NUTKS at NTCIR-12 MobileClick2: iUnit Ranking Subtask Using Topic Model

Tatsunori Yoshioka
Nagaoka University of Technology
1-1603, kamitomiokacho, Niigata, Japan
s133356@stn.nagaokaut.ac.jp

Takashi Yukawa
Nagaoka University of Technology
1-1603, kamitomiokacho, Niigata, Japan
yukawa@vos.nagaokaut.ac.jp

ABSTRACT

In this paper, NUTKS (Nagaoka University of Technology, Knowledge Systems Laboratory) reports the results of our participation at the NTCIR-12 MobileClick task, iUnit Ranking subtask. The authors have ranked iUnit by using similarity of iUnit word distribution based on LDA topics.

Our system has recorded Q-measure score as 0.7392 and nDCG@20 score as 0.6334. In baseline system, the recorded Q-measure score is 0.7411[1]. Therefore, there is no significance difference between the baseline system and our proposed approach.

According to these results, the search intent is considered as topics. However, it cannot output result of reflecting search intent diversity only using topic similarity. it is not strong enough to reflect the diversity of search intent by using only the topic similarity.

Team Name

NUTKS

Subtasks

iUnit Ranking Subtask(Japanese)

Keywords

Information Retrieval, Topic Model, Mobile Search

1. INTRODUCTION

This paper describes our work that aimed direct information summarizing in NTCIR-12 MobileClick2 task that followed by ranking element of summary in Unit ranking subtask.

NTCIR-12 MobileClick2 task is compared with the NTCIR-11 MobileClick task, which was noted as ambiguous or underspecified query [1]. Therefore, it was difficult to search the intent form the given query only. Then participants have to estimate the search intent and search intent probability with query and iUnits and Document. Our approach has assumed that LDA topics is a search intent. Moreover, which we assumed that topics distribution ratio is search probability.

The reason for using Latent Dirichlet Allocation (LDA) is to estimate query of latent by its algorithm

For intent estimation, a usual system recognizes a type of query [2]. Our approach is different. Our system estimates user's intents without query type recension. Instead it estimates the intention directly using the LDA topics.

2. BACKGROUND

2.1 Latent Dirichlet Allocation

LDA is one of Language model, which expects a document generated by many "Topic" Distribution. Moreover, a topic is comprised of *words probability distribution* and a word comprised of latent topics. [4].

This algorithm has developed by Blei et al. [5]

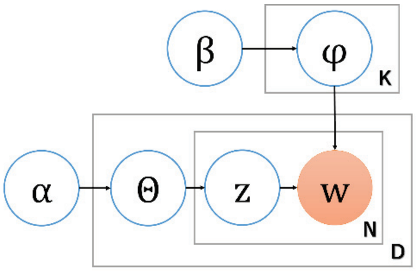

Figure 1: Graphical model representation

Figure 1. is the graphical model of LDA. Colored circle w is a word in a document. The colorless circle z is topic, θ is topic distribution, φ is word distribution, α and β is hyper parameters. K is the number of topics, D is the number of documents and N is the number of words.

Word in a document is generated according to following procedure; [4]

Firstly, word distribution of each topic generated in accordance with the Dirichlet distribution of β , which stated as φ. Secondly, topic distribution of each document generated in accordance with the Dirichlet distribution of α , as θ. Then, the topic of each word generated in accordance with the multinomial distribution of θ, stated as z. Finally, word of each word generated in accordance with the multinomial distribution of z

2.2 Previous works

In NTCIR-11 MobileClick task iUnit Retrieval Subtask, IISR has performed rule-based query classification. Their approach has classified into eight types of queries [3]. They have discriminate query types from contained words in a query: e.g. "why"," vs", "benefit" …etc. Their study has recorded highest performance. [2]

Further, Japanese iUnit ranking has not attempted in NTCIR-11

In NTCIR-10 1CLICK-2 task, NUTKS (Not our team) has performed snippet (Web page overview from search engine) similarity based on query classification. Their approach has used snippets similarity to feature of support vector machine. They have recorded one of the highest performance. [6]

3. LDA BASED iUnit RANKING

3.1 Our Approach

A query has a search intent and consist in the iUnits and document. If the meaning of iUnit and document can be analyzed, meaning can be assumed for search intent. Therefore, it is possible to estimate search intent from topic model. Consequently, we expected, it is possible to rank iUnit with iUnit to topic cosine similarity.

LDA expresses many latent meaning from word distribution. we have assumed many latent meaning on one word in Japanese queries compared to English. Therefore, our approach has followed LDA

3.2 System flow

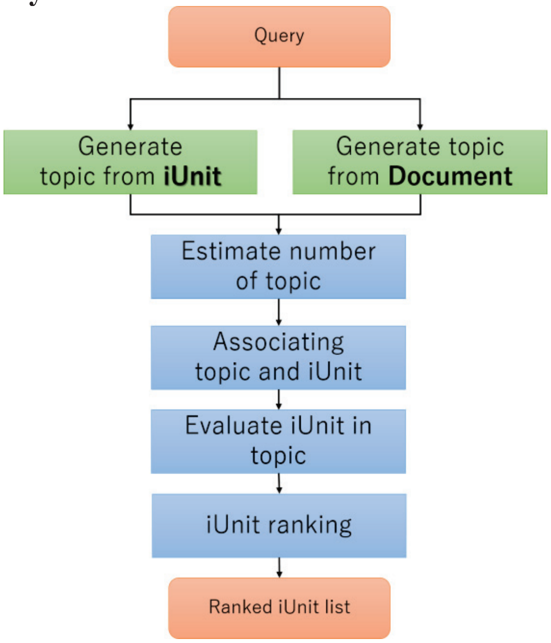
Query
Generate topic from **iUnit**
Generate topic from **Document**
Estimate number of topic
Associating topic and iUnit
Evaluate iUnit in topic
iUnit ranking
Ranked iUnit list

Figure 2: System overview

Our system ranks iUnits as shown in figure 2. Firstly, our system generates a topic that the corpus based on iUnits, correspond to the query and it generates a topic that the corpus based on documents (Search result of Bing) correspond to query. Secondly, our system estimates the number of topics according to the method described below. Thirdly, our system is to relate topic to iUnit and using that calculate similarity to iUnit word distribution and iUnit based topic. Then, our system calculates iUnit evaluation scores with the evaluation formula described below. Lastly, our system rank the iUnit using estimated topic ratio and iUnit evaluation scores.

3.3 Estimate The Number of Topics

LDA algorithm requires the number of topics. However, a query has different the number of search intents by each queries. Our approach considered the search intent and topics. Therefore, the study has to estimate the number of topics by each query.

We have assumed that the number of search intents in each query are six or less. Then, we executed LDA as each parameter consist of the number of intent from 1 to 6. And each topics extracts the top ten words from the topic distribution. After that, we generated a sparse vector based on extracted top ten words representing all corpus words. We called this as “topic feature value”.

Then, the cosine similarity for topic to topic using topic feature value for each the number of topic was calculated. The Cosine similarity was 0.1 or higher than 0.1. Then, it presumes which is same topic.

It is to execute in reverse order. If the given the number of topics to LDA and executed result the number of topic are equal, we assume that it is best the number of topics of query. And we use this LDA result for our proposed system.

3.4 Evaluate iUnit Gain

The iUnit evaluation value for each topic is determined by the following evaluation formula.

$$\mathrm{iUnitWeight} = \mathrm{sim_t}(i) \times doc\_topic\_score(i) \qquad (1)$$

sim_t(i) is cosine similarity of topic based on iUnits corpus and iUnit word distribution. doc_topic_score(i) is weighted cosine similarity of topic based on documents and iUnit word distribution. doc_topic_score(i) is determined by the following evaluation formula.

$$\mathrm{doc\_topic\_score} \ = \ \mathrm{sim_t}(\sum_{k=1}^{@30} \frac{sim_d(doc_k)}{\log_2 k}) \qquad (2)$$

k of formula represents document ranks in search result and $\mathrm{doc_k}$ is k the rank document. $\mathrm{sim_d}(doc_k)$ is topic based on documents and iUnit word distribution.

Formula (2) states that the topic appearing mostly in highly rank document is important topic based on the document. However, the topic appearing less in highly ranked document is not important topic based on document.

3.5 iUnit Ranking

Firstly, the iUnit belonging to topic based on iUnits ranks by using result of 3.3iUnit evaluation formula. Secondly, performed according to a lot of frequency topic based on iUnit in ascending order and obtained the results. Since we have assumed that our study considered few of frequency topic based on iUnit, there is less possibility of having cumbersome and noisy information. Thus, it has created important value on iUnits.

4. RESULT

Our system has obtained the following results.

Table 1 Results of NUTKS’s MobileClick runs

Run-ID	Q-Measure	nDCG@5	nDCG@20
218	0.7718	0.5564	0.6997
219	0.7197	0.4713	0.6112
355	0.7716	0.5620	0.6985
356	0.7392	0.5097	0.6334
ORG-L	0.7269		
ORG-R	0.7411		

From Table1, Run-ID 356 is our study obtained result with test data. Run-ID 255 is following same approach with training data. RunID-219 is with test data that is not using doc_topic_score which use only cosine similarity from iUnit to topic based on iUnits. RunID-218 is same as RunID-219’s approach but comprise with training data.

Table 2 Results of search intent estimation

	Average	Mode	Standard deviation
Intent	4.4141	2.0	2.9451
Document based topic (number=6, Threshold=0.1)	3.0404	3.0	0.9359
iUnit based topic (number=6, Threshold=0.1)	3.1515	4.0	0.9515
Document based topic (number=10, Threshold=0.3)	5.6970	6.0	1.15350
iUnit based topic (number=10, Threshold=0.3)	5.556	5.0	1.4302

From Table2, the number of search intent and the number of topic, both average values are approximately same in our system results as well as both modes state approximately same values. When the number of topic and threshold values are increased, the average values also remain same. However, mode is very different.

5. DISCUSSION

Our system has not classified to the query in this task and thereby we do not have weight to query, depending on query characteristic. Therefore, our system has evaluated iUnit using only iUnit to topic similarity and topic frequency. However, Q-Measure is 0.7392. It is considerable result compared to baseline systems values. Upon this value, our approach is still valid in iUnit ranking subtask. Even though baseline system has obtained 0.7411 value for Q-measure, result of our system is also more close to this value.

According to the results, our system is not performing well enough to evaluate iUnit ranking. This task must result in decentralizing iUnit to correspond to search intent. In our system, it centralizes to iUnit to correspond to topic. Therefore, it provides results of noisy and unimportant iUnit at high rank. Moreover, our system does not permit in iUnit duplication of topic based on iUnit. Thus, iUnit with many search intents consider one search intent. Intrinsically important iUnit has appeared after appearing unimportant iUnit. Therefore, our system result is mostly same as baseline system.

Our approach used similarity iUnit word distribution to topic based on document in iUnit evaluation. It affects Q-measure for value of 0.02 to use or not. Therefore, frequency of topic based on iUnits, dominates in our approach, while similarity of iUnit word distribution and topic based on document valid to iUnit evaluation.

Our approach considers search intent for topics. In light of the result, it is the number of topics and search intent approximately same. Moreover, in increasing the number of topic and threshold cause in having same result too. However, mode is different the estimated the number of topics process in decreasing the number of topics large to small amount of topics. Therefore, it ended up with large the number of topics.

6. CONCLUSION

In this paper, we discuss NUTKS result of participation in the NTCIR-12 MobileClick2 task iUnit Ranking subtask.

Our system with iUnit evaluation formula has premised that search intent and iUnit is one-to-one correspondence. Moreover, results are more centralized. Consequently, our result does not cooperate with ranking results of intent diversity. Therefore, iUnit evaluation formula should improve to reflect search intent diversity in iUnit ranking. Moreover, our approach should modify with query classification that is depending on topic weight and it gives better result of iUnit ranking.

7. ACKNOWLEDGMENT

It was very useful the software gensim(topic model library for python), MeCab (Japanese Morphological analyzer) and neologd(dictionary for MeCab). Thus we express our gratitude to each of the software authors.

8. REFERENCES

[1] Makoto P. Kato, Tetsuya Sakai, Takehiro Yamamoto, Virgil Pavlu, Hajime Morita and Sumio Fujita: Overview of the NTCIR-12 MobileClick Task, Proceedings of NTCIR-12, 2016.

[2] Makoto P. Kato, Matthew Ekstrand-Abueg, Virgil Pavlu, Tetsuya Sakai, Takehiro Yamamoto and Mayu Iwata: Overview of the NTCIR-11 MobileClick Task, Proceedings of NTCIR-11,2014

[3] Chia-Tien Chang, Yu-Hsuan Wu, Yi-Lin Tsai and Richard Tzong-Han Tsai: Improving iUnit Retrieval with Query Classification and Multi-Aspect iUnit Scoring The IISR System at NTCIR-11 MobileClick Task, Proceedings of NTCIR-11, 2014

[4] Tomoharu Iwata: Topic Models (Japanese), Kodansha, 2015

[5] David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993-1022, 2003

[6] Tatsuya Tojima and Takashi Yukawa: Query Classification System Based on Snippet Summary Similarities for NTCIR-10 1CLICK-2 Task, Proceedings of NTCIR-10, 2013