KUIDL at the NTCIR-11 IMine Task

Takehiro Yamamoto, Makoto. P. Kato, Hiroaki Ohshima, Katsumi Tanaka
Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan
+81-75-753-5969
{tyamamot, kato, ohshima, tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT

The KUIDL team participated in the Subtopic Mining subtask of the NTCIR-11 IMine task. This paper describes our approach to generating two-level hierarchical subtopics by using Web document structures. The formal run result shows that our approach achieved the best performance in terms of *H*-measure in the English Subtopic Mining subtask.

Team Name

KUIDL

Subtasks

Subtopic Mining (English, Japanese)

Keywords

document structure, diversification

1. INTRODUCTION

Kyoto University, Department of Informatics, Digital Library laboratory (KUIDL) participated in the NTCIR-11 IMine task. IMine refers to a task that explores and evaluates the technologies of satisfying different user intents behind a Web search query by mining subtopics and generating diversified search rankings [1]. The KUIDL team participated in the Subtopic Mining subtask. In this subtask, the system is expected to return a two-level hierarchy of underlying subtopic of a given topic. To tackle this problem, we propose a method of using the Web documents as a resource to obtaining intents and extracting hierarchical intents by using document structures.

2. Method

This section first presents the overview of our method and then describes the details of each step.

2.1 Overview

Figure 1 illustrates the overview of our method. Let q denote the given topic. Our system works as follows:

1. The system issues query q to a Web search engine and obtains the top n documents $D = \{d_1, \ldots, d_n\}$. In this work, the system obtains the top 500 search results by using the Bing API and obtains the corresponding documents.
2. The system clusters the documents and obtains k clusters.
3. The system ranks the clusters by using the MMR-based algorithm and selects five clusters.
4. For each selected cluster, the system extracts a first-level subtopic and 10 second-level subtopics from the documents in the cluster.
5. The system generates the output according to the first-level and second-level subtopics extracted in the step 4.

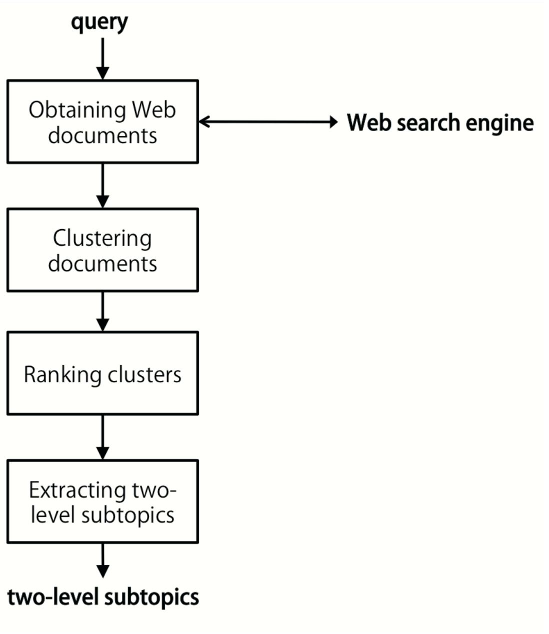

Figure 1. Overview of our method.

The rest of this section we describe the details of steps 2, 3, 4, and 5.

2.2 Clustering Documents

After obtaining a set of documents D from a search engine, the system first tries to find possible intents of a given topic. To this end, the system first groups documents D into k clusters. By applying the k-means algorithm to documents D, the system obtains a set of clusters $\mathcal{C} = \{C_1, \ldots, C_k\}$, where k denotes the number of clusters. Cosine similarity is employed when the system computes the distance between two documents. Each cluster consists of the documents grouped into the cluster. Note that $\bigcup_{C_i \in \mathcal{C}} = D$. In this work, the system generates 20 clusters.

2.3 Ranking Documents

After obtaining the set of clusters $\mathcal{C}$, the system selects the five clusters, which is the expected number of first-level subtopics in the Subtopic Mining subtask, for later subtopic extraction step. Since we want to extract diverse subtopics while keeping their importance, the system tries to select important and diverse five clusters from $\mathcal{C}$. To this end, the system applies the MMR-based algorithm [2] as described below:

$$MMR = \operatorname*{argmax}_{C_i \in \mathcal{C} \setminus S} \left[\lambda \, \mathrm{Score}(C_i) - (1 - \lambda) \max_{d_j \in S} \mathrm{Sim}(C_i, C_j) \right].$$

Table 1 Overall Results of Japanese Subtopic Mining subtask. Highest value among participating teams is shown in bold.

	Hscore	*Fscore*	*Sscore*	*H-measure*
KUIDL-S-J-1A	**0.2702**	0.2883	0.2848	0.0852

Table 2 Overall Results of English Subtopic Mining subtask. Highest value among participating teams is shown in bold.

	Hscore	*Fscore*	*Sscore*	*H-measure*
KUIDL-S-E-1A	**0.9190**	0.5670	0.5964	**0.5627**

Where $\text{Score}(C_i)$ is a score of cluster and is calculated as the number of documents in the cluster, S is a subset of $\mathcal{C}$ that are selected as output, $\mathcal{C}\backslash S$ is a subset of clusters in $\mathcal{C}$ that has not yet selected as the output, and $\text{Sim}(C_i, C_j)$ gives the similarity between clusters C_i and C_j. λ is a parameter that balances the importance of clusters and their diversity. To calculate $\text{Sim}(C_i, C_j)$, we treat the maximum distance between two documents in two clusters, C_i, C_j as the similarity between clusters. The set of the selected five clusters is hereafter denoted as $\mathcal{C}^* = \{C_1^*, \dots, C_5^*\}$. In this work, we set $\lambda = 0.5$.

2.4 Extracting Hierarchical Subtopics

After obtaining $\mathcal{C}^*$, the system obtains a first-level subtopic and 10 second-level subtopics for each cluster.

First, the system obtains first-level subtopic of a cluster C_i^* $(1 \leq i \leq 5)$. To exract first-level subtopis, we hypothesize that important subtopic appear in the *titles* of documents rather than the *bodies* of documents. Let $df_{title}(t)$ be the number of documents that contain term *t* in a cluster. The system extracts the top term according to the value of $df_{title}(t)$.

After obtaining a first-level subtopic, the system extracts second-level subtopics. To this end, we adopted the method proposed by Oyama *et al* [3]. They proposed a method of obtaining related terms of a given term by using the structures of Web documents. Their method relies on the idea that terms appear in the title of documents are likely to represent the overall subject while terms appear in the body of documents are likely to represent the detailed topic of the subject. Their method also takes into account the positions of the terms in the document structures when counting their occurrences.

Let t^{first} be the extracted first-level subtopic of a cluster and t^{second} be a candidate of second-level subtopic in a cluster. To obtain the second-level subtopics, we examine the difference of two term co-occurrence rates $P(t^{\text{second}} \mid t^{\text{first}})$ and $P(t^{\text{second}} \mid \text{intitle}(t^{\text{first}}))$ in the cluster. $P(t^{\text{second}} \mid t^{\text{first}})$ denotes the probability that term t^{second} appears in the documents that contain term t^{first} in their titles or bodies, while $P(t^{\text{second}} \mid \text{intitle}(t^{\text{first}}))$ denotes the probability that term t^{second} appears in the documents that contain term t^{first} in their titles. If the value of $P(t^{\text{second}} \mid \text{intitle}(t^{\text{first}}))$ is significantly larger than the value of $P(t^{\text{second}} \mid t^{\text{first}})$, term t^{second} is likely to be a second-level subtopic of t^{first}. To estimate the statistical significance of the difference between the two rates, we use the χ^2 test and compute the χ_0^2 value to find the second-level subtopics. We extract the top ten terms according to the value of χ_0^2 and treat them the second level subtopic of t^{first}.

2.5 Output Two-Level Subtopics

From the method described in the above sections, the system obtain first-level subtopic t_i^{first} and 10 sepcond-level subtopics t_i^{second} for each cluster C_i^*. To generate the output, we simply concatenate the topic and the first-level subtopic and the second-level subtopic which white spaces. For example, if the topic is "Apple" and first-level and second-level subtopics are "computer" and "iPhone 5", "Apple computer iPhone 5" will be generated as an output.

3. Evaluation Results

In this section we first describe the settings of our method and then present our results.

We participated the Japanese and English Subtopic Mining subtasks. For each subtask, we submitted one run, namely KUIDL-S-J-1A and KUIDL-S-E-1A, as the official runs.

Tables 1 and 2 show the overall results of Japanese and English Subtopic Mining subtasks, respectively. The tables shows the results of four metrics: *Hscore*, *Fscore*, *Sscore* and *H*-measure [1].

As shown in Tables 1 and 2, our methods achieved the highest *Hscore* in both Japanese and English subtasks. *Hscore* measures the quality of the hierarchical structure by whether the second-level subtopic is correctly assigned to the appropriate first-level subtopic. These results indicate that our method that utilizes document structures to obtain second-level subtopics has work well to find hierarchical intents.

Our method also achieved the highest *H*-measure in the English Subtopic Mining subtask. Since *H*-measure is the combination of *Hscore*, *Fscore* and *SScore*, our method got the highest *H*-measure in the English Subtopic Mining subtask due to the high *Hscore*.

4. Conclusion

In this paper, we presented our method of generating two-level hierarchical subtopics. Our method employs the documents obtained from a Web search engine and extracts hierarchical subtopics by using document structures..

5. REFERENCES

[1] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. Kato, H. Ohshima and K. Zhou. Overview of the NTCIR-11 IMine Task. In *NTCIR-11*, 2014.

[2] J. Carbonell and J. Goldstein. The use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335-336, 1998.

[3] S. Oyama and K. Tanaka. Query modification by discovering topics from web page structures. In *Proceedings of the 6th Asia Pacific Web Conference*, pp.553-564, 2004.