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

Introduction

MobileClick2 task aims at solving information retrieval problem specifically in mobile platform. Returning a list of links relevant to the query is not suitable because of two reasons: 1) the small screen size of mobile platform is not suitable for displaying large amount of information and 2) the slow network speed and computing power of mobile platform greatly increase the time burden of users visiting the links one by one to locate the final desired information. The two subtasks, namely, iUnit ranking and iUnit summarization address the two problems stated above. The iUnit ranking subtask focuses on ordering the iUnits (information unit). Displaying only the top ranked iUnits can solve the small screen size problem. The iUnit summarization subtask requires a two-layered structure with a brief summary and link of intents (interpretation) in the first layer and detailed description of different intents in the second layer. Users can easily visit a particular interested intent.

iUnit Ranking Model

Web content extraction

We aim at extracting useful Web content from the visible components. We use tag-based method and text feature-based method to achieve this work.

For **tag-based method**, we consider those HTML tags with which useful content is associated. The tags include **paragraph (p tag), title (h tag), cell in table (td tag) and list (li tag)**. The text inside these tags is extracted and is saved to a plain text document.

For **text feature-based method**, we use Boilerpipe [3] package which exploits shallow text features such as the number of words and link density to distinguish the actual content from boilerplate text.

Intent discovery

We use Latent Dirichlet Allocation(LDA) [1] to discover hidden intents. Our intent discovery approach is inspired by the work of He et al.[2] which finds facets for search result diversification.

The next step is to rank the set of intents based on the latent topic distribution of the query(**P(i|q)**). We treat the query as a short document and use the trained LDA model to get the inferred latent topic distribution.

By randomly sampling the latent topics for each word in the query, we get a distribution of the latent topics whose normalized form represents the latent topic distribution of the query.

$$P(i|q) = \frac{c_i}{nr}$$

Web document re-ranking

We attempt to associate one intent for each Web document. It is assigned to an intent **i*** in which the document has the highest topic probability.

$$i^* = \arg\max_{i \in I} P(i|d)$$

The documents are re-ranked using a round-robin manner strategy. At each round, the top Web document of all intent document groups is picked according to the intent order and then added to a list. We set a **document weighting** for each document:

$$w_d = |D| + 1 - ranking(d)$$

iUnit ranking

We design an **importance measure I(u,d)** to capture the importance of the iUnit **u** in the Web document **d**:

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

Using the **document weighting** and iUnit **importance measure**, we calculate the iUnit score of each iUnit:

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

Finally, all iUnits are ranked using this score.

iUnit Summarization Model

Associate Web document and iUnit to intent label

We conduct intent label keyword search on the collection of Web content documents. The Web document is assigned to an intent document group $\mathbf{D_{i*}}$ of the intent **i*** in which the Web doucment has the highest probability.

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

f(i,d) is the importance of the intent **i** in the Web document **d**:

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

For the remaining Web documents with no words of intent label inside the content, the term weight **tf-idf** is adopted for calculating the **cosine similarity** between themselves and each Web document in each intent document group. They are assigned to the intent document group with the highest average similarity:

$$P(i|d') = \frac{\sum_{d'' \in D_i} sim(\boldsymbol{v}_{d'}, \boldsymbol{v}_{d''})}{|D_i|}$$

Constructing two-layered structure

We construct the first layer by concatenating iUnits based on the ranking obtained from the iUnit ranking subtask until reaching a length limit of 420.

The intent label that links to the second layer of each intent is placed at tail.

We construct a second layer for each intent. The corresponding iUnits for the intent layer is placed again in the order obtained in the iUnit ranking subtask.

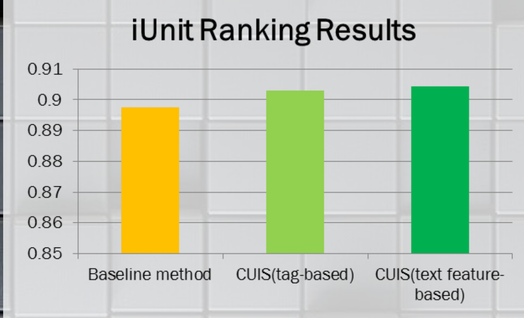
Results

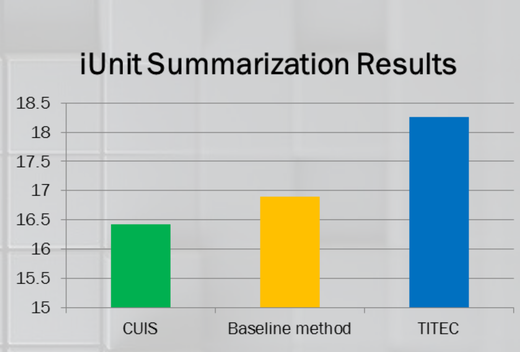
iUnit Ranking

Using the tag-based method for extracting Web content achieves a score of 0.903. The score improves to 0.9042 using the text feature-based method while keeping the same settings. Both scores are slightly higher than the best baseline method which has a score of 0.8975.

iUnit Summarization

The score of our iUnit summarization model is 16.4195. It is close to the best base-





References

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

[2] He, Jiyin, Edgar Meij, and Maarten de Rijke. "Result diversification based on query-specific cluster ranking." *Journal of the American Society for Information Science and Technology* 62.3 (2011): 550-571.

[3] Kohlschütter, Christian, Peter Fankhauser, and Wolfgang Nejdl. "Boilerplate detection using shallow text features." *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010.