Context-Aware Online Commercial Intention Detection

Derek Hao Hu1, Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen
HKUST, Microsoft Research

Outline

- Motivation
- Difficulties
- Problem Formulation
- Solution
- Experiment
- Conclusion

Motivation

- People with commercial intention vs. Search engine
  - Online purchase
  - Online research before actual transactions
- Most web users start their online behaviors by submitting a web query to a search engine

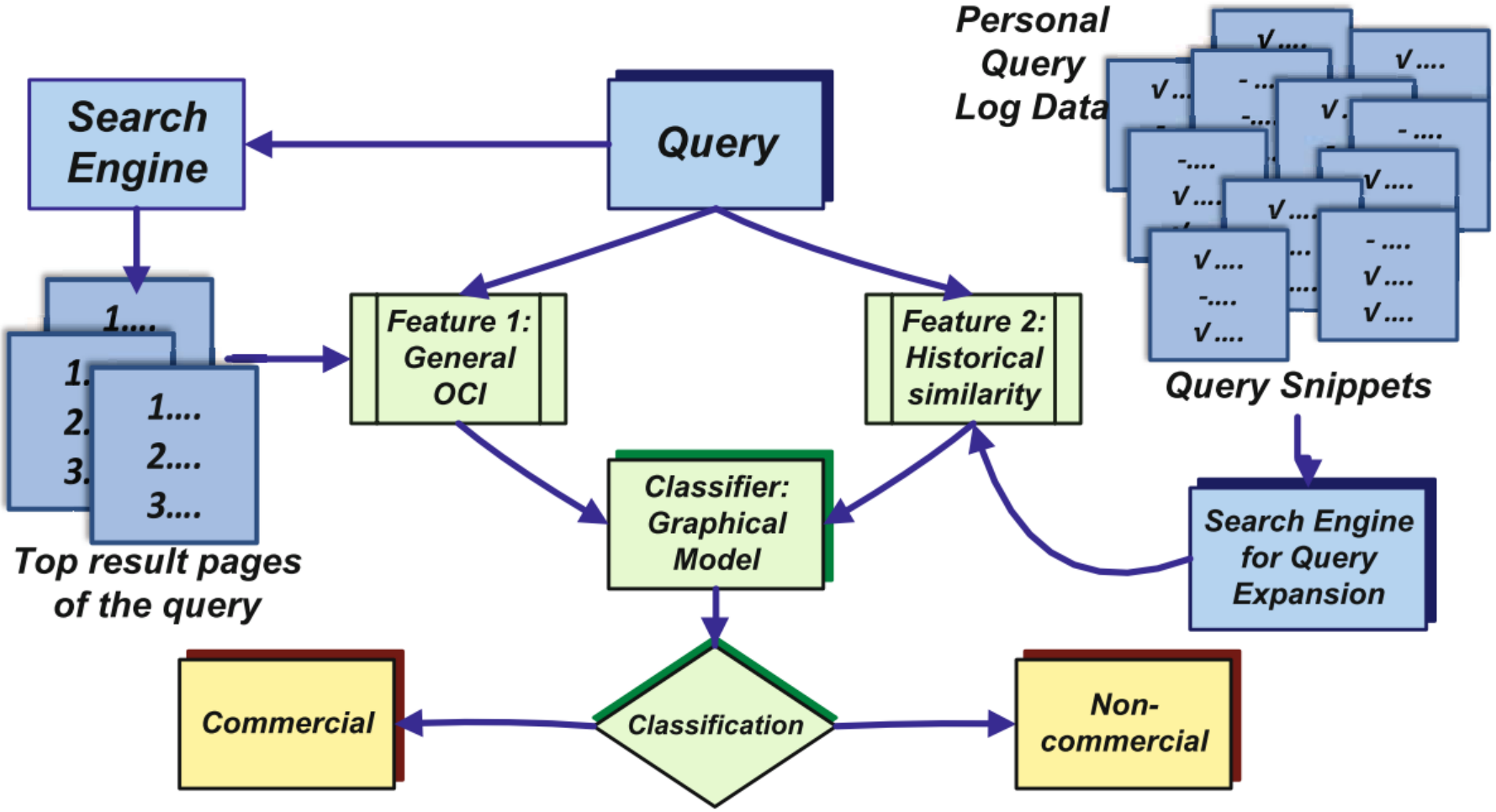
Difficulties

- Queries are very short
  - 93% queries contains less than 4 terms
- Web query often has multiple meanings (ambiguous)
- Intention of a web query can vary for different context

Problem Formulation

- Same as the first work in [1]
- Binary classification problem:
  - Query → {Commercial, Non-Commercial}
  - Query:
    - Text of query term
    - User query history
    - Query timestamp
    - Clickthrough log

Overview


Search Engine
Query
Personal Query Log Data
Query Snippets
Feature 1: General OCI
Feature 2: Historical similarity
Top result pages of the query
Classifier: Graphical Model
Search Engine for Query Expansion
Commercial
Classification
Non-commercial

Overview

- Query log: <U,T,Q,[C]>
  - U: User ID (IP address)
  - T: Query timestamp
  - Q: Text of query term
  - C: Clickthrough log

Modeling Query Logs

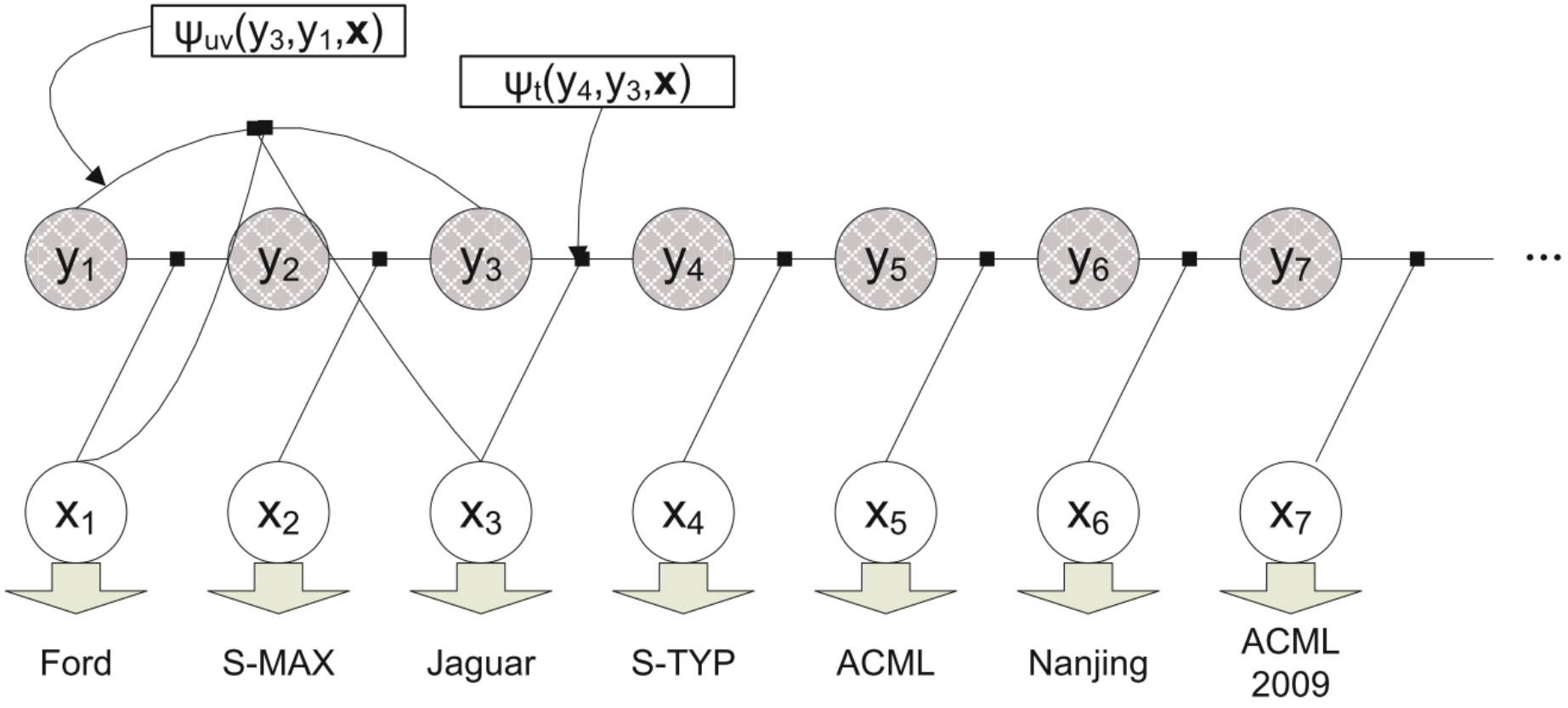
- Skip-chain Conditional Random Field (SCCRF)
- $p(y|x)$ x is the observed personal query log of length L; y is the label of the query
- y_t is the OCI value of t^{th} query
- Threshold: 0.5

Modeling Query Logs

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(x)} \prod_{t=1}^{n} \Psi_t(y_t, y_{t-1}, \mathbf{x}) \prod_{(u,v)\in I} \Psi_{uv}(y_u, y_v, \mathbf{x})$$

- Ψ_t : linear-chain edges
- Ψ_{uv} : skip edges
- $Z(x)$: normalization factor

SCCRF


$\psi_{uv}(y_3,y_1,\mathbf{x})$
$\psi_t(y_4,y_3,\mathbf{x})$
y_1
y_2
y_3
y_4
y_5
y_6
y_7
...
x_1
x_2
x_3
x_4
x_5
x_6
x_7
Ford
S-MAX
Jaguar
S-TYP
ACML
Nanjing
ACML 2009

Modeling Query Logs

$$\Psi_t\left(y_t, y_{t-1}, \mathbf{x}\right) = \exp\left(\sum_k \lambda_{1k} f_{1k}\left(y_t, y_{t-1}, \mathbf{x}, t\right)\right)$$

$$\Psi_{uv}\left(y_u, y_v, \mathbf{x}\right) = \exp\left(\sum_k \lambda_{2k} f_{2k}\left(y_u, y_v, \mathbf{x}, u, v\right)\right)$$

- λ_{1k}, λ_{2k} : parameters
- f_{1k}, f_{2k} : features function

Modeling Semantic Similarities between Queries

- "First order" query expansion:
  - Retrieve the result pages of two queries as documents and get TFIDF vector A and B, then use cosine similarity:

$$\theta = \arccos \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

- "Second order" query expansion

$$f(y_u, y_v, x) = \max_{1 \le i \le m, 1 \le j \le m} g(S_{ui}, S_{vj})$$

  - $g(S_{ui}, S_{vj})$: value of similarities between query snippet
  - S_{ui} : ith query snippet of the uth query

Modeling Semantic Similarities between Queries

- Kernel function for Similarity
  - Top n returned Web page p_i
  - Get TFIDF vector v_i for each page p_i
  - Truncate v_i to include m highest terms (m=50)
  - $C(x)$ be the centroid of L2 normalized vectors

$$C(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{v_i}{\|v_i\|_2}$$

  - $QE(x)$ be the L2 normalization of $C(x)$

$$QE(x) = \frac{C(x)}{\|C(x)\|_2}$$

  - Similarity: $K(x, y) = QE(x) \cdot QE(y)$

Algorithm

- Input:
  - N: the length of a query log,
  - Each query item is represented by $\{x_i, y_i\}$
  - x_i is the i^{th} query and y_i is the corresponding i^{th} label for x_i.
  - Q, which is a newly asked query.
- Output:
  - P , which is the probability for Q as being commercial intended.

Algorithm

- Assumption:
  - Assume all the queries in the personal query log we considered here are issued by the same user or user group
- Parameters:
  - θ: suggests the confidence parameter for us to add the skip edges
  - L: the length of the personal query log training data

Algorithm

1: **for** $i = 1$ to $N - L + 1$ **do**
2: Initialize the i^{th} training data as empty.
3: **for** $j = 0$ to $L - 1$ **do**
4: Add the $(i + j)^{th}$ query x_{i+j} to the i^{th} training data.
5: **end for**
6: **end for**
7: **for** $i = 1$ to N **do**
8: Issue the query x_i to the search engine to get the top P landing pages. P can be tuned to reflect more information from landing pages. To simplify, we set $P = 10$ in our experiments.
9: Compute the corresponding OCI value of these landing pages from the baseline method
10: Use these values as features for f_1.
11: **end for**

Algorithm

Train the corresponding SCCRF model from the training set created.
for $i = N - L + 2$ to N **do**
    Add the query x_i to the test personal query log.
end for
Add the query Q to the test personal query log. Now it contains L terms.
for $i = 1$ to L **do**
    for $j = 1$ to $i - 1$ **do**
        Compute the semantic similarity of T_i and T_j, i.e. $K(T_i, T_j) = QE(T_i) \cdot QE(T_j)$ as defined.
        if $K(T_i, T_j) > \theta$ **then**
            Add a skip edge between y_i and y_j, corresponding to the feature function $f_2(y_i, y_j, \mathbf{x})$.
        end if
    end for
end for

Experiment

- AOL query log dataset (http://www.gregsadetsky.com/aol-data/)
- Live Search collected in March 2008
- 100 users at least 100 queries

Labeler	AOL Commercial	AOL Non-commercial	Live Commercial	Live Non-Commercial
1	1238	8627	919	8819
2	1430	8435	1025	8713
3	1117	8748	973	8765
Sum	1247	8306	936	8738

Experiment

Baseline classifier [1]

Dataset	Precision	Recall	F1-Measure
AOL	0.817	0.796	0.806
Live Search	0.802	0.836	0.809

Experiment

Proposed Algorithm with varying parameter θ, L=50

θ	AOL (Variance)	Live Search (Variance)
$\theta = 0.01$	0.863 (0.002)	0.872 (0.003)
$\theta = 0.02$	0.887 (0.005)	0.878 (0.003)
$\theta = 0.04$	0.892 (0.003)	0.881 (0.004)
$\theta = 0.08$	0.901 (0.005)	0.893 (0.002)
$\theta = 0.1$	**0.913 (0.002)**	0.901 (0.004)
$\theta = 0.2$	0.912 (0.005)	**0.908 (0.003)**
$\theta = 0.4$	0.902 (0.004)	0.883 (0.006)
$\theta = 0.8$	0.871 (0.003)	0.852 (0.008)
Baseline	0.806	0.809

Experiment

Proposed Algorithm with varying parameter L, θ=0.1

L	AOL (Variance)	Live Search (Variance)
$L = 5$	0.872 (0.010)	0.871 (0.013)
$L = 10$	0.893 (0.011)	0.878 (0.010)
$L = 15$	0.882 (0.009)	0.891 (0.005)
$L = 20$	0.901 (0.005)	0.891 (0.003)
$L = 25$	0.910 (0.004)	0.897 (0.007)
$L = 30$	**0.913 (0.002)**	0.901 (0.004)
$L = 40$	0.909 (0.003)	**0.903 (0.005)**
$L = 50$	0.905 (0.003)	0.902 (0.003)
Baseline	0.806	0.809

Experiment

Training Time with varying parameter L, θ=0.1

L	AOL Time	Live Search Time
$L = 5$	1.7s	1.7s
$L = 10$	3.0s	4.1s
$L = 15$	4.9s	5.2s
$L = 20$	6.2s	6.8s
$L = 25$	9.0s	10.2s
$L = 30$	11.1s	11.7s
$L = 40$	14.0s	14.1s
$L = 50$	15.3s	16.3s

Pentium Core 2 Dual 2.13GHz CPU

Experiment

Comparison of first-order vs. second-order query expansion

Dataset	Baseline	First-Order	Second-Order
AOL	0.806	0.825 (0.007)	0.913(0.002)
Live Search	0.809	0.826 (0.006)	0.901(0.004)

Reference

1. Dai, H.K., Zhao, L, Nie, Z., Wen, J.R., Wang, L. and Li, Y., Detecting online commercial intention (oci), WWW, 2006
2. Hu, D.H., Yang, Q. and Li, Y., An algorithm for analyzing personalized online commercial intention, ADKDD, 2008