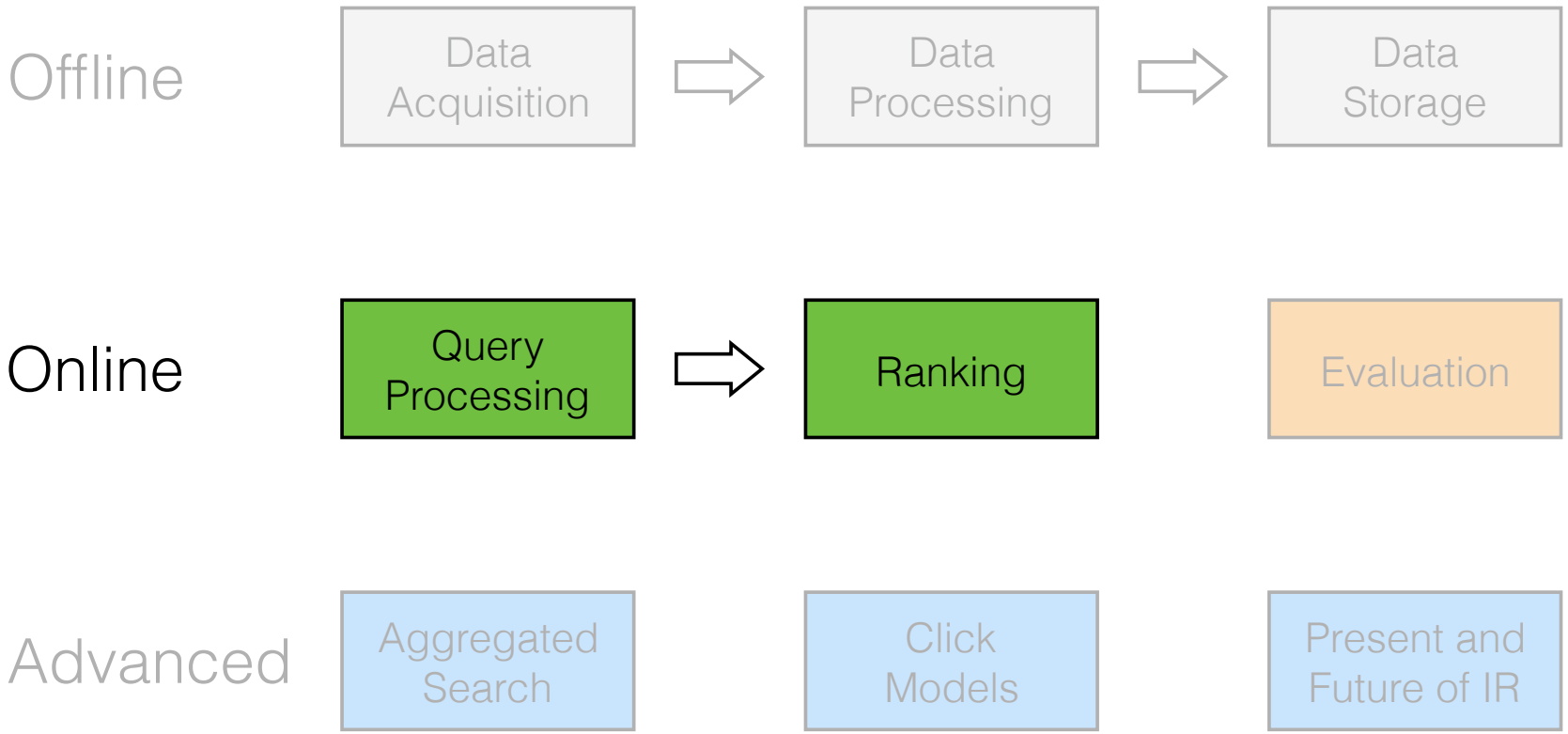
Information Retrieval

Query Processing

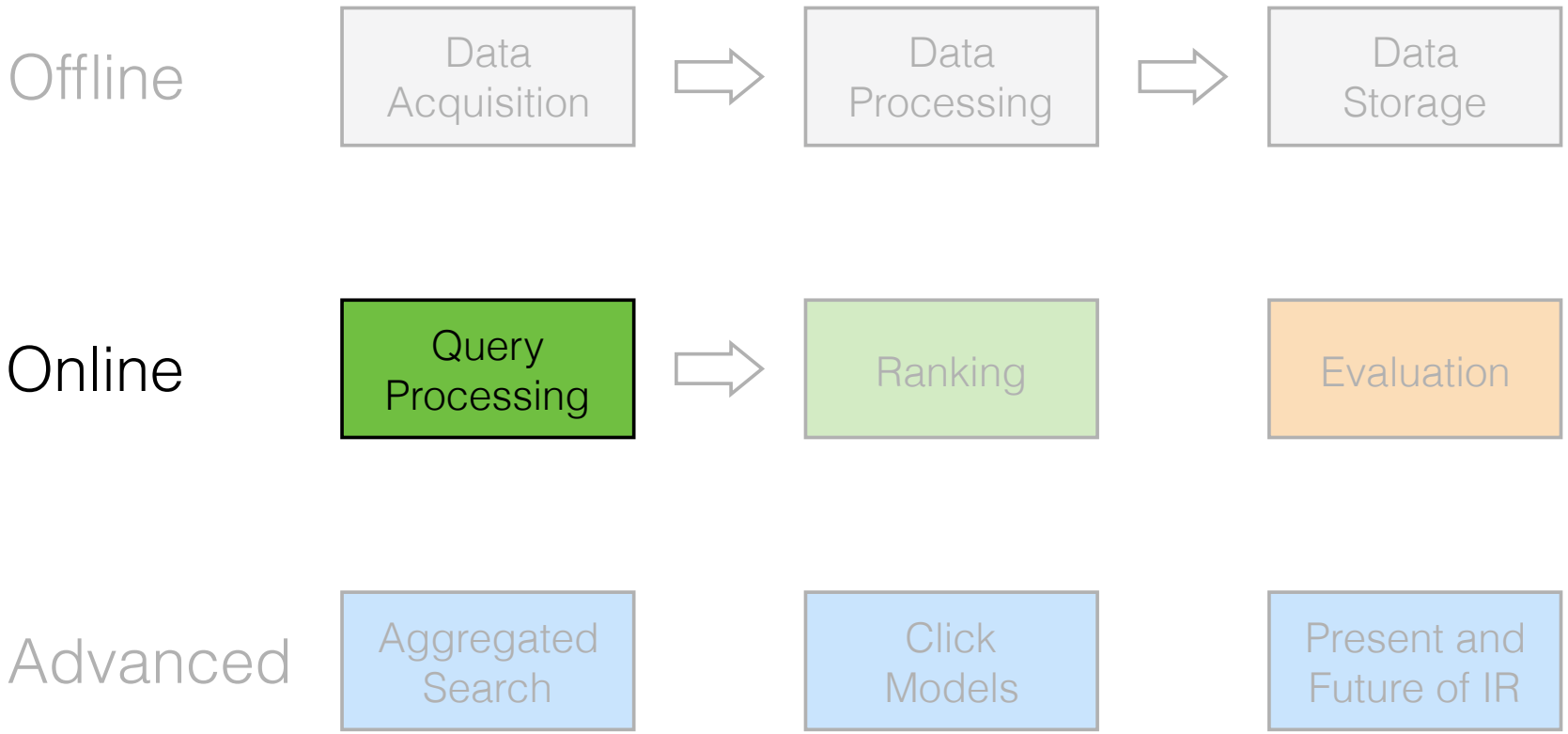
Ilya Markov
i.markov@uva.nl

University of Amsterdam

This block


Offline
Data Acquisition
Data Processing
Data Storage
Online
Query Processing
Ranking
Evaluation
Advanced
Aggregated Search
Click Models
Present and Future of IR

This lecture



Outline

Outline

Spell checking

all images videos shopping news more search tools

about 17 results

Did you mean: **information retrieval**

F. Cai and M. de Rijke, "A Survey of Query Auto Completion in Information Retrieval"

Outline

Simple typos

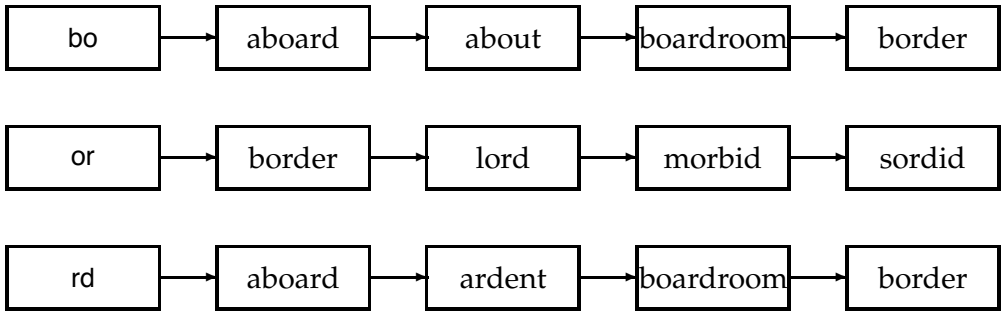
- extenssions $\longrightarrow$ extensions (*insertion error*)
- poiner $\longrightarrow$ pointer (*deletion error*)
- marshmellow $\longrightarrow$ marshmallow (*substitution error*)
- brimingham $\longrightarrow$ birmingham (*transposition error*)

Use an edit distance, e.g., Damerau-Levenshtein distance

Consider only words that...

- start with the same letter
- are of the same or similar length

k-gram index optimization



1. Consider the misspelled word "`bord`"
2. In one pass, find all words that contain at least two bi-grams of "`bord`": `aboard`, `boardroom`, `border`
3. For each candidate, calculate the Jaccard coefficient $|A \cap B|/|A \cup B|$
4. For "`boardroom`" it is $2/(8+3-2)$
5. All required numbers can be obtained efficiently

Manning et al., "Introduction to Information Retrieval"

Outline

Soundex code

1. Keep the first letter (in uppercase)
2. Replace these letters with hyphens: a, e, o, i, u, y, h, w
3. Replace the other letters by numbers as follows
    1. b, f, p, v
    2. c, g, j, k, q, s, x, z
    3. d, t
    4. l
    5. m, n
    6. r
4. Delete adjacent repeats of a number
5. Delete the hyphens
6. Keep the first three numbers or pad our with zeros

Soundex code example

- extenssions $\longrightarrow$ E235; extensions $\longrightarrow$ E235
- poiner $\longrightarrow$ M625; pointer $\longrightarrow$ M625
- marshmellow $\longrightarrow$ B655; marshmallow $\longrightarrow$ B655
- brimingham $\longrightarrow$ P560; birmingham $\longrightarrow$ P536

Outline

Noisy channel model

1. A person chooses a word w to output (i.e., write), based on a probability distribution $P(w)$
2. The person tries to write the word w
3. The noisy channel (e.g., the person's brain) causes the person to output the word e with probability $P(e \mid w)$

Dealing with multiple corrections

- Rank corrections by $P(w \mid e)$

$$P(w \mid e) = \frac{P(e \mid w)P(w)}{P(e)} \propto P(e \mid w)P(w)$$

- $P(w)$ is the probability of the word w in a collection

$$P(w) = \frac{tf(w)}{\sum_{w_i \in C} tf(w_i)}$$

- $P(e \mid w)$ can be estimated in different ways, e.g., by assigning the same probability to errors with the same edit distance

Outline

Considering context

- Rank corrections by $P(e \mid w)\hat{P}(w)$
- Where $\hat{P}(w) = \lambda P(w) + (1 - \lambda)P(w \mid w_p)$
- Example: "`fish tink`"
  - Possible corrections "`think`", "`tank`"
  - $P(tank \mid fish) > P(think \mid fish)$
  - Correct as "`fish tank`"

Spell checking summary

- Simple typos
  - Edit distance
  - k-gram index optimization
- Homophones
  - Soundex code
- Multiple corrections
  - Noisy channel model
- Considering context

Outline

Outline

Thesauri

- Controlled vocabulary with canonical terms
- Manual thesauri, e.g., WordNet
- **Automatically derived thesauri**

Term association measures

- Dice's coefficient

$$\frac{2 \cdot n_{ab}}{n_a + n_b} \propto \frac{n_{ab}}{n_a + n_b}$$

- Mutual information

$$\log \frac{P(a,b)}{P(a)P(b)} = \log N \cdot \frac{n_{ab}}{n_a \cdot n_b} \propto \frac{n_{ab}}{n_a \cdot n_b}$$

- Expected mutual information

$$P(a,b) \log \frac{P(a,b)}{P(a)P(b)} = \frac{n_{ab}}{N} \cdot \log N \cdot \frac{n_{ab}}{n_a \cdot n_b} \propto n_{ab} \cdot \log N \cdot \frac{n_{ab}}{n_a \cdot n_b}$$

- Pearson's χ^2

$$\frac{(n_{ab} - N \cdot P(a) \cdot P(b))^2}{N \cdot P(a) \cdot P(b)} = \frac{\left(n_{ab} - N \cdot \frac{n_a}{N} \cdot \frac{n_b}{N}\right)^2}{N \cdot \frac{n_a}{N} \cdot \frac{n_b}{N}} \propto \frac{\left(n_{ab} - \frac{1}{N} \cdot n_a \cdot n_b\right)^2}{n_a \cdot n_b}$$

Term association example

MIM	*EMIM*	χ^2	*Dice*
zoologico	water	arlsq	species
zapanta	species	happyman	wildlife
wrint	wildlife	outerlimit	fishery
wpfmc	fishery	sportk	water
weighout	sea	lingcod	fisherman
waterdog	fisherman	longfin	boat
longfin	boat	bontadelli	sea
veracruzana	area	sportfisher	habitat
ungutt	habitat	billfish	vessel
ulocentra	vessel	needlefish	marine
needlefish	marine	damaliscu	endanger
tunaboat	land	bontebok	conservation
tsolwana	river	taucher	river
olivacea	food	orangemouth	catch
motoroller	endanger	sheepshead	island

Croft et al., "Search Engines, Information Retrieval in Practice"

Thesauri discussion

- Pros: does not need user input
- Cons: expands each term separately

Outline

Relevance feedback

1. The user issues a (short, simple) query
2. The system returns an initial set of retrieval results
3. **Some returned results are identified as relevant or non-relevant**
4. The system computes a better representation of the information need based on this feedback
5. The system displays a revised set of retrieval results

Types of feedback

1. Relevance feedback
   - Users explicitly mark relevant and non-relevant results
2. Pseudo-relevance feedback
   - The top-k results are assumed to be relevant
3. Implicit relevance feedback
   - Relevant and non-relevant results are identified based on user behavior

Relevance feedback example

1. Badmans **Tropical Fish**

   A freshwater aquarium page covering all aspects of the **tropical fish** hobby. ... to Badman's **Tropical Fish**. ... world of aquariology with Badman's **Tropical Fish**. ...

2. **Tropical Fish**

   Notes on a few species and a gallery of photos of African cichlids.

3. The **Tropical** Tank Homepage - **Tropical Fish** and Aquariums

   Info on **tropical fish** and **tropical** aquariums, large **fish** species index with ... Here you will find lots of information on **Tropical Fish** and Aquariums. ...

4. **Tropical Fish** Centre

   Offers a range of aquarium products, advice on choosing species, feeding, and health care, and a discussion board.

5. **Tropical fish** - Wikipedia, the free encyclopedia

   Tropical fish are popular aquarium **fish** , due to their often bright coloration. ... Practica Fishkeeping • **Tropical Fish** Hobbyist • Koi. Aquarium related companies: ...

6. **Tropical Fish** Find

   Home page for **Tropical Fish** Internet Directory ... stores, forums, clubs, **fish** facts, **tropical fish** compatibility and aquarium ...

7. Breeding **tropical fish**

   ... intrested in keeping and/or breeding **Tropical**, Marine, Pond and Coldwater **fish**. ... Breeding **Tropical Fish** ... breeding **tropical**, marine, coldwater & pond **fish**. ...

8. FishLore

   Includes **tropical** freshwater aquarium how-to guides, FAQs, **fish** profiles, articles, and forums.

9. Cathy's **Tropical Fish** Keeping

   Information on setting up and maintaining a successful freshwater aquarium.

10. **Tropical Fish** Place

    Tropical Fish information for your freshwater **fish** tank ... great amount of information about a great hobby, a freshwater **tropical fish** tank. ...

- Pseudo-relevance feedback
  `tropical (26), fish (28), aquarium (8), freshwater (5), breeding (4), information (3), species (3)`
- (Implicit) relevance feedback
  `breeding (4), fish (4), tropical (4), marine (2), pond (2), coldwater (2)`

Croft et al., "Search Engines, Information Retrieval in Practice"

Relevance feedback implementation

- Depends on a retrieval model
- Will be discussed during the next lectures

Outline

Using query-log

- Find associated terms in user queries
- Short pieces of text are easier to analyze
- Example for "`tropical fish`"

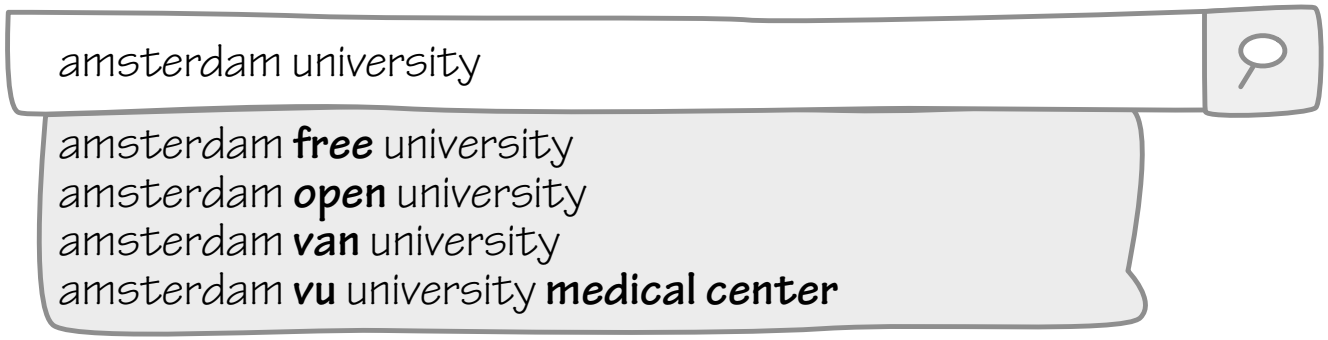
  `stores, pictures, live, sale, types, clipart, blue, freshwater, aquarium, supplies`

Query expansion summary

- Thesauri and term association measures
- Relevance and pseudo-relevance feedback
- Using query-log

Outline

Query suggestion



F. Cai and M. de Rijke, "A Survey of Query Auto Completion in Information Retrieval"

Query suggestion

- Similar to query expansion
- In practice, mainly based on query-logs
  - Session-based query suggestion
  - Click-through-based query suggestion
  - It is always useful to add some sort of query similarity

Outline

Adjacency and co-occurrence

- Adjacency in the same search session $P(q \rightarrow s)$
- Co-occurrence in the same search session $P(q, s)$

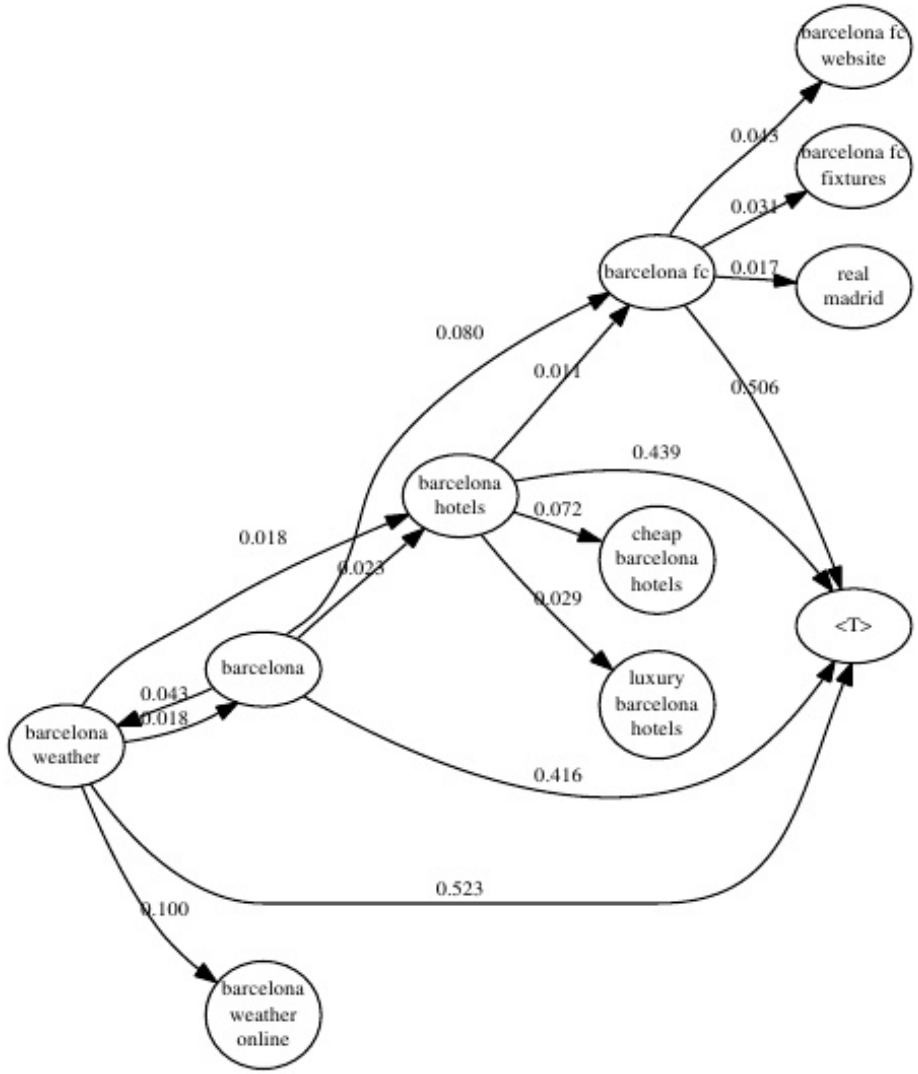
Query flow graph

- $G(Q, E)$ – query flow graph
- Q – set of all queries
- $E \subseteq Q \times Q$ – query transitions
- $r : E \to \mathbb{N}$ – the number of times a transition was observed
- $w(i, j)$ – weight of a transition

$$w(i,j) = \frac{r(i,j)}{\sum_{k:(i,k)\in E} r(i,k)}$$

Query flow graph

- Start with the initial query
- Perform a random walk with the transition probabilities $w(i, j)$
- Suggest queries based on the posterior probabilities



Picture taken from `http://www.slideshare.net/ChaToX/agei`

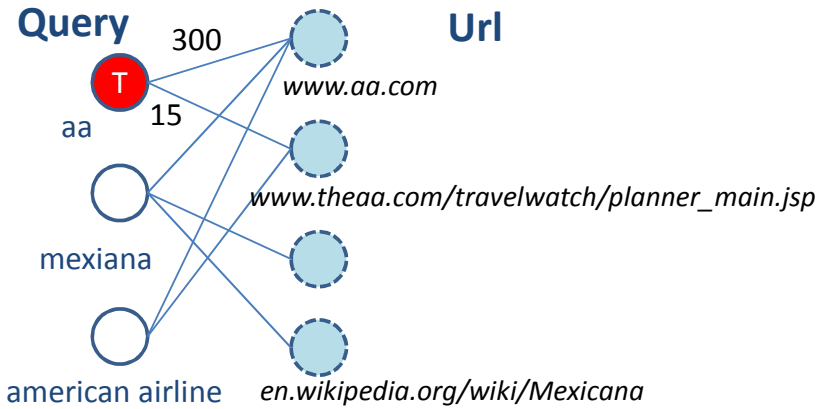
Outline

Clustering

1. Cluster queries based on clicked URLs
2. Suggest queries from the same cluster

Bipartite graph

- $G(V_1 \cup V_2, E)$ – bipartite graph
- V_1 – set of all queries
- V_2 – set of all URLs
- $E \subseteq V_1 \times V_2$ – edge from a query to a clicked URL
- $w(i, k)$ – click frequency for query i and result k



Q. Mei et al., "Query Suggestion Using Hitting Time"

Bipartite graph

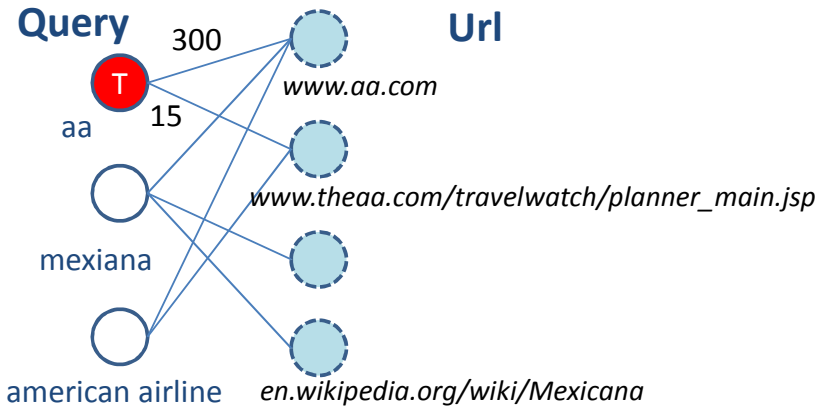
- Given a query, construct a subgraph with n queries using depth-first search
- Perform a random walk on this subgraph

$$p_{ij} = \sum_{k \in V_2} \frac{w(i,k)}{\mathcal{Z}_i} \frac{w(k,j)}{\mathcal{Z}_j}$$

- For each query accumulate time

$$h_i(t+1) = \sum_j p_{ij} h_j(t)$$

- Suggest queries with the smallest final time



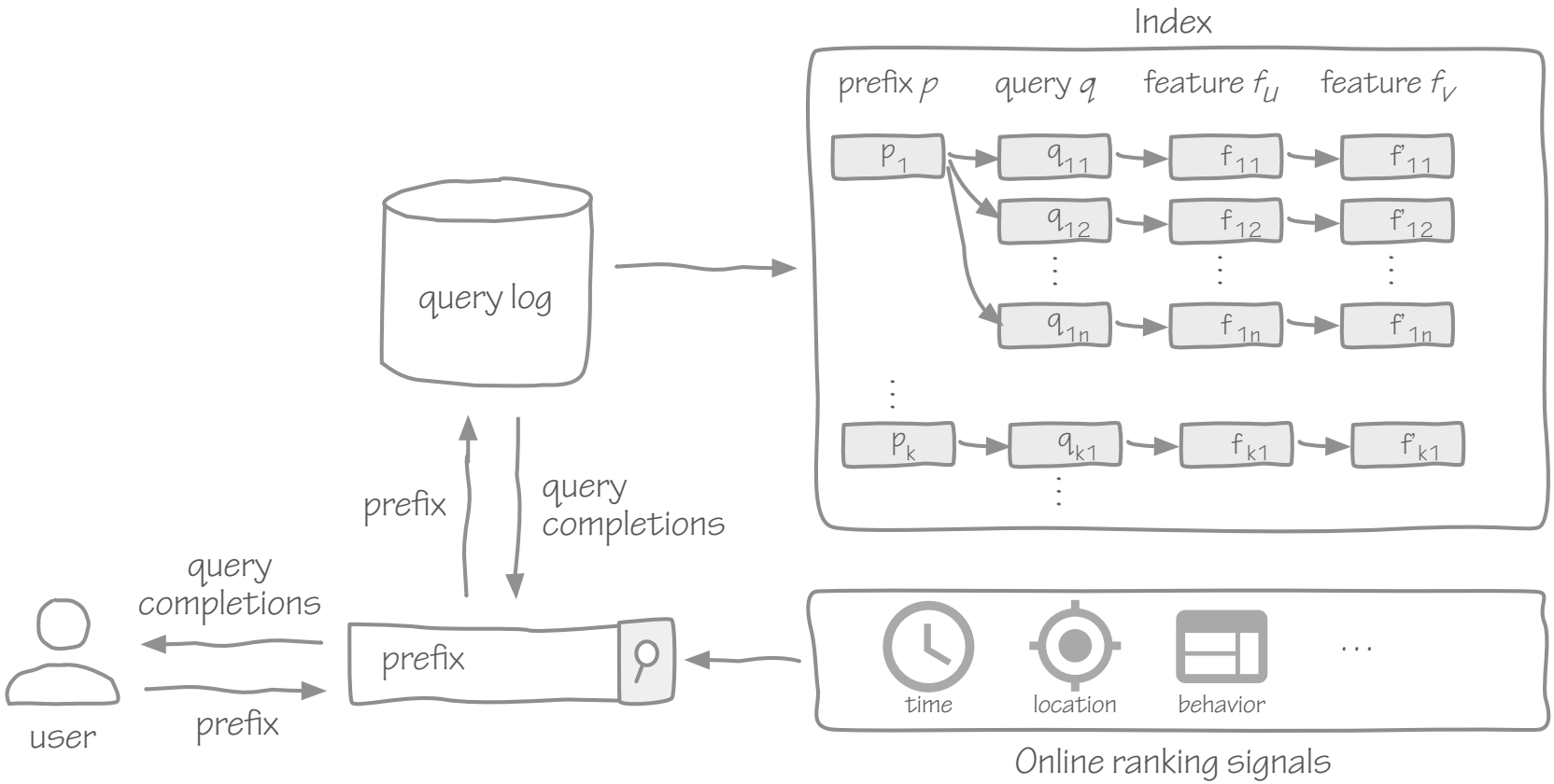
Q. Mei et al., "Query Suggestion Using Hitting Time"

Query suggestion summary

- Session-based query suggestion
  - Adjacency and co-occurrence
  - Query flow graph
- Click-through-based query suggestion
  - Co-clicked URLs and clustering
  - Bipartite graph
- It is always useful to add some sort of query similarity

Outline

Query auto-completion



F. Cai and M. de Rijke, "A Survey of Query Auto Completion in Information Retrieval"

Outline

Most popular completion

$$MPC(p) = \underset{q \in C(p)}{\text{argmax}} \frac{f(q)}{\sum_{q_i \in Q} f(q_i)}$$

- p – prefix
- q – query
- $C(q)$ – candidate completions
- Q – all queries
- $f(q)$ – frequency of query q

Outline

Time-sensitive QAC

$$TS(q,t) = \underset{q \in C(p)}{\text{argmax}} \frac{\hat{f}_t(q)}{\sum_{q_i \in Q} \hat{f}_t(q_i)}$$

- t – time
- $\hat{f}_t(q)$ – estimated frequency of query q at time t

Time-sensitive QAC (cont'd)

$$\hat{f}_{t+1} = \lambda \cdot f_t + (1 - \lambda) \cdot \bar{f}_{t-1}$$

- f_t – observed frequency at time t
- $\bar{f}_{t-1}$ – smoothed frequency at time $t - 1$

$$\hat{f}_t = \lambda \cdot f_t^{trend} + (1 - \lambda) \cdot f_t^{period}$$

- f_t^{trend} – predicted popularity of query q based on recent trends
- f_t^{period} – predicted popularity of query q based on periodicity

Outline

User-centered QAC 1

$$UC1(q) = \lambda \cdot cos(v_q, v_C) + (1 - \lambda) \cdot MPC(q)$$

- v_q – vector representation of query q
- C – context (e.g., previous user queries in the current session)
- cos – cosine similarity
- MPC – most popular completion

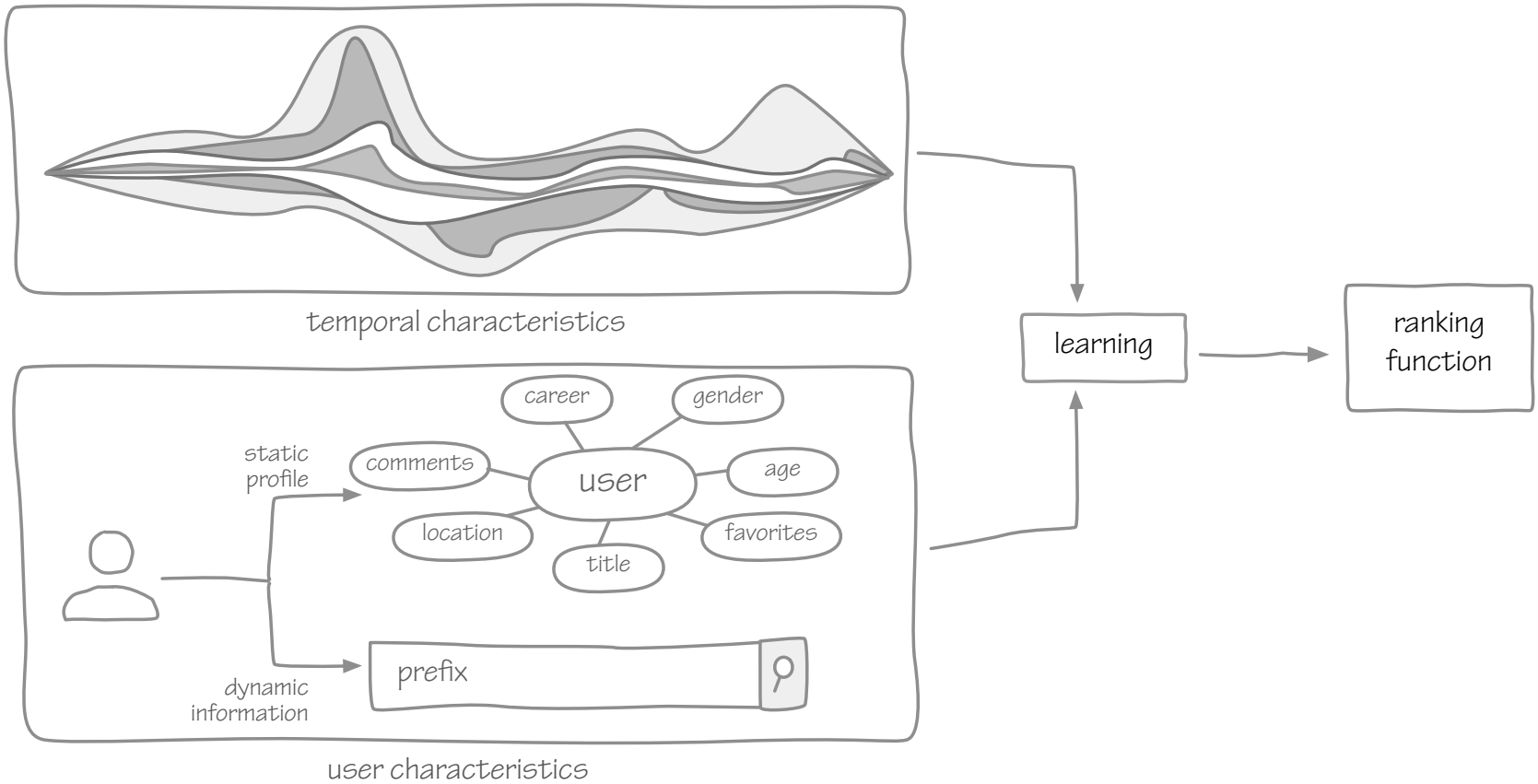
User-centered QAC 2

$$UC2(q) = \lambda \cdot sim(q, Q_s) + (1 - \lambda) \cdot sim(q, Q_u)$$

- *sim* – some similarity score (e.g., cosine)
- Q_s – previous user queries in the current session
- Q_u – all previous user queries

Outline

Learning-based QAC



F. Cai and M. de Rijke, "A Survey of Query Auto Completion in Information Retrieval"

Learning-based QAC

- Popularity-based features
  - Previous observations
  - Future predictions
- Semantic features
  - Semantic relatedness of terms in queries
  - Temporal correlation between queries
- User behavior features

QAC summary

- Frequency-based QAC
- Time-sensitive QAC
- User-centered QAC
- Learning-based QAC

Outline

Query processing summary

- Spell checking
- Query expansion
- Query suggestion
- Query auto-completion

More query processing

- Analyze syntactical structure
- Extract entities
- Interpret semantics

Materials

- Croft et al., Chapter 6.2
- Manning et al., Chapters 3.3–3.4, 9.2
- L. Meng
  A Survey on Query Suggestion
  International Journal of Hybrid Information Technology, 2014
- F. Cai, M. de Rijke
  A Survey of Query Auto Completion in Information Retrieval
  Foundations and Trends in Information Retrieval, 2016

Remaining lectures in this block