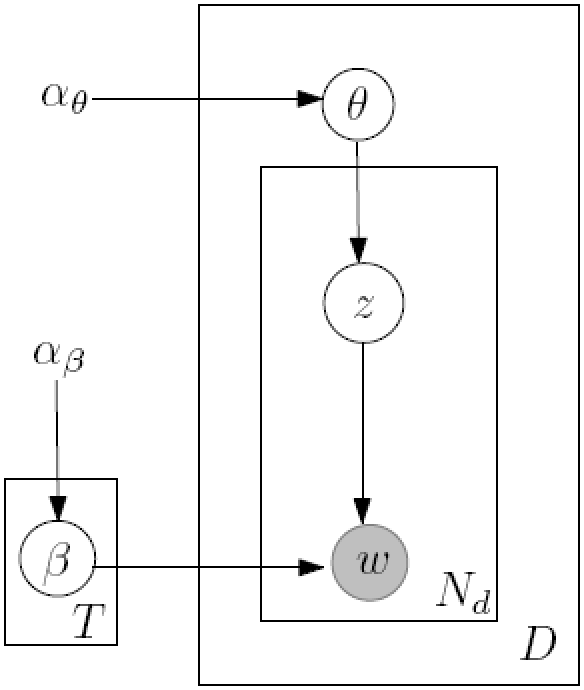
Topic

Word Dist.

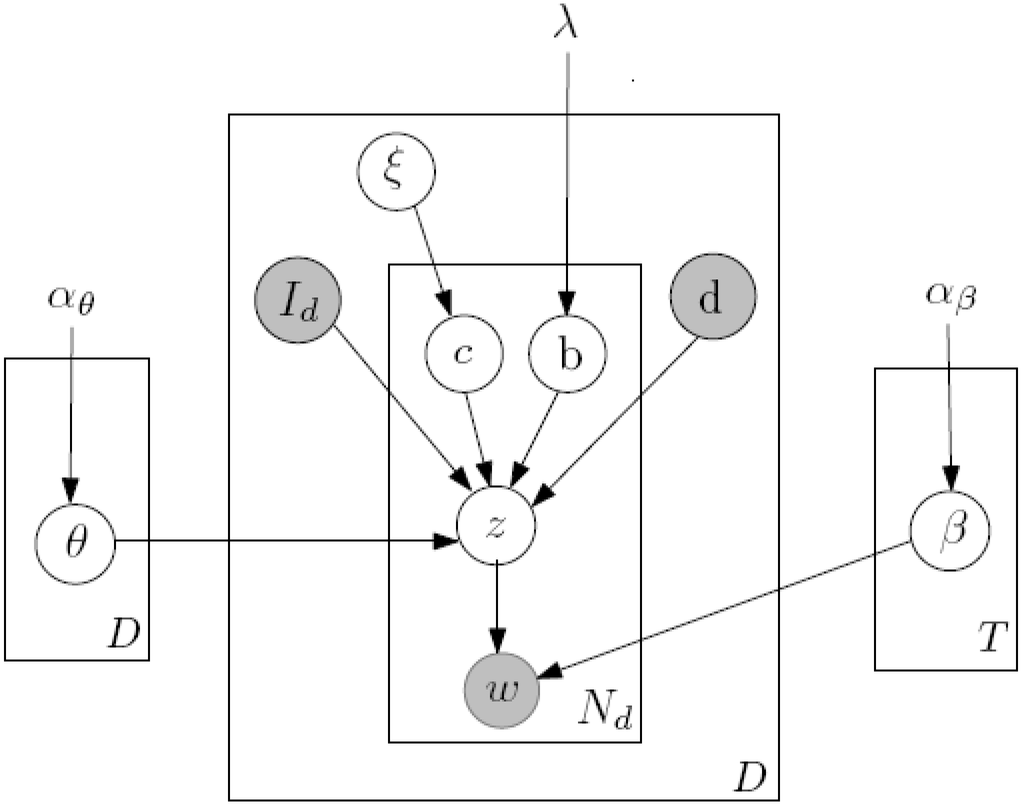
Citing Page

Topic Dist.

LDA vs HTM



(a) LDA



(b) HTM

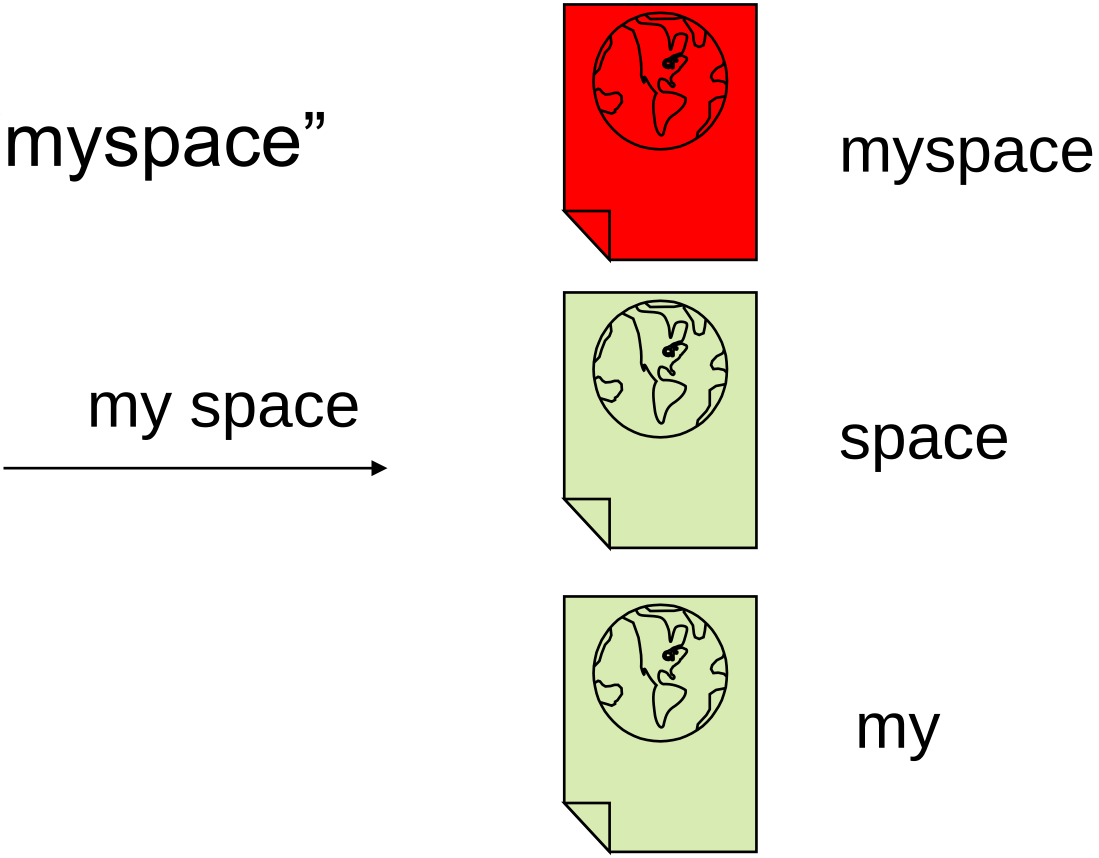
Query Understanding: Query Refinement (SIGIR 2008)

Query Understanding

- Spelling Error Correction
- Query Refinement
  - E.g., “ny times” → “new york times”
- Query Classification
  - Based on Semantics (Sport, etc)
  - Based on Type (Name Query, etc)
- Query Segmentation
  - E.g., “harry porter book” → “[harry porter] book”

Mismatching between Query Term and Document Term

I want to search “myspace”



my space

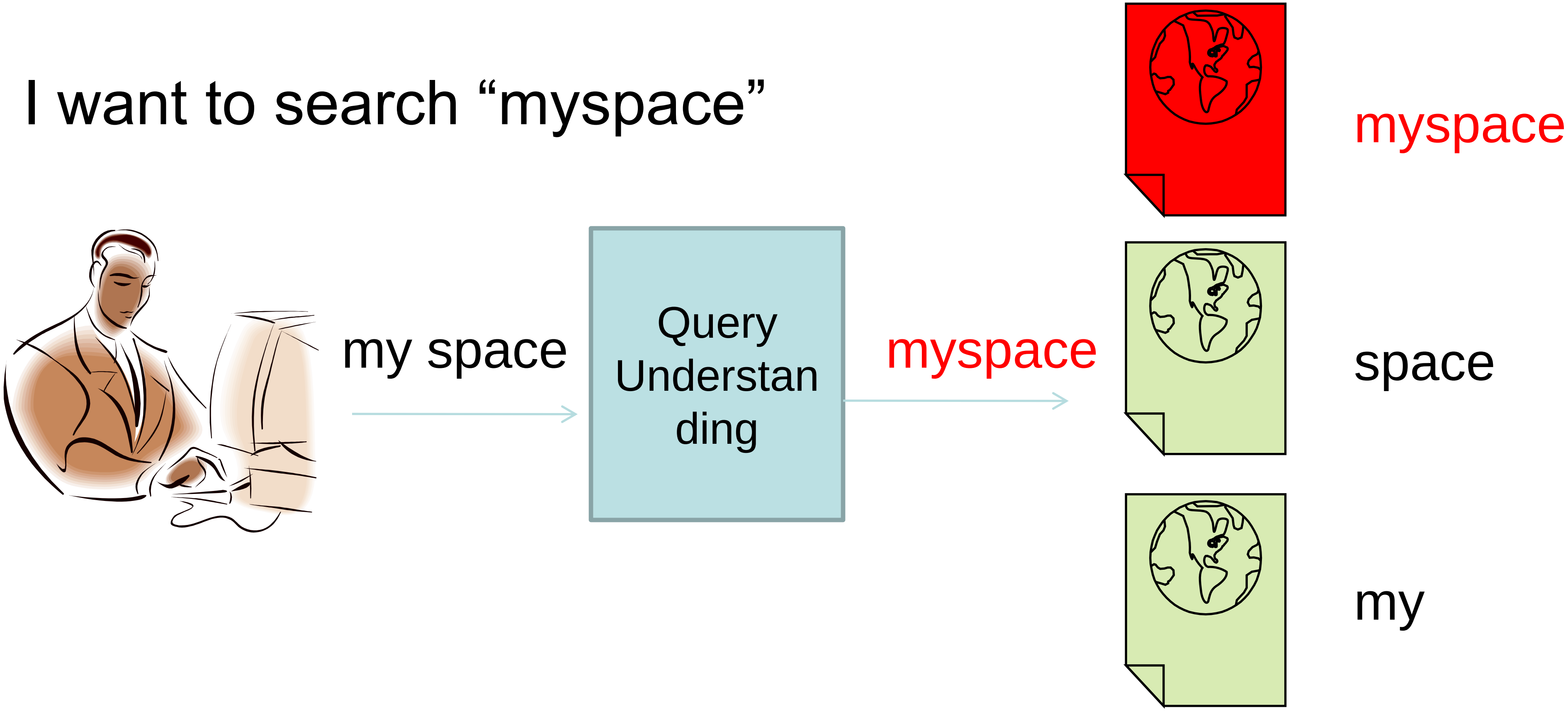
myspace

space

my

Understanding the Intent and Solving the Mismatch

I want to search “myspace”



my space

Query Understanding

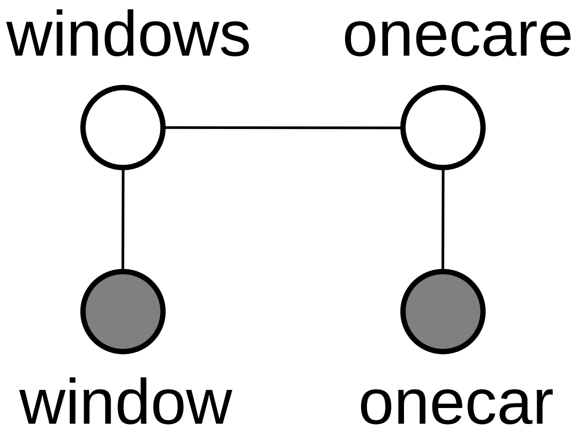
myspace

myspace

space

my

Structured Prediction Problem



“Ideal” word sequence

Observed “noisy” word sequence

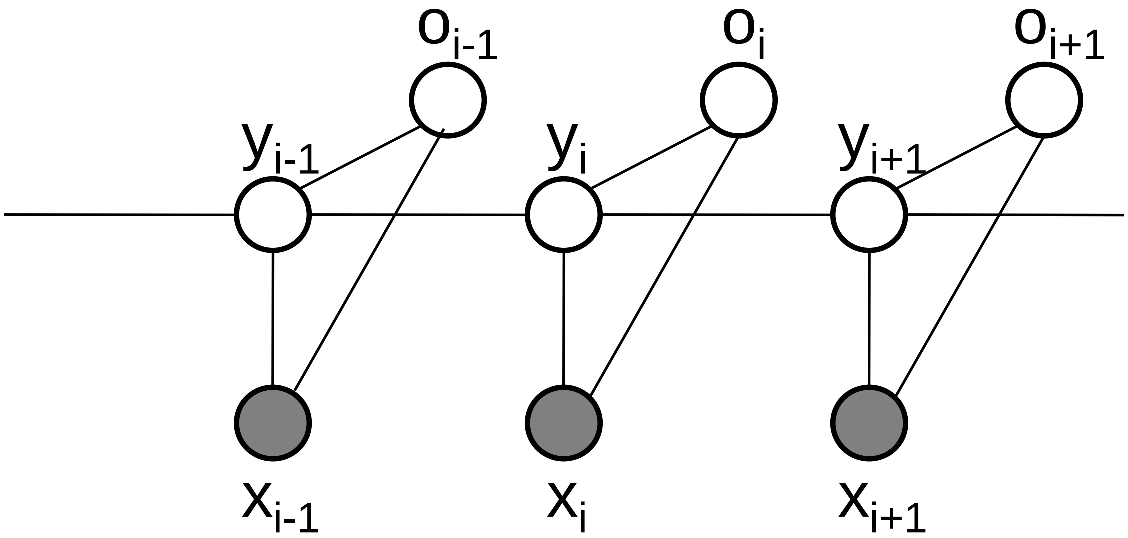
$$y^* = \arg\max_y \Pr(y|x)$$

“ideal” query word sequence

original query word sequence

Conditional Random Fields for Query Refinement

Introducing Refinement Operations



$$\Pr(y, o|x) = \frac{1}{Z(x)} \prod_{i=1}^{n} \phi(y_{i-1}, y_i)\phi(y_i, o_i, x)$$

Operations

Spelling: insertion, deletion, substitution, transposition, …

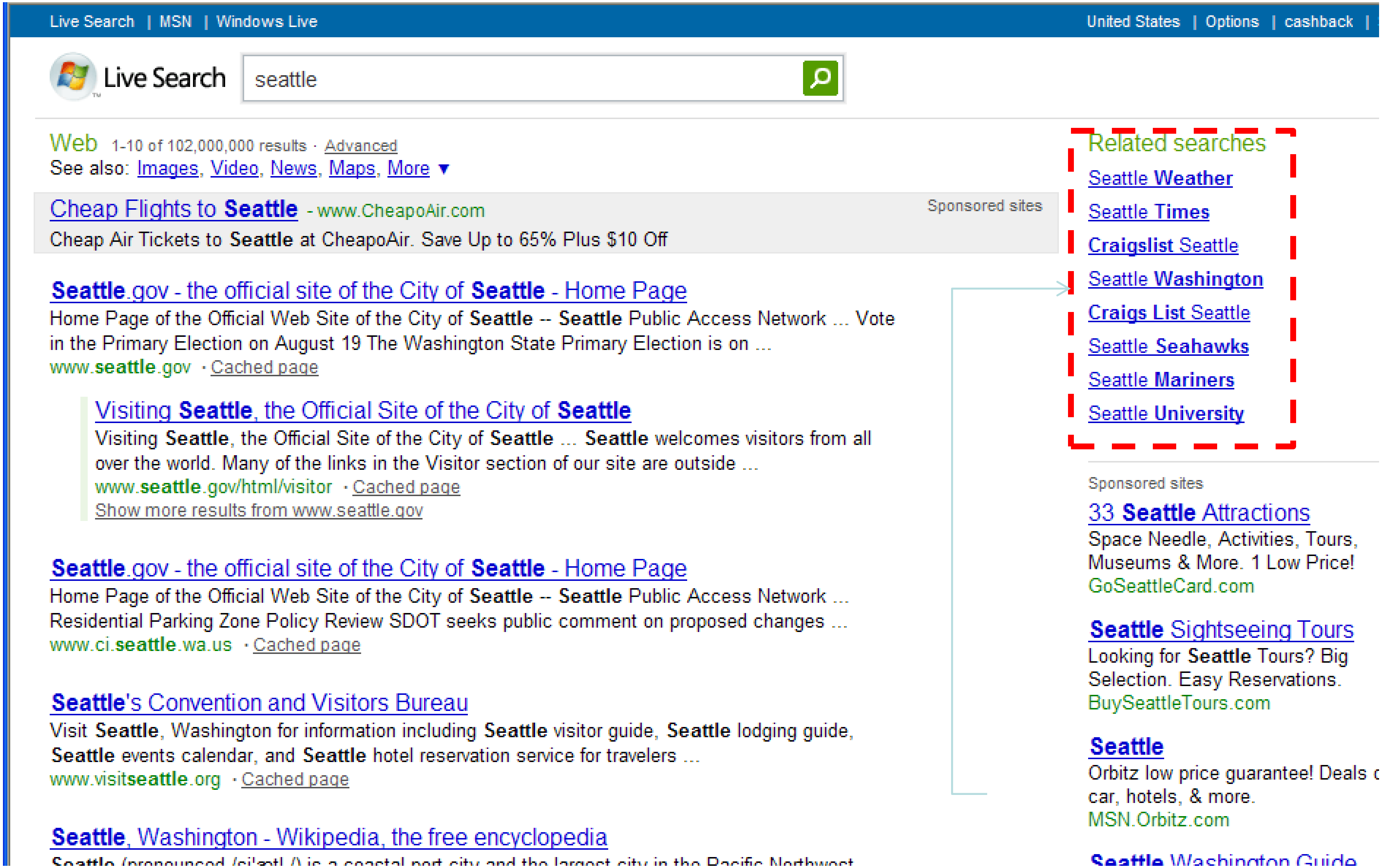
Word Stemming: +s/-s, +es/-es, +ed/-ed, +ing/-ing, …

Result Presentation:
Query Suggestion
(KDD 2008 Best Application Paper)

Result Presentation

- Snippet Generation
- Query Suggestion
- Result Clustering and Classification

Query Suggestion


Live Search | MSN | Windows Live
United States | Options | cashback |
Live Search
seattle

Web 1-10 of 102,000,000 results · Advanced

See also: Images, Video, News, Maps, More ▼

Cheap Flights to **Seattle** - www.CheapoAir.com Sponsored sites

Cheap Air Tickets to **Seattle** at CheapoAir. Save Up to 65% Plus $10 Off

Seattle.gov - the official site of the City of **Seattle** - Home Page

Home Page of the Official Web Site of the City of **Seattle** -- **Seattle** Public Access Network ... Vote in the Primary Election on August 19 The Washington State Primary Election is on ...

www.**seattle**.gov · Cached page

Visiting **Seattle**, the Official Site of the City of **Seattle**

Visiting **Seattle**, the Official Site of the City of **Seattle** ... **Seattle** welcomes visitors from all over the world. Many of the links in the Visitor section of our site are outside ...

www.**seattle**.gov/html/visitor · Cached page

Show more results from www.seattle.gov

Seattle.gov - the official site of the City of **Seattle** - Home Page

Home Page of the Official Web Site of the City of **Seattle** -- **Seattle** Public Access Network ... Residential Parking Zone Policy Review SDOT seeks public comment on proposed changes ...

www.ci.**seattle**.wa.us · Cached page

Seattle's Convention and Visitors Bureau

Visit **Seattle**, Washington for information including **Seattle** visitor guide, **Seattle** lodging guide, **Seattle** events calendar, and **Seattle** hotel reservation service for travelers ...

www.visit**seattle**.org · Cached page

Seattle, Washington - Wikipedia, the free encyclopedia

Related searches

- Seattle **Weather**
- Seattle **Times**
- **Craigslist** Seattle
- Seattle **Washington**
- **Craigs List** Seattle
- Seattle **Seahawks**
- Seattle **Mariners**
- Seattle **University**

Sponsored sites

33 **Seattle** Attractions

Space Needle, Activities, Tours, Museums & More. 1 Low Price!

GoSeattleCard.com

Seattle Sightseeing Tours

Looking for **Seattle** Tours? Big Selection. Easy Reservations.

BuySeattleTours.com

Seattle

Orbitz low price guarantee! Deals car, hotels, & more.

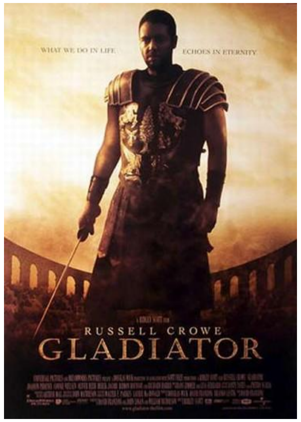
MSN.Orbitz.com

Search Intent and Context

- Suppose a user raises a query "*gladiator*"

History?

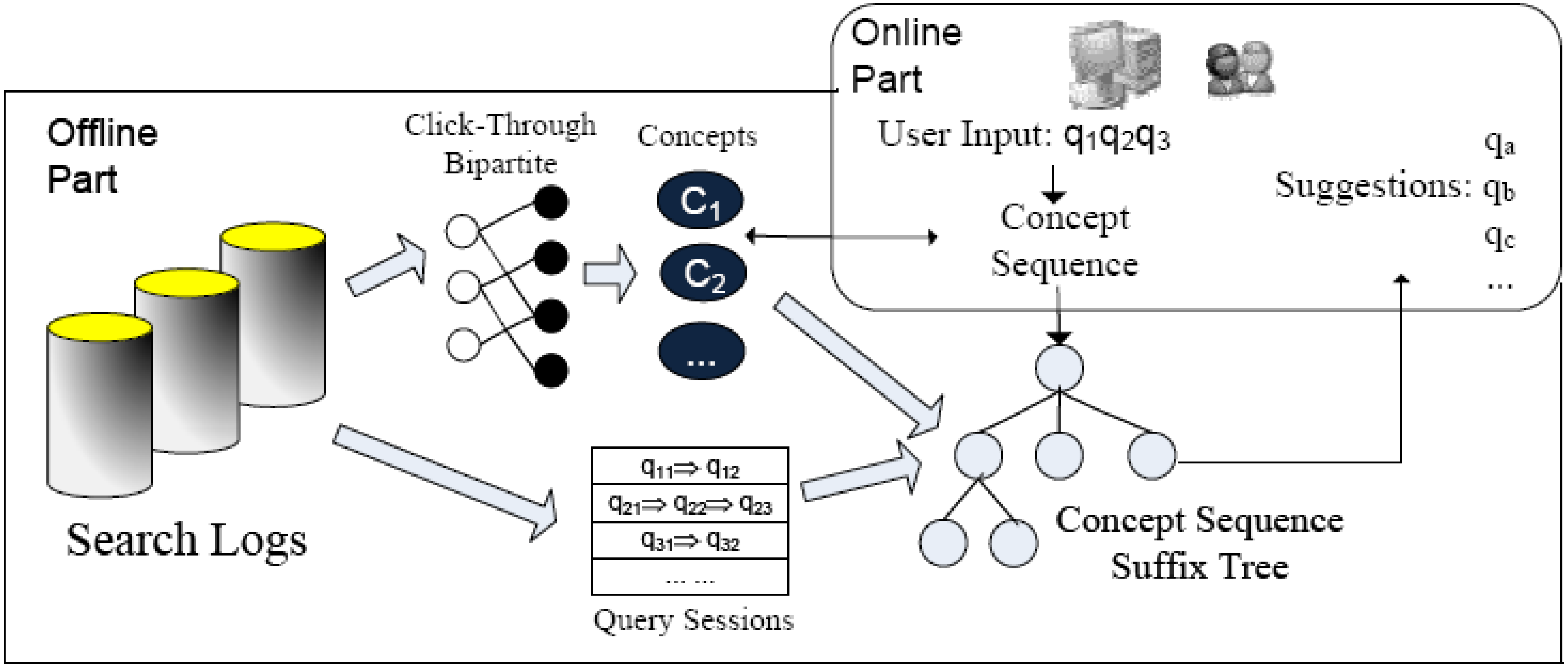
People?



Film?

- If we know the user raises query *"beautiful mind"* before *"gladiator"*
  - User is likely to be interested in the film
  - User is likely to be searching the films played by Russell Crowe.

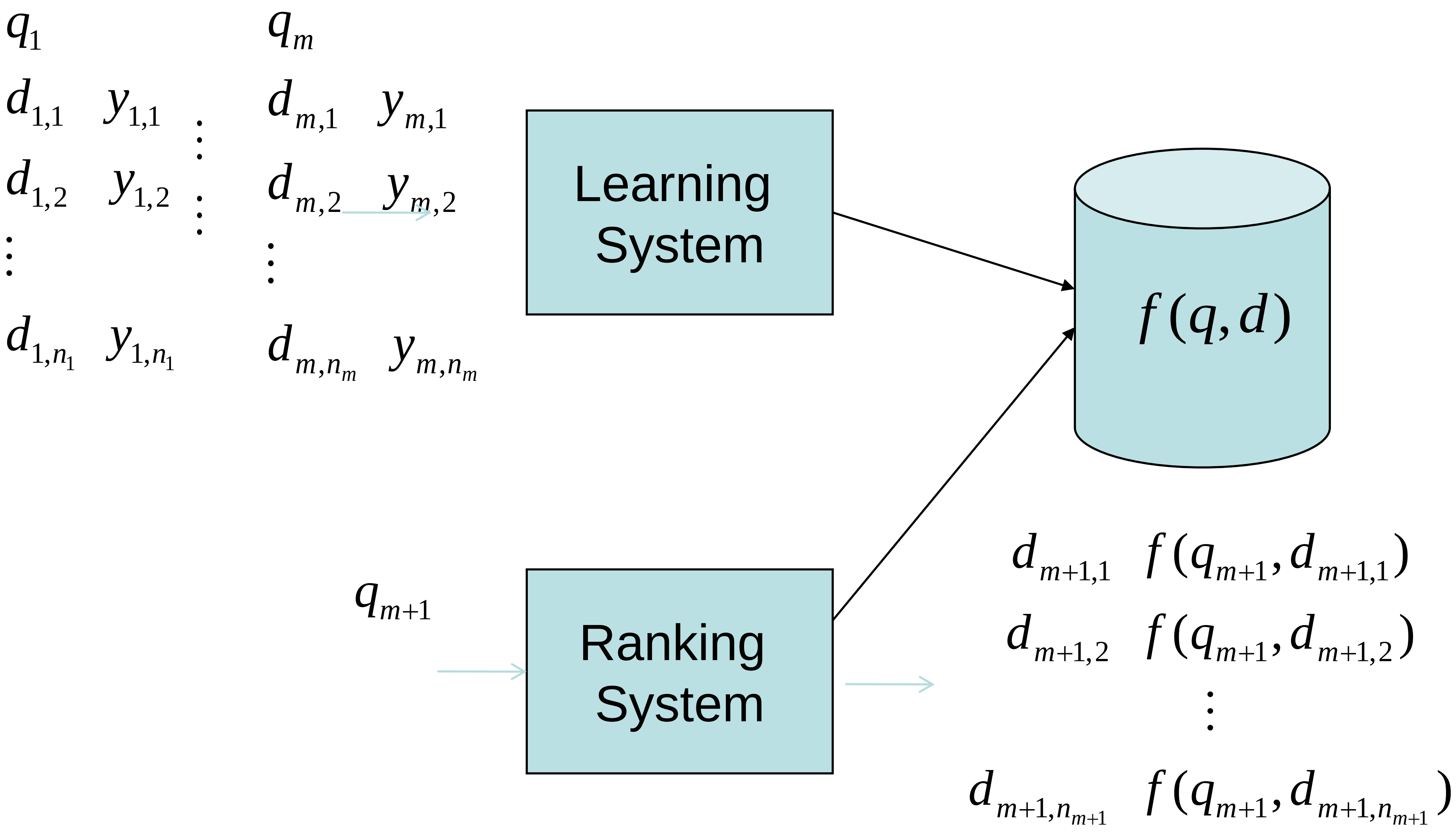
Framework



- Offline part: model learning
  - Summarizing queries into concepts by clustering click-through bipartite
  - Mining frequent patterns from session data and building a concept sequence suffix tree
- Online part: query suggestion

Learning to Rank:
ListWise and Global Ranking
(ICML 2008, NIPS 2008, etc)

Learning to Rank



$$\begin{array}{lll} q_1 & & q_m \\ d_{1,1} \quad y_{1,1} & \vdots & d_{m,1} \quad y_{m,1} \\ d_{1,2} \quad y_{1,2} & \vdots & d_{m,2} \quad y_{m,2} \\ \vdots & & \vdots \\ d_{1,n_1} \quad y_{1,n_1} & & d_{m,n_m} \quad y_{m,n_m} \end{array}$$

Learning System

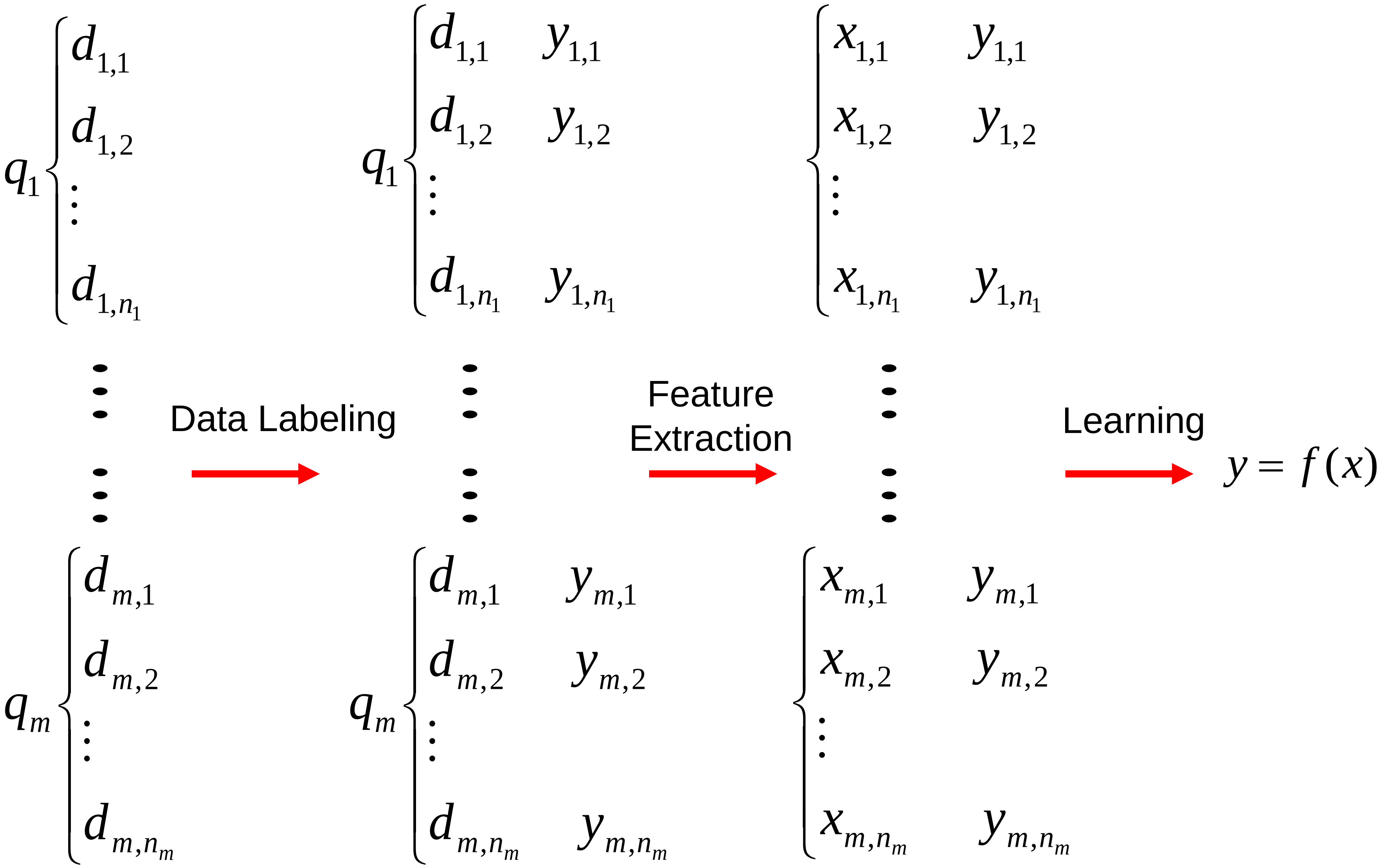
$f(q,d)$

q_{m+1}

Ranking System

$$\begin{array}{ll} d_{m+1,1} & f(q_{m+1}, d_{m+1,1}) \\ d_{m+1,2} & f(q_{m+1}, d_{m+1,2}) \\ & \vdots \\ d_{m+1,n_{m+1}} & f(q_{m+1}, d_{m+1,n_{m+1}}) \end{array}$$

Learning Process



$$q_1 \begin{cases} d_{1,1} \\ d_{1,2} \\ \vdots \\ d_{1,n_1} \end{cases} \quad \vdots \quad q_m \begin{cases} d_{m,1} \\ d_{m,2} \\ \vdots \\ d_{m,n_m} \end{cases}$$

Data Labeling

$$q_1 \begin{cases} d_{1,1} & y_{1,1} \\ d_{1,2} & y_{1,2} \\ \vdots & \\ d_{1,n_1} & y_{1,n_1} \end{cases} \quad \vdots \quad q_m \begin{cases} d_{m,1} & y_{m,1} \\ d_{m,2} & y_{m,2} \\ \vdots & \\ d_{m,n_m} & y_{m,n_m} \end{cases}$$

Feature Extraction

$$\begin{cases} x_{1,1} & y_{1,1} \\ x_{1,2} & y_{1,2} \\ \vdots & \\ x_{1,n_1} & y_{1,n_1} \end{cases} \quad \vdots \quad \begin{cases} x_{m,1} & y_{m,1} \\ x_{m,2} & y_{m,2} \\ \vdots & \\ x_{m,n_m} & y_{m,n_m} \end{cases}$$

Learning

$$y = f(x)$$

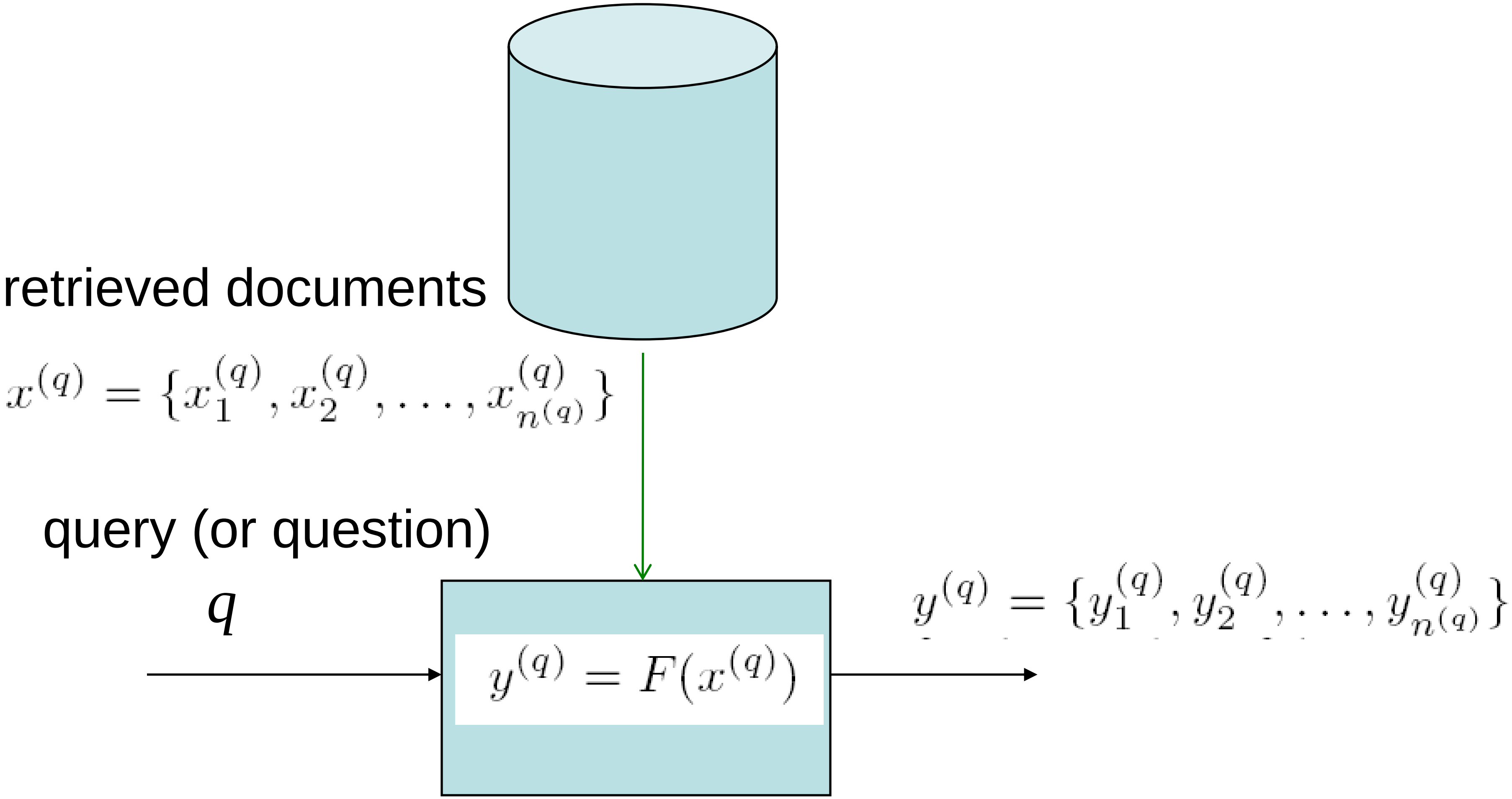
Previous Approach: Pairwise Approach

- Transforming ranking to classification
  - – Ranking SVM (Herbrich et al, 2000)
  - – RankBoost (Freund et al, 2003)
  - – *Ranknet* (Burges et al, 2005)
  - – IR-SVM (Cao et al, 2005)
  - – Frank (Tsai et al, 2006)

Our Proposal = Listwise Approach

- Probabilistic Model
  - ListNet (Cao, et al 2007)
  - ListMLE (Xia, et al 2008)
- Optimizing Upper Bounds of IR Measures
  - AdaRank (Xu & Li, 2007)
  - SVM-MAP (Yue, et al, 2007)
  - PermuRank (Xu, et al 2008)
- Approximation of IR Measures
  - SoftRank (Taylor 2007)
  - ApproxRank (Qin, et al, to appear)

Global Ranking Problem



retrieved documents

$x^{(q)} = \{x_1^{(q)}, x_2^{(q)}, \ldots, x_{n^{(q)}}^{(q)}\}$

query (or question)

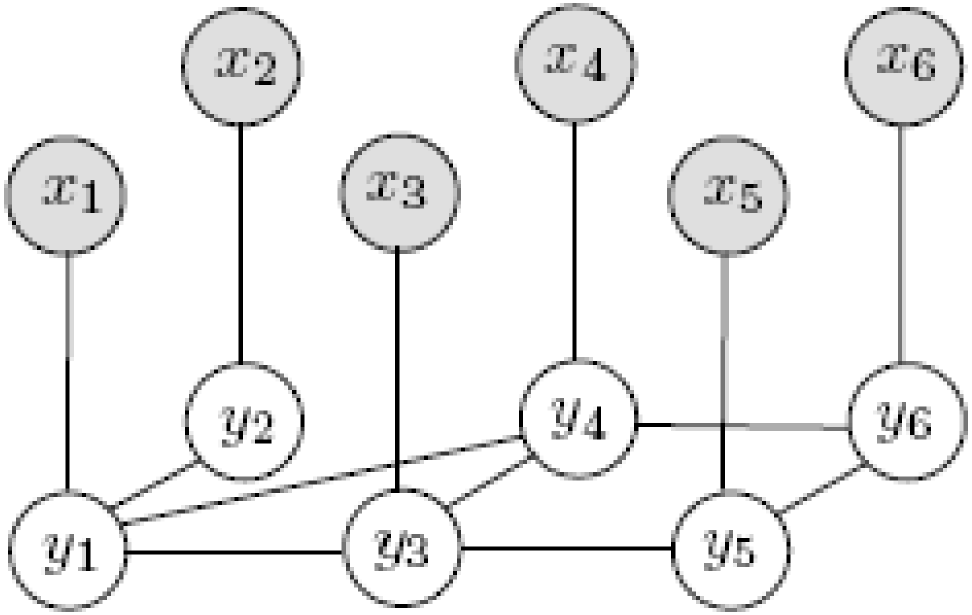
q

$y^{(q)} = F(x^{(q)})$

$y^{(q)} = \{y_1^{(q)}, y_2^{(q)}, \ldots, y_{n^{(q)}}^{(q)}\}$

Global Ranking Using Continuous CRF

$$\Pr(y^{(q)}|x^{(q)}) = \frac{1}{Z(x^{(q)})}\exp\left\{\sum_{i}\sum_{k=1}^{K_1}\alpha_k h_k(y_i^{(q)}, x^{(q)}) + \sum_{i,j}\sum_{k=1}^{K_2}\beta_k g_k(y_i^{(q)}, y_j^{(q)}, x^{(q)})\right\}$$



Summary

- Statistical Learning is Important for Web Search
- Statistical Learning in Web Search
  - Importance Ranking: BrowseRank
  - Anti-Spam: Temporal Classifier
  - Query Understanding: CRF for Query Reformulation
  - Web Page Understanding: HyperText Topic Model
  - Result Presentation: Context Aware Query Suggestion
  - Learning to Rank: Listwise Approach and Global Ranking

References

- Yuting Liu, Bin Gao, Tie-Yan Liu, Ying Zhang, Zhiming Ma, Shuyuan He, Hang Li, BrowseRank: Letting Users Vote for Page Importance, Proc. of SIGIR 2008, 451-458. *SIGIR’ 08 Best Student Paper Award*.
- Jiafeng Guo, Gu Xu, Hang Li, Xueqi Cheng, A Unified and Discriminative Model for Query Refinement, Proc. of SIGIR 2008, 379-386.
- Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhohng Chen, Hang Li, Context-Aware Query Suggestion by Mining Click-Through and Session Data, Proc. of KDD 2008, 875-883, *SIGKDD’ 08 Best Application Paper Award.*
- Guoyang Shen, Bin Gao, Tie-Yan Liu, Guang Feng, Shiji Song, and Hang Li, Detecting Link Spam using Temporal Information, Prof. of ICDM-2006, 1049-1053.
- Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, Hang Li, Listwise Approach to Learning to Rank –Theory and Algorithm, Proc. of ICML 2008, 1192-1199.
- Tao Qin, Tie-Yan Liu, Xu-Dong Zhang, De-Sheng Wang, Hang Li, Global Ranking Using Continuous Conditional Random Fields, Proc. of NIPS 2008, to appear.

Pao-Lu Hsu Lecture on Statistical Machine Learning and Applications

Speaker: Prof. Trevor Hastie

Talk: Regularization Paths

Time: 10:00am-12:00am, December 3, 2008

Place: Peking University Hall (北大百年讲堂会议厅)

Web Site: http://iria.pku.edu.cn/PMSIT/

Contact: Pao-Lu-Hsu@163.com

Thank You!

hangli@microsoft.com