IRCE at the NTCIR-12 IMine-2 Task

Ximei Song
University of Tsukuba
songximei@slis.tsukuba.ac.jp

Yuka Egusa
National Institute for Educational Policy Research
yuka@nier.go.jp

Hitomi Saito
Aichi University of Education
hsaito@auecc.aichi-edu.ac.jp

Masao Takaku
University of Tsukuba
masao@slis.tsukuba.ac.jp

ABSTRACT

The IRCE team participated in the IMine-2 task at the NTCIR-12 workshop. We submitted one Chinese language run and five Japanese language runs for the Query Understanding subtask. Our methods exploited online text corpora BaiduPedia for the Chinese language run and Japanese Wikipedia for the Japanese language runs. The approaches employed in the Chinese and Japanese language topics are differed. This paper discusses our approaches to the Query Understanding subtask of the NTCIR-12 IMine-2 task.

Team Name

IRCE

Subtasks

IMine-2 (Chinese and Japanese)

Keywords

query understanding, intents, search results diversification

1. INTRODUCTION

Search results diversification has recently emerged as a research topic [4]. Users might have different reasons for a search even when they have submitted the same query, and users might seek different interpretations for an ambiguous query. Moreover, users might be interested in different subtopics of multi-faceted topics. Result diversification deals with ambiguous or multi-faceted queries by providing documents that cover as many subtopics of a query as possible. In the NTCIR-12 IMine-2 task [8], we proposed methods for diversifying search results and experimented with evaluation metrics to measure diversity.

The IRCE (Information Retrieval Cognitive Study Evaluation) team participated in the IMine-2 task of the NTCIR-12, which included the Query Understanding (QU) subtask for the Japanese and Chinese languages. Our goals were to identify the relationship between the diversification and user behavior and how it takes effect (or not) for user study [1]. In the QU subtask, our goal was to acquire 10 subtopics for given queries that were ambiguous and broad. We used different methods for the Chinese language and the Japanese language in the QU subtask.

2. RELATED WORKS

Previous NTCIR workshops targeted several tasks regarding intents: NTCIR-9 INTENT task, NTCIR-10 INTENT2 task, and NTCIR-11 IMINE task. In NTCIR-9 [6], 42 Chinese language runs were submitted from 13 teams and 14 Japanese language runs were submitted from five teams. In NTCIR-10 [5], the Subtopic Mining subtask had 23 Chinese language runs from six groups and eight Japanese language runs from two groups. In NTCIR-11 [3], the Subtopic Mining subtask received 19 Chinese language runs from five groups and five Japanese language runs from two groups.

The FRDC team [9] at the NTCIR-11 IMine Task had two strategies for the subtasks of Subtopic Mining for the Chinese language. One was effective for finding first-level subtopics of ambiguous queries (high F-score), but it failed to demonstrate high S-score and H-score performances. It was based on document-clustering technology and no external knowledge was involved. The document-clustering method was a clustering of the candidate queries to obtain second-level subtopics followed by generation of the first-level subtopics based on the second-level results. The method used the following four steps: (1) clustering using the open sources toolkit *Cluto*, (2) refining the clustering result using the LDA model [2], (3) selecting the optimal clustering result, and (4) generating the first-level subtopics. The second strategy attained a high H-score. It used BaiduPedia as the knowledge base and employed document clustering and classification technologies. This method used the following four steps: (1) classifying using BaiduPedia, (2) document clustering by threshold-based clustering method, (3) merging the classification with the clustering results, and last (4) ranking the subtopics.

We employed different approaches than the related works above. As described below in Sections 3 and 4, we used text corpora of BaiduPedia for the Chinese language topics and Japanese Wikipedia for the Japanese language topics.

3. QUERY UNDERSTANDING (QU) FOR THE CHINESE LANGUAGE

3.1 Dataset

In the QU subtask, we adopted a strategy that uses BaiduPedia as the knowledge base for the Chinese language topics. Because BaiduPedia does not provide bulk download files, we used BaiduPedia only as an online resource. We collected the contents from BaiduPedia and extracted the text via a scraping technique.

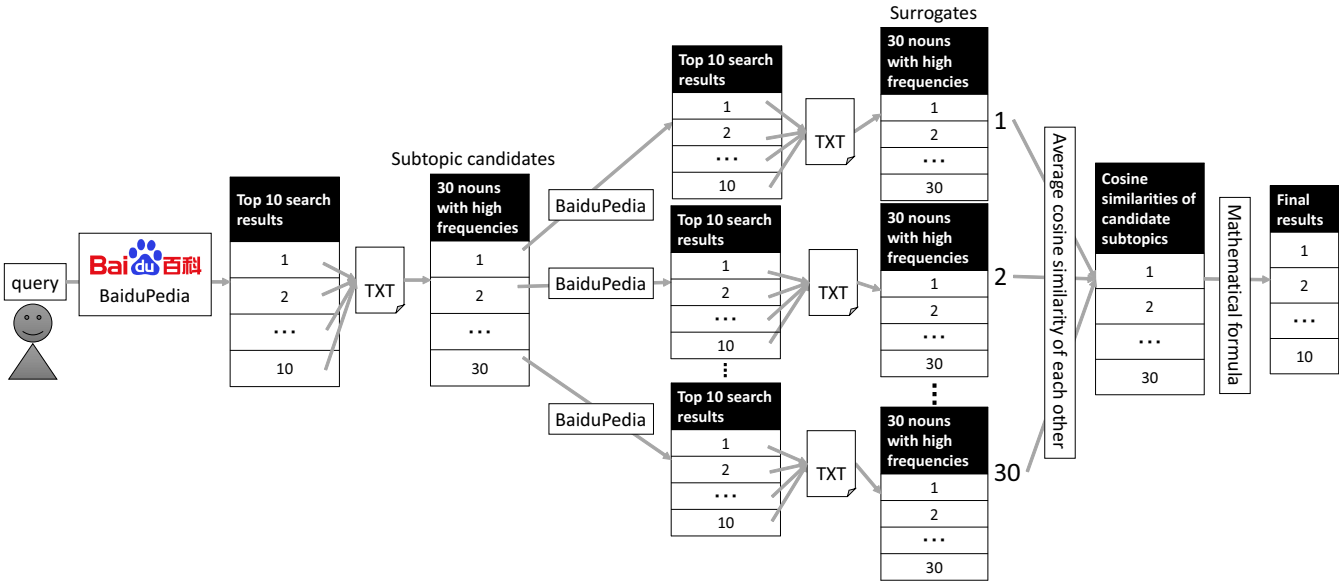

Figure 1: Subtopic extraction process for the Chinese language topics

3.2 Methods

Our method exploits BaiduPedia as the knowledge base. Figure 1 shows the overall process used for the Chinese language topics. The method initially retrieves the top 10 articles from the BaiduPedia search results in response to a query. First, the top 10 articles of the BaiduPedia search results were combined and treated as a single document. Word segmentation was performed using the *jieba* tool[1]. After stopwords were excluded, we selected 30 nouns with high frequencies in a document as the subtopic candidates.

Second, for each subtopic candidate, the top 10 articles in the BaiduPedia search results were again obtained and combined to create a single document. After word segmentation, we converted that document into a word vector representation. This word vector representation comprised 30 nouns with high frequencies in a document, which was treated as a surrogate of a subtopic candidate. Using the TF-IDF method, a surrogate's score could be calculated.

Third, the cosine similarities [7] of the surrogates of a subtopic candidate were calculated. The average cosine similarity of a subtopic candidate was treated as the score of the subtopic candidate. The subtopic candidate with the highest score is the one that was the most similar to the other subtopic candidates. To diversify the results, we assumed that lower scores were better final results. Therefore, we applied the mathematical formula of one minus the average similarity as the final score. Last, we chose the top 10 words from the subtopic candidates using the descending order of the final scores as the final results.

4. QUERY UNDERSTANDING (QU) FOR THE JAPANESE LANGUAGE

4.1 Datasets

For the Japanese language runs, we used the dataset of the Japanese Wikipedia data dump files[2] as of December 17, 2013.

The dataset was indexed to be searchable for article title, fulltext, article titles that redirected to the article, and category titles that attached to the article.

The retrieval system was implemented using the Apache Solr engine and MySQL server[3].

4.2 Methods

Figure 2 shows the overall process that used for the Japanese language runs. First, we generated a ranked list of articles with weights on a given topic query. In this search process, the query was executed to the indices of the fields of the article title, fulltext, article titles that redirected to the article, and the category titles that attached to the article. The cosine similarities between a query and each field of an article were computed as field scores. Then, a score for an article was computed with the weighted totals of the similarity scores as follows:

$$Score_{article} = S_{title} \times W_{title} + S_{redirect} \times W_{redirect} + S_{category} \times W_{category} + S_{text} \times W_{text}$$

where S_{field} is a similarity score of a field, and W_{field} is a weighting value of a field. We used weights $W_{title} = 10$, $W_{redirect} = 10$, $W_{category} = 10$, and $W_{text} = 1$ in the submitted runs. The top 100 articles of a given query were retrieved. We assumed that the categories attached to the articles were subtopic candidates for the given query.

Second, we converted the $Score_{article}$ of each article into the $Score_{category}$ of a subtopic candidate. When multiple categories were assigned to an article, the original score was divided by the number of categories. $Score_{category}$ values were accumulated and reordered for each category in descending

[1]https://github.com/fxsjy/jieba

[2]http://dumps.wikimedia.org/jawiki/

[3]The implementation is available at https://github.com/cres-project/irce-wikipedia

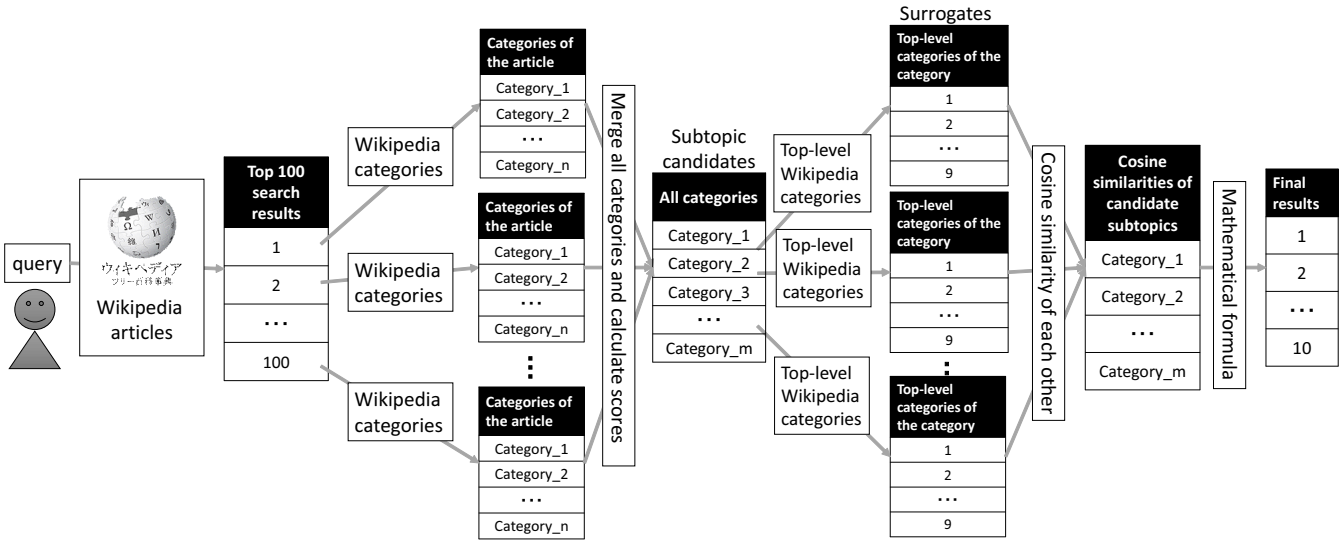

Figure 2: Subtopic extraction process for the Japanese language topics

order. We used the ranked list of categories (subtopic candidates) as baseline for the search results of an original given query.

Third, to diversify the results, we built a surrogate for each subtopic candidate. We used the following nine top-level categories in Japanese Wikipedia to build a surrogate:

1. 学問 (Academia)
2. 技術 (Technology)
3. 自然 (Nature)
4. 社会 (Society)
5. 地理 (Geography)
6. 人間 (Humans)
7. 文化 (Culture)
8. 歴史 (History)
9. 総記 (Generals)

Each category was converted into a weighted vector representation using the distance between the category and the top-level categories. In this process, we propagated a score value $Score_{category}$ to weight each surrogate using a hierarchical distance from the original category to a top-level category. We assumed that this vector representation was a surrogate of the category. The cosine similarity of a given pair of surrogates could be calculated.

Last, we extracted the final ranked lists. We used several strategies to generate the final ranked lists as follows.

Baseline.

The final ranked list was generated only using $Score_{category}$. The baseline run was submitted as IRCE-QU-J-5S.

Diversity.

The final ranking was selected using $Score_{category}$ and the cosine similarity through their linear combinations as follows.

$$Rank_1 = \max\{Score_{category}\}$$

$$Score(surrogate) = \max\{Similarity(surrogate, x), x = Rank_1, \cdots, Rank_i\}$$

$$Score_{final} = \alpha \times Score(surrogate) + (1-\alpha) \times Score_{category}$$

where $Rank_i$ is a ranked subtopic, and $Similarity(surrogate, x)$ is a cosine similarity between one surrogate and another surrogate x. $Score_{final}$ was calculated individually, and the subtopic candidates with the highest scores were chosen for the final ranked list. The parameter α represents the blending ratio of relevance (baseline) and diversification. This process continued until it reached ten subtopics.

The submitted runs IRCE-QU-J-1S, IRCE-QU-J-2S and IRCE-QU-J-3S were generated with the parameters $\alpha = 0.8$, 0.2, and 0.5, respectively.

Another variation of $Score(surrogate)$ was used as follows.

$$Score'(surrogate) = \sum_{x=Rank_1}^{i} Similarity(surrogate, x)$$

The submitted run IRCE-QU-J-4S was generated using this formula with the parameter $\alpha = 0.5$.

5. RESULTS AND DISCUSSION

Our team submitted one Chinese language run and five Japanese language runs.

5.1 Chinese language run

Table 1 shows the evaluation results of the Chinese language run.

Table 1: Evaluation results of the Chinese language run

Run ID	I-rec@10	D-nDCG@10	D#-nDCG@10
IRCE-QU-C-1S	0.4827	0.4290	0.4558

There are four topic types of IMine-2 topics: ambiguous, faceted, task-oriented, and vertical-oriented. The evaluations results of our run per topic type are shown in Table 2, which indicates that there were no significant differences among topic types.

Table 2: Evaluation results of the Chinese language run per topic type

Topic types	I-rec@10	D-nDCG@10	D#-nDCG@10
Ambiguous	0.4827	0.4085	0.4456
Faceted	0.4650	0.4249	0.4450
Task-oriented	0.4713	0.4103	0.4408
Vertical-oriented	0.4959	0.4437	0.4698

We selected the following four topics to conduct failure analysis of the Chinese language run: "**哀歌**" (IMINE2-C-006), "**白眉大侠单田芳**" (IMINE2-C-074), "**圣诞节怎么过**" (IMINE2-C-023), and "**爱回家粤语**" (IMINE2-C-066). The evaluation results of our run of these topics had the lowest values. In particular, the results of our run on the topics "**哀歌**" (IMINE2-C-006) and "**白眉大侠单田芳**" (IMINE2-C-074) were evaluated as 0.0 in D#-nDCG metrics.

Although the judged subtopics of the topic IMINE2-C-006 "**哀歌**" were songs, network novels, published books, songs' information, the Bible, and movies, the majority of the subtopics results from our run was dominated by a particular person's name. Just three subtopic candidates of the original 30 candidates from BaiduPedia covered a subtopic of **歌曲** (songs), another subtopic candidate covered a subtopic of **歌曲资源信息** (songs' information), and the others were not covered. Because these four subtopic candidates were similar to each other, they were ranked lower in the final results.

In the case of topic IMINE2-C-074 "**白眉大侠单田芳**", the judged subtopics were downloads, Chinese storytelling, videos, listening to recordings online, adapted dramas, and resources. Just one subtopic candidate of the original 30 candidates from BaiduPedia covered a subtopic of **评书** (Chinese storytelling).

In the case of topic IMINE2-C-023 "**圣诞节怎么过**", the judged subtopics were regions, methods, romances, lovers, decorations, event marketing, and gifts. The subtopic candidates for the topic from our run did not cover these subtopics at all.

In the case of topic IMINE2-C-066 "**爱回家粤语**", the judged subtopics were downloads, watching videos online, video tapes, and related information. The subtopic candidates for the topic from our run covered only the subtopic of **在线观看** (watching videos online).

In sum, these results indicate that obtaining candidates only from BaiduPedia seems to be insufficient. Future work should focus on approaches that add other resources, such as query suggestions.

5.2 Japanese language runs

Table 3 shows the evaluation results of the Japanese language runs.

As described in Section 4, we employed slightly different strategies and parameters to our Japanese runs. There seems to be no significant differences among them in terms of evaluation metrics. More diverse resources might be needed to achieve better results.

There are four topic types in the IMine-2 topics: ambiguous, faceted, task-oriented, and vertical-oriented. The evaluation results of our runs per topic type are shown in Tables 4, 5 and 6.

Table 3: Evaluation results of the Japanese language runs

Run ID	I-rec@10	D-nDCG@10	D#-nDCG@10
IRCE-QU-J-1S	0.4102	0.2706	0.3404
IRCE-QU-J-2S	0.4043	0.3167	0.3605
IRCE-QU-J-3S	0.3900	0.3300	0.3600
IRCE-QU-J-4S	0.4169	0.3100	0.3634
IRCE-QU-J-5S	0.3903	0.3387	0.3644

Table 4: Evaluation results of I-rec@10 of the Japanese language runs per topic type

Run ID	Ambiguous	Faceted	Task-oriented	Vertical-oriented
IRCE-QU-J-1S	0.4859	0.4404	0.2674	0.4471
IRCE-QU-J-2S	0.4822	0.4558	0.2674	0.4118
IRCE-QU-J-3S	0.5042	0.3964	0.2463	0.4131
IRCE-QU-J-4S	0.5150	0.4270	0.2754	0.4502
IRCE-QU-J-5S	0.4942	0.4084	0.2517	0.4068

Table 5: Evaluation results of D-nDCG@10 of the Japanese language runs per topic type

	Ambiguous	Faceted	Task-oriented	Vertical-oriented
IRCE-QU-J-1S	0.3693	0.2913	0.1781	0.2437
IRCE-QU-J-2S	0.4382	0.3625	0.2055	0.2606
IRCE-QU-J-3S	0.4808	0.3697	0.1907	0.2790
IRCE-QU-J-4S	0.4368	0.3499	0.1963	0.2568
IRCE-QU-J-5S	0.4922	0.3828	0.1983	0.2815

Table 6: Evaluation results of D#-nDCG@10 of the Japanese language runs per topic type

Run ID	Ambiguous	Faceted	Task-oriented	Vertical-oriented
IRCE-QU-J-1S	0.4233	0.3722	0.2376	0.3388
IRCE-QU-J-2S	0.4572	0.4178	0.2491	0.3362
IRCE-QU-J-3S	0.4894	0.3977	0.2265	0.3397
IRCE-QU-J-4S	0.4736	0.3954	0.2502	0.3535
IRCE-QU-J-5S	0.4901	0.4090	0.2343	0.3305

Tables 4, 5, and 6 indicate that our runs performed relatively better for ambiguous, faceted, and vertical-oriented topics than for task-oriented topics, suggesting that the method has some disadvantages concerning task-oriented topics.

6. CONCLUSIONS

This paper reports and discusses our methodologies and results in the Query Understanding (QU) subtask of the IMine-2 task in the NTCIR-12 workshop. Our best runs for the Chinese and Japanese language topics were evaluated in D#-nDCG as 0.4558 and 0.3644, respectively. These evaluation results suggest that our methodology has room for improvement regarding some topics. Our method uses simple text features from text corpora in Chinese and Japanese language encyclopedic articles. Utilizing other features, such as query suggestions and query logs, remains for future work. Our team aims to develop a research platform for user-centered evaluations [1]. The results discussed herein facilitate building it using system-oriented evaluations.

Acknowledgements

This work was supported by JSPS KAKENHI Grant No. 25730193.

7. REFERENCES

[1] Y. Egusa, M. Takaku, and H. Saito. How to evaluate searching as learning. In *Proceedings of Searching as Learning Workshop (IIiX 2014 workshop)*, pages 40–43, 2014.

[2] W. Li, L. Sun, Y. Feng, and D. Zhang. Smoothing LDA model for text categorization. In *Proceedings of AIRS 2008*, pages 83–94, 2008.

[3] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. P. Kato, H. Ohshima, and K. Zhou. Overview of the NTCIR-11 IMine task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11*, pages 8–23, 2014.

[4] D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 781–790, 2010.

[5] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the NTCIR-10 INTENT-2 task. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-10*, pages 94–123, 2013.

[6] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT task. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 82–105, 2011.

[7] S. Tata and J. M. Patel. Estimating the selectivity of *tf-idf* based cosine similarity predicates. *SIGMOD Rec.*, 36(2):7–12, June 2007.

[8] T. Yamamoto, Y. Liu, M. Zhang, Z. Dou, K. Zhou, I. Markov, M. P. Kato, H. Ohshima, and S. Fujita. Overview of the NTCIR-12 IMine-2 task. In *Proceedings of the NTCIR-12*, 2016.

[9] Z. Zheng, S. Song, Y. Meng, and J. Sun. FRDC at the NTCIR-11 IMine task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11*, pages 36–40, 2014.