HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

WEB MINING

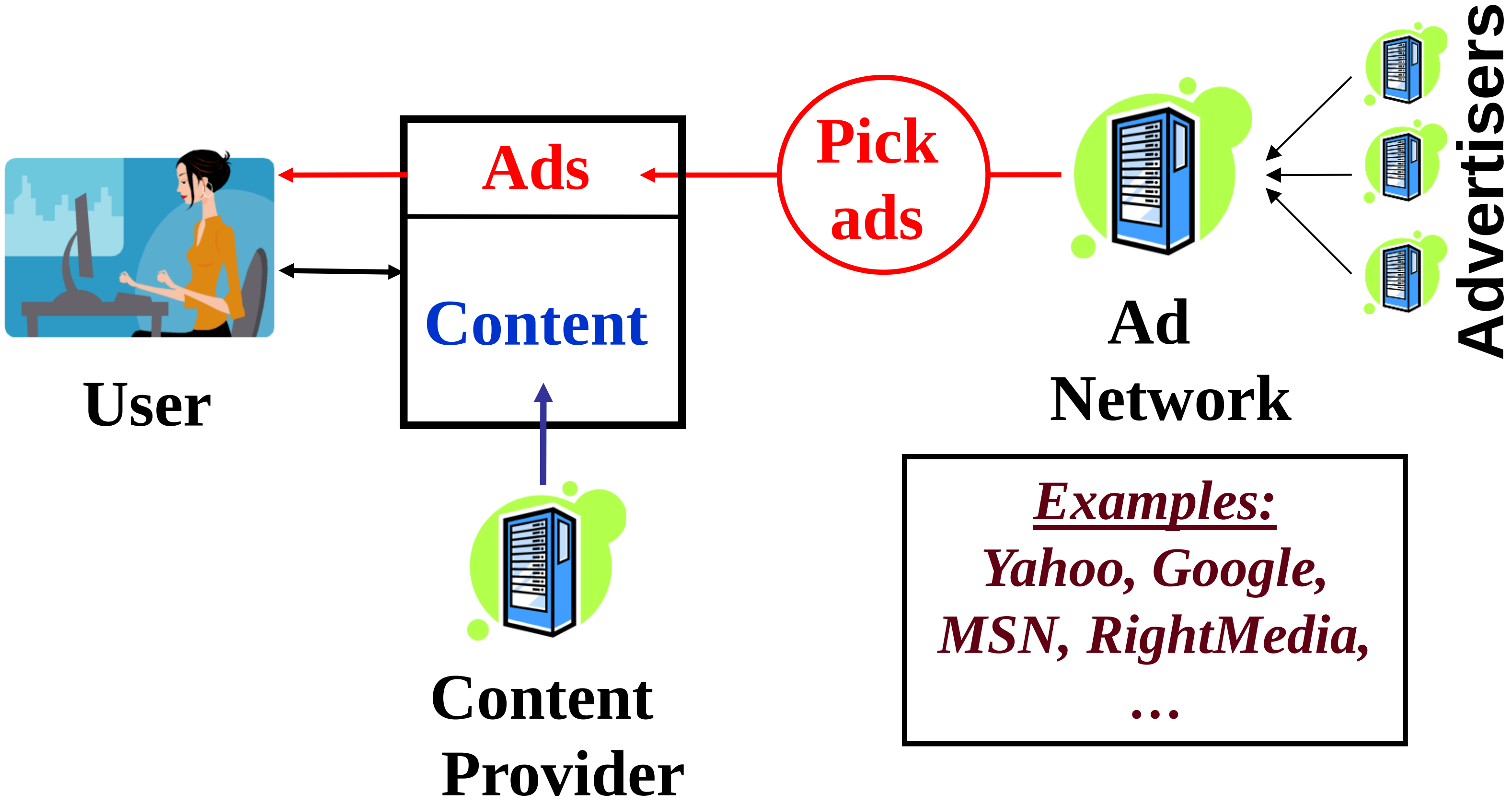
LECTURE 09: ONLINE AD & QUERY MINING

ONE LOVE. ONE FUTURE.

Agenda

1. Online advertising
2. Search engine advertising
3. Query Mining

1. Online Advertising

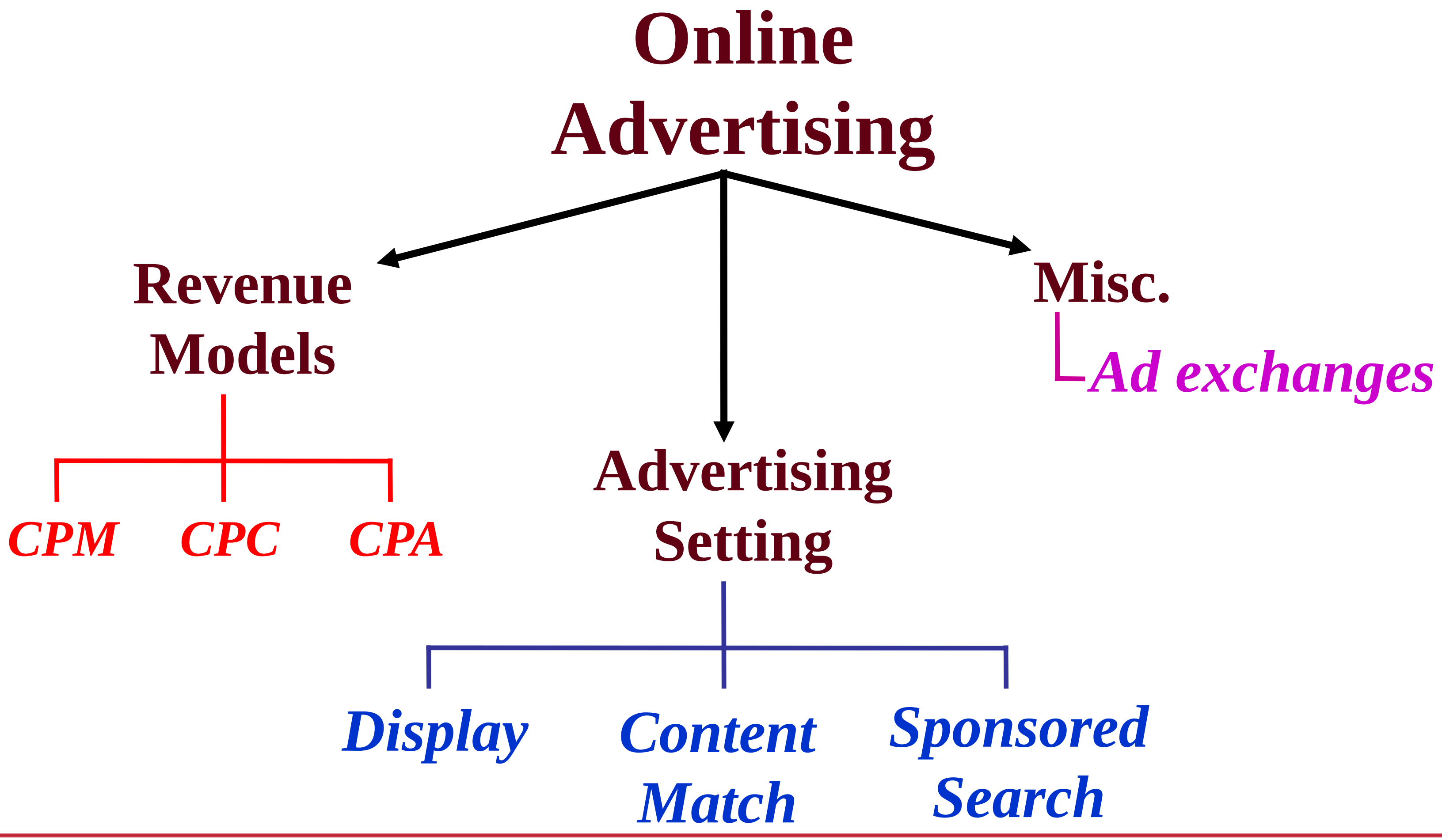


Examples:

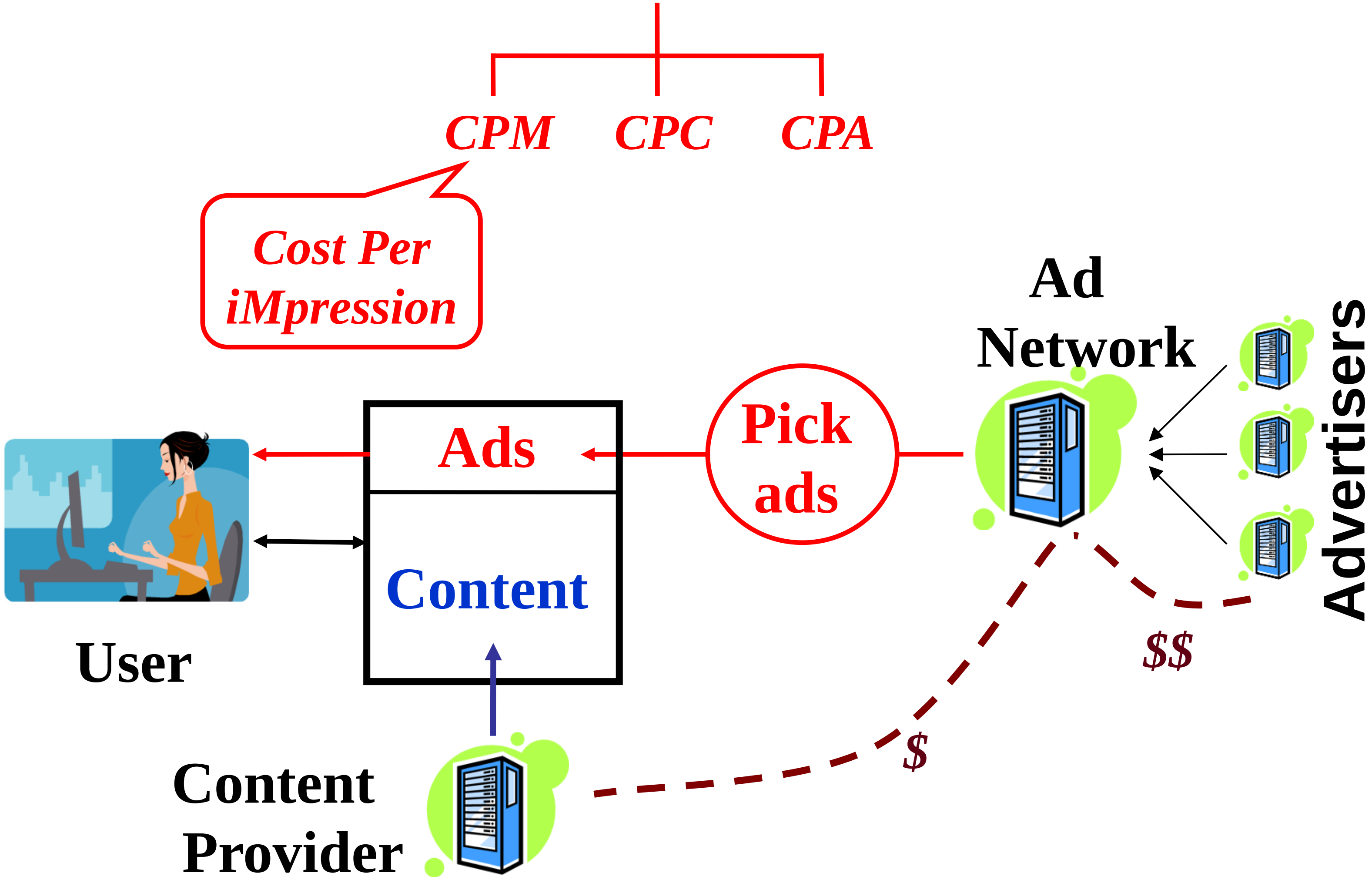
Yahoo, Google, MSN, RightMedia, …

from Agarwal, D (2009)

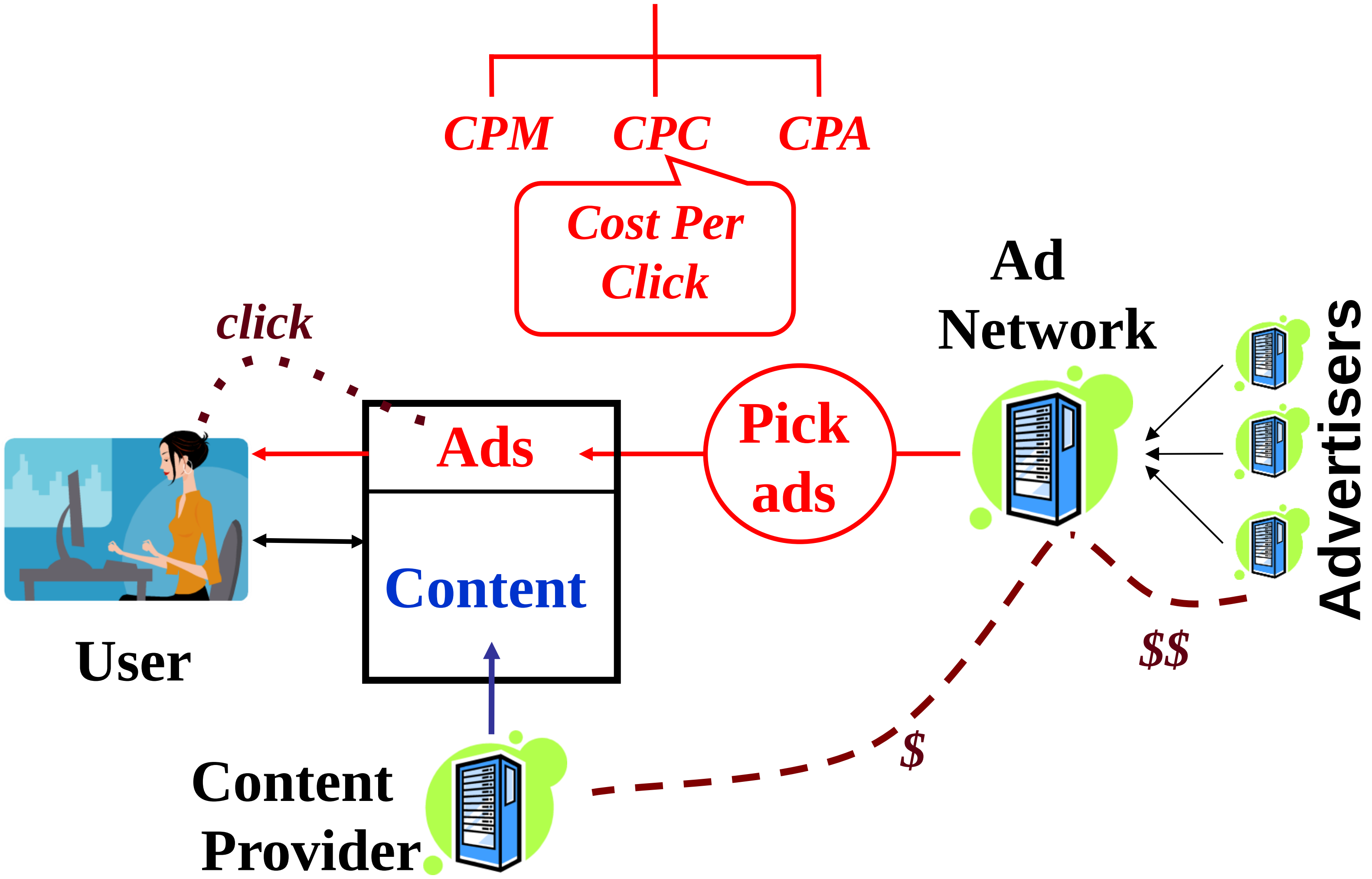
Online advertising model



CPM

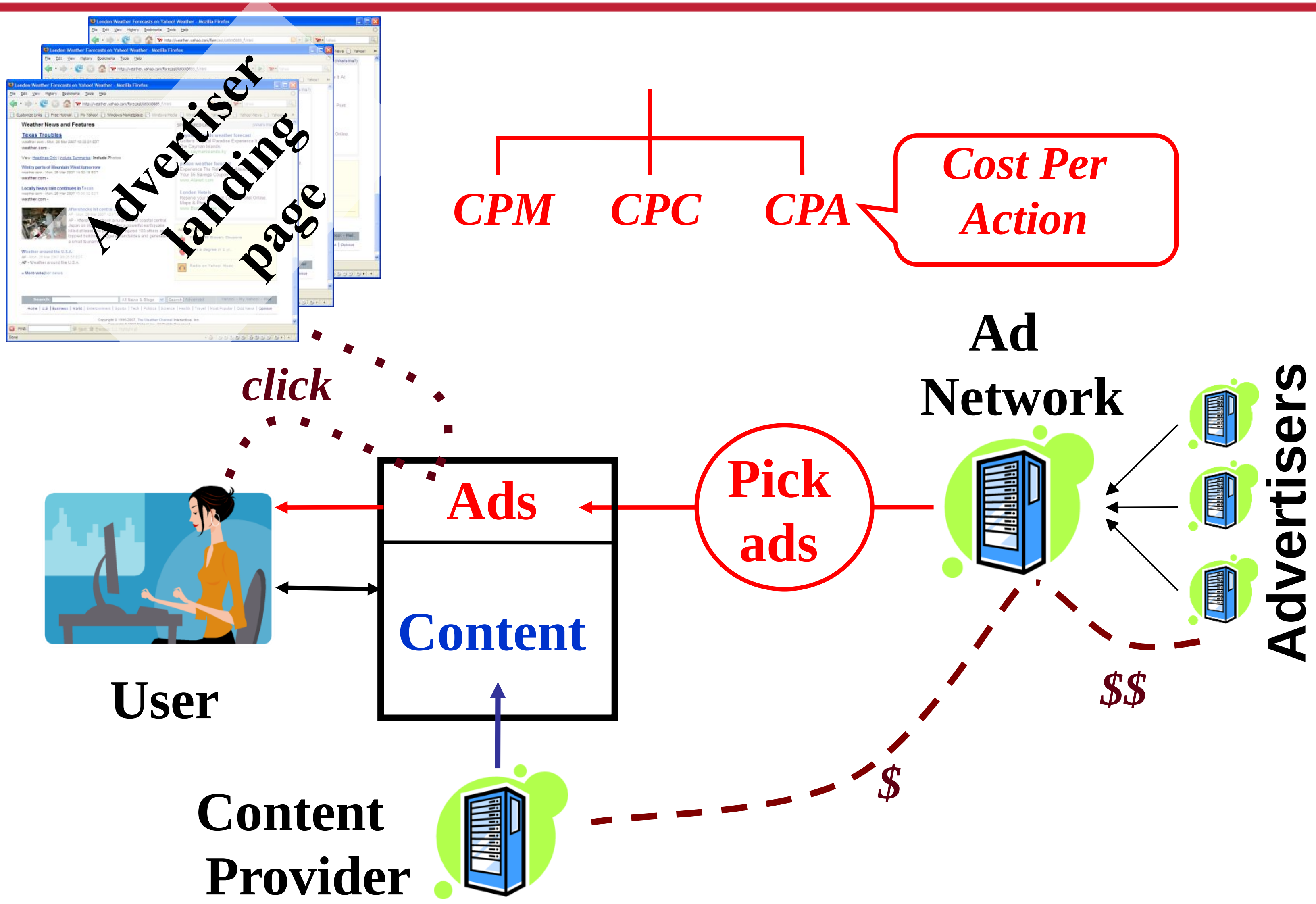

CPM
CPC
CPA
Cost Per iMpression
Ad Network
Advertisers
Pick ads
Ads
Content
User
Content Provider
$$
$

CPC

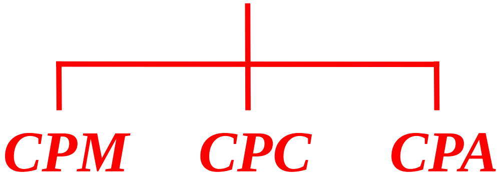


CPM
CPC
CPA
Cost Per Click
click
Ads
Content
User
Pick ads
Ad Network
Advertisers
$$
$
Content Provider

CPA


Advertiser landing page
CPM
CPC
CPA
Cost Per Action
Ad Network
click
Ads
Content
Pick ads
Advertisers
User
$$
$
Content Provider

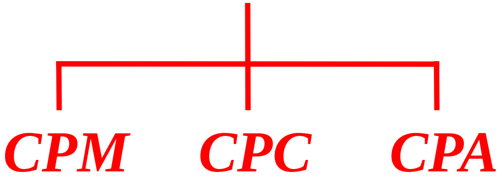
Revenue - CPM



CPM CPC CPA

- Assume that an ad is shown N items at the same position
- CPM: Revenue = N * CPM

Revenue - CPC



CPM CPC CPA

- Assume that an ad is shown N items at the same position
- CPM: Revenue = N * CPM
- CPC: Revenue = N * **CTR** * CPC

Depends on the auction mechanism

Click-through Rate (probability of clicking on an ad)

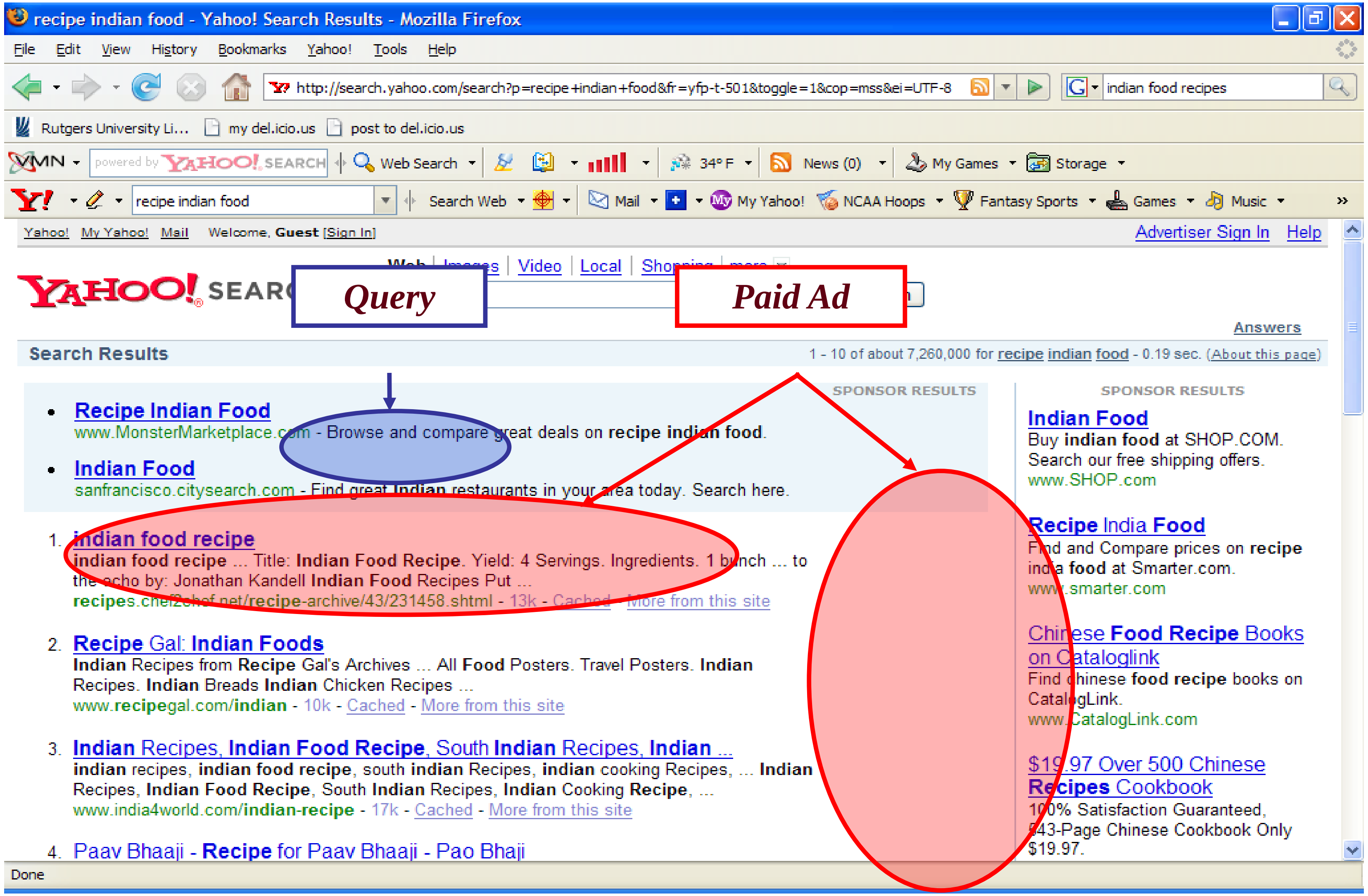
Revenue - CPA

CPM CPC CPA

- Assume that an ad is shown N items at the same position
- CPM: Revenue = N * CPM
- CPC: Revenue = N * CTR * CPC
- CPA: Revenue = N * CTR * **Conv. Rate** * CPA

Conversion Rate (the probability that the user takes an action when viewing the ad page)

2. Search engine advertising



Query

Paid Ad

Search Results

1 - 10 of about 7,260,000 for recipe indian food - 0.19 sec. (About this page)

SPONSOR RESULTS

- **Recipe Indian Food**
  www.MonsterMarketplace.com - Browse and compare great deals on recipe indian food.
- **Indian Food**
  sanfrancisco.citysearch.com - Find great Indian restaurants in your area today. Search here.

1. **indian food recipe**
   indian food recipe ... Title: **Indian Food Recipe**. Yield: 4 Servings. Ingredients. 1 bunch ... to the echo by: Jonathan Kandell **Indian Food** Recipes Put ...
   recipes.chef2chef.net/**recipe**-archive/43/231458.shtml - 13k - Cached - More from this site

2. **Recipe** Gal: **Indian Foods**
   Indian Recipes from **Recipe** Gal's Archives ... All **Food** Posters. Travel Posters. **Indian** Recipes. **Indian** Breads **Indian** Chicken Recipes ...
   www.recipegal.com/**indian** - 10k - Cached - More from this site

3. **Indian** Recipes, **Indian Food Recipe**, South **Indian** Recipes, **Indian** ...
   indian recipes, **indian food recipe**, south **indian** Recipes, **indian** cooking Recipes, ... **Indian** Recipes, **Indian Food Recipe**, South **Indian** Recipes, **Indian** Cooking **Recipe**, ...
   www.india4world.com/**indian-recipe** - 17k - Cached - More from this site

4. Paav Bhaaji - **Recipe** for Paav Bhaaji - Pao Bhaji

SPONSOR RESULTS

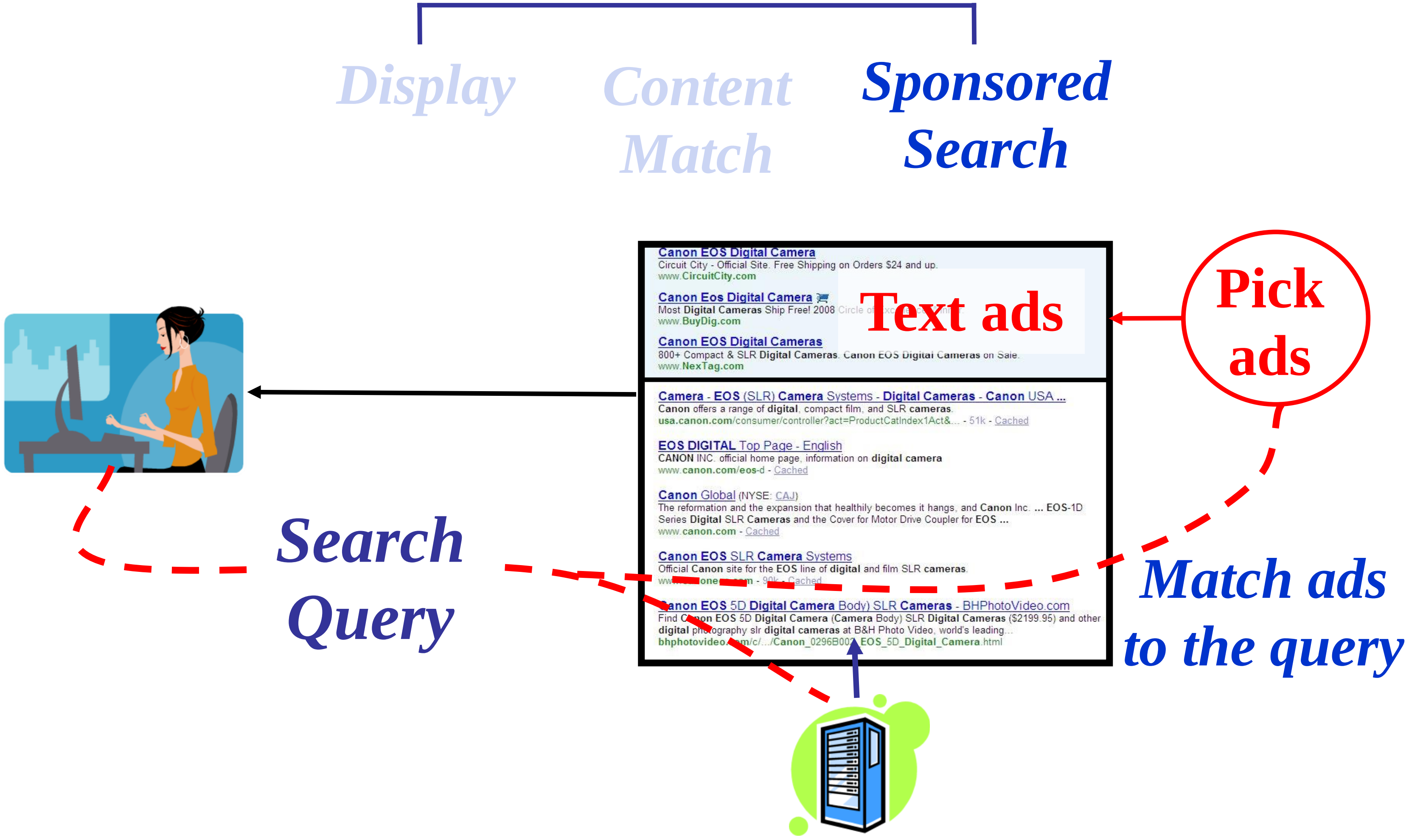
Indian Food
Buy **indian food** at SHOP.COM. Search our free shipping offers.
www.SHOP.com

Recipe India **Food**
Find and Compare prices on **recipe** india **food** at Smarter.com.
www.smarter.com

Chinese **Food Recipe** Books on Cataloglink
Find chinese **food recipe** books on CatalogLink.
www.CatalogLink.com

$19.97 Over 500 Chinese **Recipes** Cookbook
100% Satisfaction Guaranteed. 443-Page Chinese Cookbook Only $19.97.

Search engine advertising model



Canon EOS Digital Camera
Circuit City - Official Site. Free Shipping on Orders $24 and up.
www.CircuitCity.com

Canon Eos Digital Camera
Most **Digital Cameras** Ship Free! 2008 Circle of Excellence Winner.
www.BuyDig.com

Canon EOS Digital Cameras
800+ Compact & SLR **Digital Cameras**. **Canon EOS Digital Cameras** on Sale.
www.NexTag.com

Camera - EOS (SLR) Camera Systems - Digital Cameras - Canon USA ...
Canon offers a range of **digital**, compact film, and SLR **cameras**.
usa.canon.com/consumer/controller?act=ProductCatIndex1Act&... - 51k - Cached

EOS DIGITAL Top Page - English
CANON INC. official home page, information on **digital camera**
www.canon.com/eos-d - Cached

Canon Global (NYSE: CAJ)
The reformation and the expansion that healthily becomes it hangs, and **Canon** Inc. **... EOS**-1D Series **Digital** SLR **Cameras** and the Cover for Motor Drive Coupler for **EOS ...**
www.canon.com - Cached

Canon EOS SLR Camera Systems
Official **Canon** site for the **EOS** line of **digital** and film SLR **cameras**.
ww[illegible]one[illegible]m - 90k - Cached

Canon EOS 5D Digital Camera Body) SLR Cameras - BHPhotoVideo.com
Find Canon **EOS** 5D **Digital Camera** (**Camera** Body) SLR **Digital Cameras** ($2199.95) and other **digital** photography slr **digital cameras** at B&H Photo Video, world's leading...
bhphotovideo.com/c/.../Canon_0296B002_EOS_5D_Digital_Camera.html

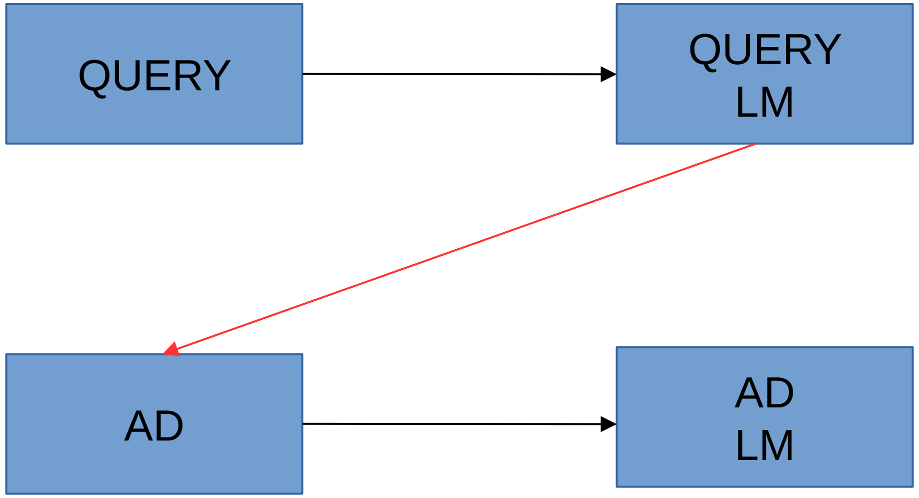
Maximize revenue

- The problem of advertising company
- Select ads for maximum revenue
  - Match the query
  - Advertising costs
  - Ad page quality

Scoring based on content

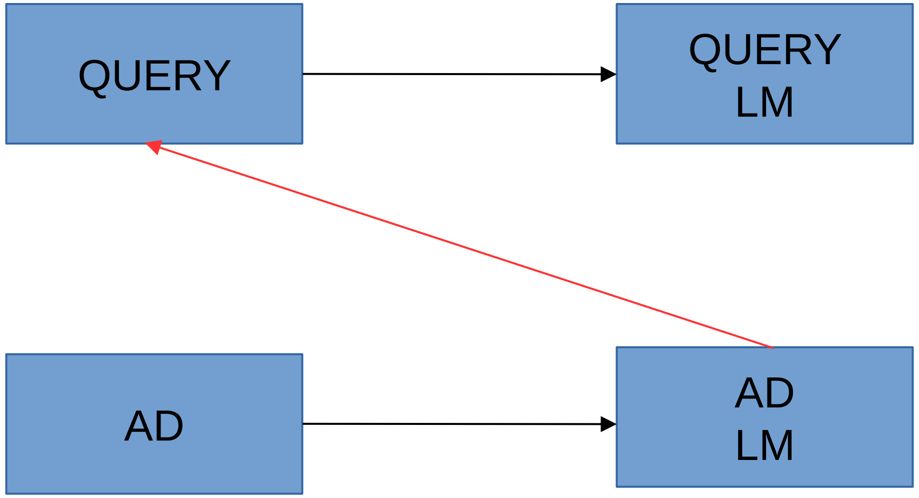
- Consider advertising like a text
- Compare query similarity to ads
- Methods
  - Vector space model
  - Language model

Language model



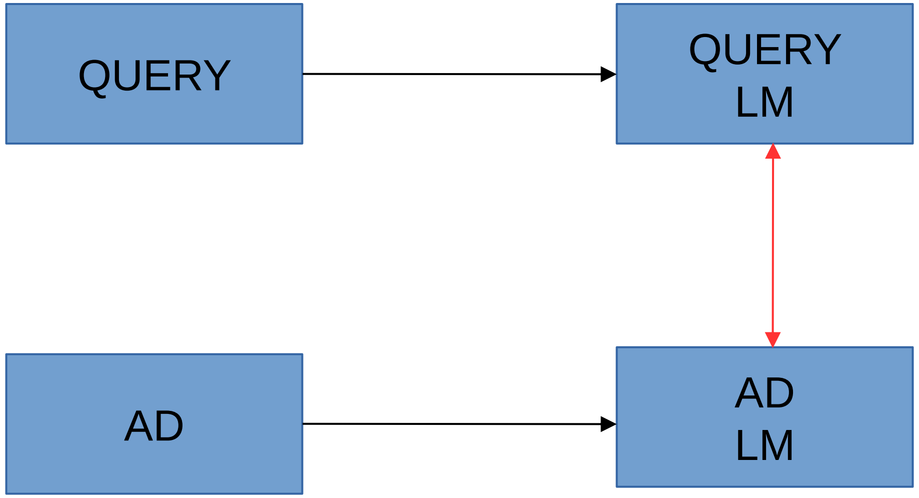
P(ad|query LM)

Language model (cont.)



P(query|ad LM)

Language model (cont.)



KL(ad LM;query LM)

Pros and Cons

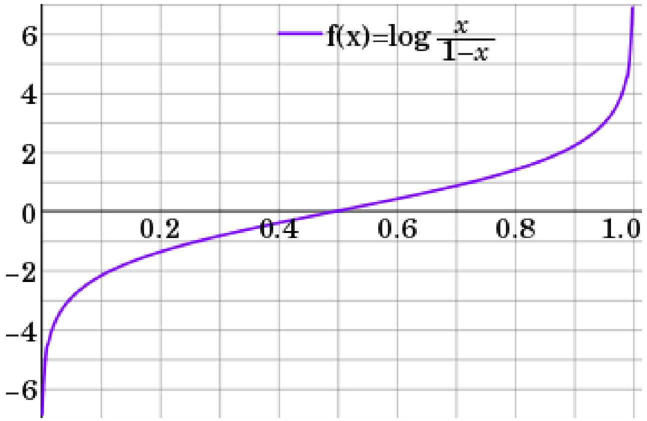
- Pros
  - Simple model
  - Suitable for short popular query
- Cons:
  - Hardly handle rare queries (long tail)
  - Hardly process in real-time
  - Not using user feedback

Score based on user feedback

- Query set Q
- Ad page set A
- For each query $q \in Q$ and ad page $a \in A$, compute the probability that user clicks on ad page Pr(click| q, a)
- Using user feedback to estimate probabilities

Logistic Regression

- Representation of query and advertising content in vectors (bag of words)
- Pr(click| q, a) = f($\boldsymbol{q}$, $\boldsymbol{a}$; θ)
- Logistic Regression:
  - Log-odds (Pr(click |q, a)) = $\boldsymbol{q}' \boldsymbol{W} \boldsymbol{a}$
  - Estimate $\boldsymbol{W}$ using user feedback as training data

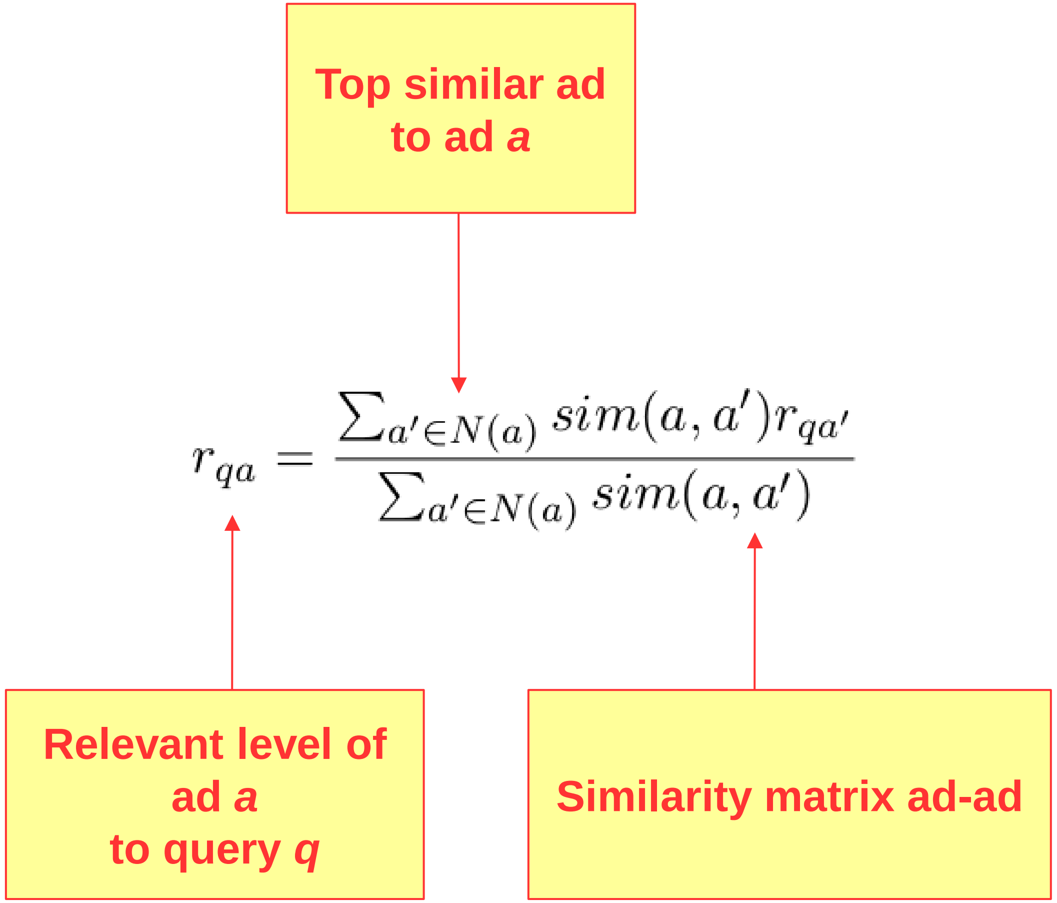


from Wikipedia

Collaborative filtering

- Interactive matrix query, advertising
- Use latent user feedback (click on ad page)
- For each query *q* and ad page *a*, predict user interest
- Collaborative filtering
  - Using kNN
  - Represent ads by query to calculate similarity

Collaborative filtering (cont.)



Top similar ad to ad *a*

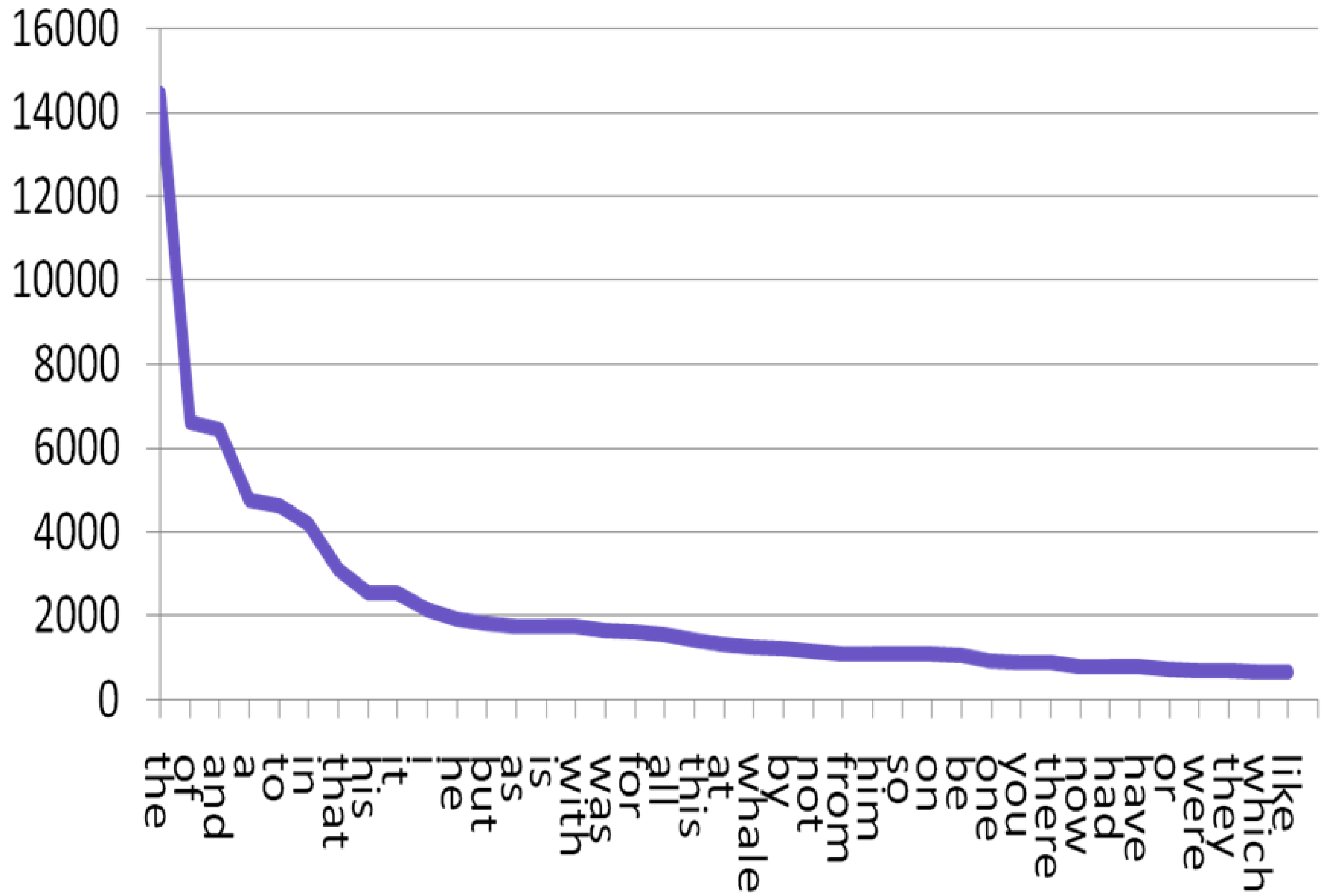
$$r_{qa} = \frac{\sum_{a' \in N(a)} sim(a, a') r_{qa'}}{\sum_{a' \in N(a)} sim(a, a')}$$

Relevant level of ad *a* to query *q*

Similarity matrix ad-ad

3. Query Mining

- Google: 40,000 query/s



Query features

- A query contains an average of 2.4 words
- 21% of internet traffic comes from search engines
- User feedback
  - 50% click on first result
  - Users mostly only use the first two results

Query features (cont.)

- Users often edit query
- Search trends shift from entertainment to e-commerce, in which product search accounts for 1/5
- The distribution of vocabulary on the query and on the website content is different → what users search for is different from what is available on the internet

Query logging

- User information
- Query content
- List of relevant documents
- Selected documents of user

Query preprocessing

- Identify query session
- Filter bot query
- Standardize query

Identify query session

- Classify pairs of consecutive queries into classes :
  - Same query content but different search scope
  - Query Generalization
  - Query fine-tuning for a more precise query
  - Query detailing
  - New query content

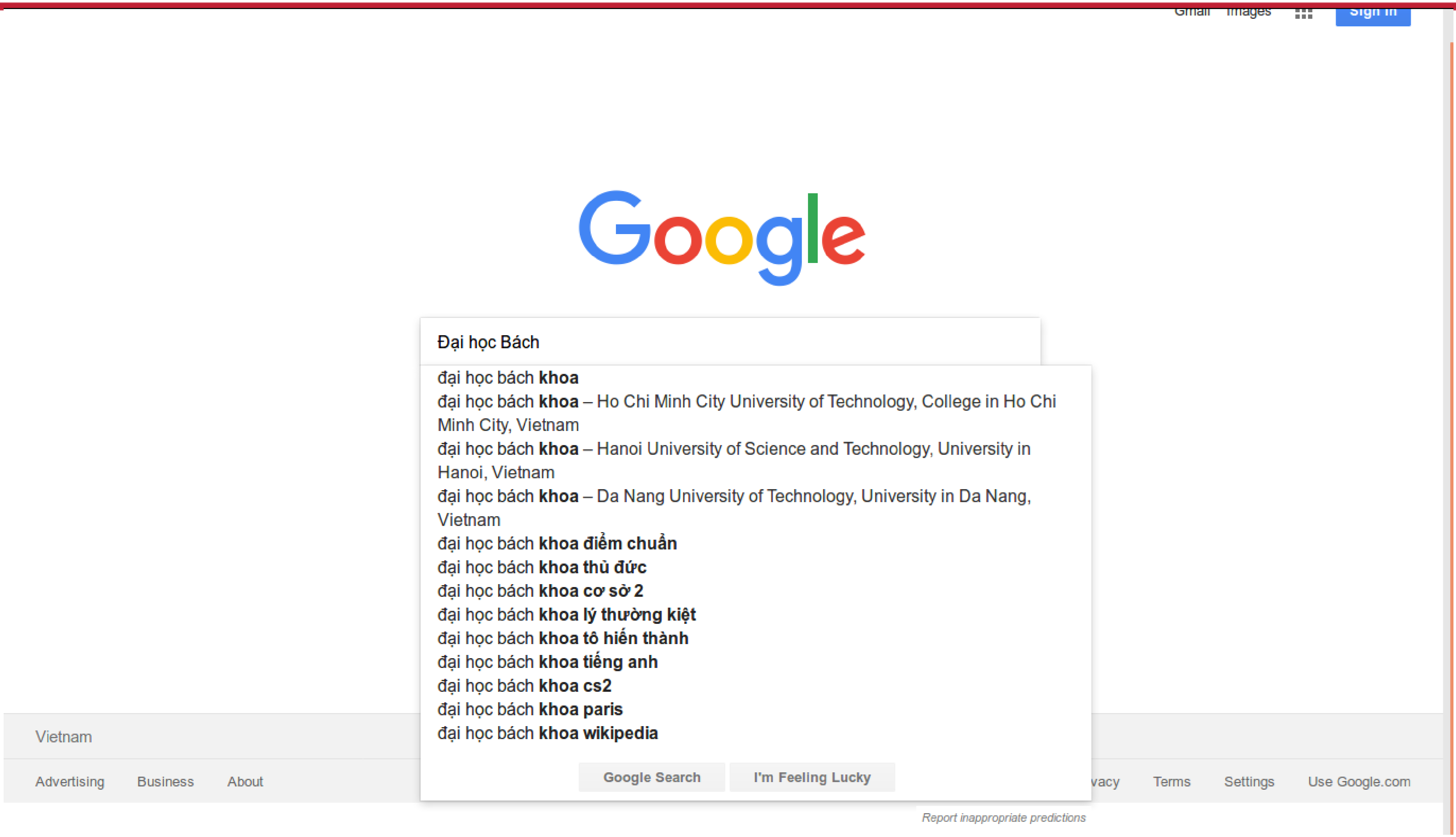
Filter bot query

- Query generated by bot to collect search engine results
- Duplicate content
- Unusually high query rate and/and recurring query frequency

Standardize query

- Remove stopwords
- Convert to lower case
- Standardize number
- Stemming
- For Vietnamese
    - Restore accent
    - Tokenize

Application 1: Query suggestion



Đại học Bách

đại học bách **khoa**
đại học bách **khoa** – Ho Chi Minh City University of Technology, College in Ho Chi Minh City, Vietnam
đại học bách **khoa** – Hanoi University of Science and Technology, University in Hanoi, Vietnam
đại học bách **khoa** – Da Nang University of Technology, University in Da Nang, Vietnam
đại học bách **khoa điểm chuẩn**
đại học bách **khoa thủ đức**
đại học bách **khoa cơ sở 2**
đại học bách **khoa lý thường kiệt**
đại học bách **khoa tô hiến thành**
đại học bách **khoa tiếng anh**
đại học bách **khoa cs2**
đại học bách **khoa paris**
đại học bách **khoa wikipedia**

Language model

- Learn language model on query data

  $argmax_w\ P(w|w_0,w_1,\ldots w_{n-1},w_n)$
- Require large query dataset
- The basic unit of the language model
  - word (tokenize)
  - syllable
  - demisyllabel (‘ch’, ‘ang’)
  - Character

n-gram language model

- Unigram

  $P(w) = (count(w)+1) / (sum_{w'}\ count(w')+V)$

- Bigram

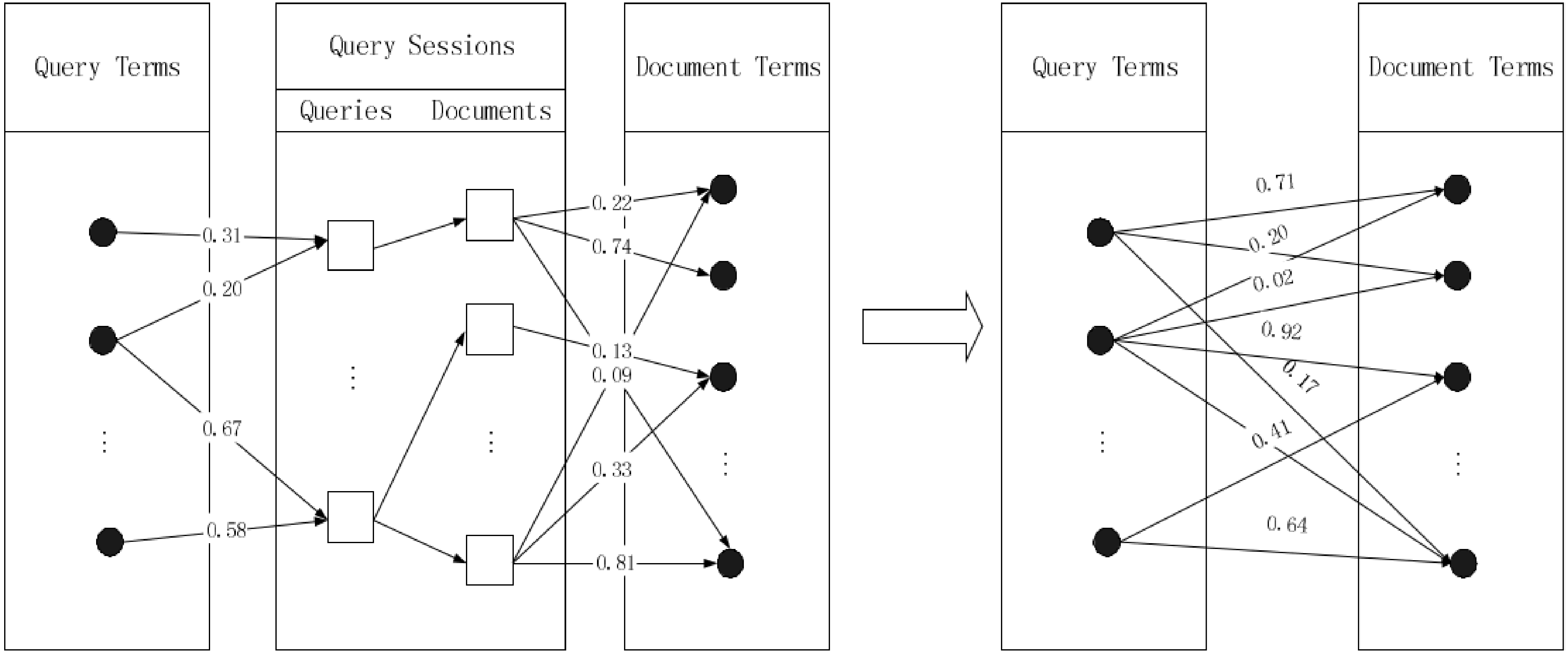
  $P(w_0,w_1) = P(w_1|w_0)*P(w_0)$

  $P(w_1|w_0) = (count(w_0,w_1)+1) / (sum_{w'}\ count(w_0,w')+V)$

Application 2: Extend query

- User queries often do not contain enough information
- Query expansion based solely on textual content may not meet user needs properly
  - Using user feedback
- Assumption: If a query containing one keyword leads to related documents containing another keyword, it is likely that the two keywords are related.

Extend query model



from Hang Cui et al 2003

Extend query model (cont.)

$$P(w_j^{(d)} \mid w_i^{(q)}) = \frac{P(w_j^{(d)}, w_i^{(q)})}{P(w_i^{(q)})}$$

$$= \frac{\sum_{\forall D_k \in S} P(w_j^{(d)}, w_i^{(q)}, D_k)}{P(w_i^{(q)})}$$

$$= \frac{\sum_{\forall D_k \in S} P(w_j^{(d)} \mid w_i^{(q)}, D_k) \times P(w_i^{(q)}, D_k)}{P(w_i^{(q)})}$$

Extend query model (cont.)

$$P(w_j^{(d)} \mid w_i^{(q)}, D_k) = P(w_j^{(d)} \mid D_k)$$

$$P(w_j^{(d)} \mid w_i^{(q)}) = \frac{\sum_{\forall D_k \in S} P(w_j^{(d)} \mid D_k) \times P(D_k \mid w_i^{(q)}) \times P(w_i^{(q)})}{P(w_i^{(q)})}$$

$$= \sum_{\forall D_k \in S} P(w_j^{(d)} \mid D_k) \times P(D_k \mid w_i^{(q)})$$

$P(w_j^{(d)} \mid D_k)$: probability of $w_j^{(d)}$ given selected D_k
$P(D_k \mid w_i^{(q)})$: probability of D_k to be selected if $w_i^{(q)}$ appears in query

Extend query model (cont.)

$$P(D_k \mid w_i^{(q)}) = \frac{f_{ik}^{(q)}(w_i^{(q)}, D_k)}{f^{(q)}(w_i^{(q)})}$$

$$P(w_j^{(d)} \mid D_k) = \frac{W_{jk}^{(d)}}{\max\limits_{\forall t \in D_k}(W_{tk}^{(d)})}$$

$$P(w_j^{(d)} \mid w_i^{(q)}) = \sum_{\forall D_k \in S} (P(w_j^{(d)} \mid D_k) \times \frac{f_{ik}^{(q)}(w_i^{(q)}, D_k)}{f^{(q)}(w_i^{(q)})})$$

$f_{ik}^{(q)}(w_i^{(q)}, D_k)$: number query session in which query contain $w_i^{(q)}$ and D_k is seleted

$f^{(q)}(w_i^{(q)})$: number of query session in which query contain $w_i^{(q)}$

$W_{jk}^{(d)}$: Weight of $w_j^{(d)}$ in document D_k

Extend query model (cont.)

$$CoWeight_Q(w_j^{(d)}) = \ln(\underset{w_t^{(q)} \in Q}{\div} (P(w_j^{(d)} \mid w_t^{(q)}) + 1))$$

1. Extract term in query Q

2. Find documents related to any term

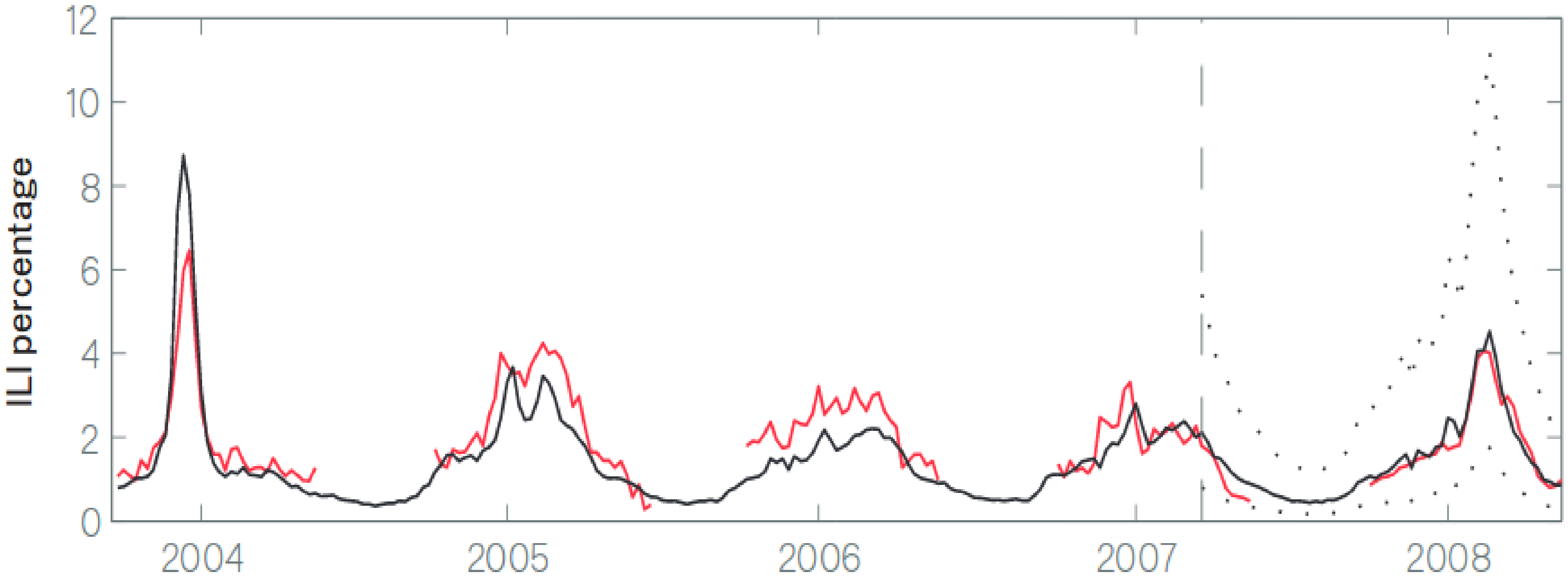
3. For each term in each document, use the formula to measure relevance to query Q

4. Using top n highest score term to construct query Q’

5. Search with query Q’

Application 3: Disease warning

- https://www.google.org/flutrends
- Based on related queries
- The number of people looking for information about the disease is proportional to the number of people who are sick



Jeremy Ginsberg et al 2009

HUST

hust.edu.vn fb.com/dhbkhn

THANK YOU !