CUIS at the NTCIR-12 MobileClick2 Task

Kwun Ping LAI#, Wai LAM#, Lidong BING^

#Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong
^Machine Learning Department, Carnegie Mellon University

Presenter: Kwun Ping LAI
Introduction

- **Problems of IR in mobile platform**
  - small screen size
  - slow network speed and computing power

- **MobileClick Solution**
  - iUnit ranking
  - iUnit summarization
iUnit Ranking

- **Input**
  - a query
  - a list of iUnits
  - a collection of retrieved HTML documents

- **Output**
  - a list of iUnits with ranking
iUnit Ranking Model

1. Web content extraction
2. Intent discovery
3. Web document re-ranking
4. iUnit ranking
Web content extraction

- Aims at removing unimportant parts such as menu, navigation bar, scripts, etc.

- Tag-based
  - paragraph (p tag)
  - title (h tag)
  - cell in table (td tag)
  - list (li tag)

- Text feature-based
  - Boilerpipe

Intent discovery

- Use Latent Dirichlet Allocation (LDA) to discover hidden intent
  - Train a LDA model using the Web content documents
  - Rank the intents based on $P(i|q)$
    - $q$: query
    - $i$: intent
  - Query folding-in
    - $q = \{q_1, q_2, \ldots, q_k\}$, where $k = |q|$
    - $P(i|q)$ is inferred by the LDA model
    - $P(q_j|i)$ is inferred by the LDA model
    - Set a uniform prior
Intent discovery

- Random sampling of the latent topic $n$ times for each $q_i$
  \[ P(i|q) = \frac{c_i}{n r} \]
  - $c_i$: total number of latent topics of the intent $i$ being sampled
  - $n$: number of sampling of the query
  - $r$: total number of terms of the query

- Inspired by the work of He et al. (2011)

Web document re-ranking

- Associate one intent to each Web document
  \[ i^* = \arg \max_{i \in I} P(i|d) \]
  - \( d \): document
  - \( i \): intent

- the Web document with same intent form an intent document group \( G_i \)

- Web document re-ranking using Round-Robin
  - document weighting
    \[ w_d = |D| + 1 - \text{ranking}(d) \]
  - \(|D|\): the total number of documents
  - \( \text{ranking}(.) \): the new ranking score of the document
Web document re-ranking

**Algorithm 1** Web document re-ranking

1: function RR-RANK(\(|D|, G, R_I\))
2: input: The total number of Web document \(|D|\); intent document group \(G\); intent ranking \(R_I\)
3: output: New intent-based ranking of all Web document
4: ranking = \(\emptyset\)
5: while \(|\text{ranking}| \neq |D|\) do
6:     for each \(i \in R_I\) do
7:         if \(G_i \neq \emptyset\) then
8:             ranking.add(\(G_i\).poll())
9:         end if
10:     end for
11: end while
12: return ranking
13: end function
iUnit Ranking

- Importance measure $I(u,d)$
  - The importance of iUnit $u$ in each document $d$

$$I(u, d) = \frac{\sum_{w \in u} exist(w, d)}{|u|}$$

- $exist(w, d)$: indicator function denoting the existence of the word $w$ in the document $d$

$$exist(w, d) = \begin{cases} 
1 & w \in d \\
0 & w \notin d 
\end{cases}$$

- Overall score

$$score(u) = \sum_{d \in D_q} w_d I(u, d)$$

- $u$: iUnit, $w_d$: document weighting of the document $d$
Result of iUnit Ranking Model

iUnit Ranking Results

- Baseline method: 0.89
- CUIS(tag-based): 0.91
- CUIS(text feature-based): 0.9
iUnit Summarization

**Input**
- a query
- a list of iUnits
- a collection of retrieved HTML documents
- a list of intent labels

**Output**
- a two-layered structure
iUnit Summarization Model

Constructing 1\textsuperscript{st} layer

Associate web document and iUnit to intent label

Constructing 2\textsuperscript{nd} layers
Constructing 1\textsuperscript{st} layer

- Concatenate the iUnits in the order obtained in iUnit ranking result until reaching character limit (420 characters)
- Append all intent labels at the tail
Associate Web document and iUnit to intent label

**Web document containing words of intent label**

- Probability of the intent given the Web document

\[ P(i|d) = \frac{f(i, d)}{\sum_{i' \in I_q} f(i', d)} \]

- The intent importance in the Web document

\[ f(i, d) = \min_{w \in i} tf(w, d) \]

**Web document not containing words of intent label**

\[ P(i|d') = \frac{\sum_{d'' \in D_i} \text{sim}(\nu_{d'}, \nu_{d''})}{|D_i|} \]
Constructing 2\textsuperscript{nd} layers

- Construct a layer for each intent
- Concatenate the iUnits corresponding to the intent in the order obtained in iUnit ranking result
Result of iUnit Summarization Model

iUnit Summarization Results

CUIS | Baseline method | TITEC

15 | 16 | 17
THANK YOU!