Academic Term Search Support System for Beginners in Inquiry-based Learning

Yasunobu Sumikawa[1], Ryohei Ikejiri[2], and Yuhei Yamauchi[2]

[1] Depat. of Computer Science, Takushoku University, 815-1 Tatemachi, Hachioji-shi, Tokyo, Japan
ysumikaw@cs.takushoku-u.ac.jp
[2] Interfaculty Initiative in Information Studies, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{ikejiri,yamauchi}@iii.u-tokyo.ac.jp

Abstract. Previous studies on journal searching have proposed models to assist those with experience in publishing articles in journals. We propose a search engine to retrieve academic terms using non-academic terms to support beginner students while searching for academic journals. The proposed search engine supports beginner students with insufficient knowledge regarding academic terms, which limits their ability to find articles that they want to read. Our search engine uses a classifier to retrieve the appropriate academic terms from sentences without using academic terms, and a ranking algorithm that uses the Wikipedia graph structure.

Keywords: Academic terms searching · inquiry-based learning · digital library · search engine

1 Introduction

Knowledge discovery is considered as one of the most important skills to be acquired by not only professional academic researchers who structure knowledge systems, but also by many others, and its importance has been widely recognized. Research is being conducted on organizing inquiry activities as a learning activity in schools [4, 8]; in these activities, students formulate their own hypotheses, investigate previous research, and conduct experiments to prove their hypotheses. According to Pedaste et al. [8], inquiry activities consist of five phases: orientation, conceptualization, investigation, conclusion, and discussion. The process of conceptualization includes two sub-phases question generation and hypothesis generation. Therefore, a literature review of previous studies is considered for conceptualization. However, conceptualization is one of the most difficult phases of the inquiry activity process. For example, numerous students have difficulty starting a research project, defining a topic, or narrowing a topic [3]. Although they have been educated to develop literacy in information retrieval for these problems, it is assumed that students have a certain level of academic knowledge in the research field. It has been pointed out that students need support to

supplement their academic knowledge [9]. A review of previous studies is an important phase in inquiry activities; it requires a high level of competence because discovering the appropriate academic articles without academic terms used in the articles is difficult. Therefore, appropriate support for literature research in an inquiry activity is required, especially for high school and university students who do not have any research experience or publication history.

This study proposes a search engine to support high school and university students in conducting literature research. Because the target learners of this study are high school and university students who have never published a paper, we assume the following: 1) We cannot use the content of previously published articles or citation-related information, as suggested in previous studies. 2) Users have insufficient knowledge of academic terms. We assume that a literature search consists of two stages: selecting the appropriate keywords to search for articles and using the keywords for the literature search. Our algorithm aims to support the selection of the appropriate keywords to search for articles. To achieve this goal, our algorithm uses a sentence that summarizes the information input by the learner. The algorithm then extracts named entities and creates feature vectors using the entire text. These two types of information are used to calculate the similarity between the input text and academic terms using two algorithms that use the graph structure of Wikipedia and a text classifier. Finally, the algorithm outputs academic terms useful for article retrieval.

2 Related Works

Support for academic article retrieval has been studied in the field of information retrieval. There are two main approaches for this support: one that intervenes with professional supporters, and one that improves the ability to search for information. For the former, an educational model was developed in which librarians ask questions to clarify students' interests, provide knowledge on the topic, and support information retrieval [9]. For the latter, there are studies about query expansion based on data sources, applications, and expansion techniques over the past few decades [1]. Recent research has shown that various data sources are divided into categorized into four categories: documents used in the retrieval process, hand-built knowledge resources, external text collections and resources, and hybrid data sources. Furthermore, there are three types of expansion approaches: manual, automatic or interactive [1]. As a type of automatic approach, there are numerous techniques using word and topic models and citation contexts to recommend citation, see the details in the survey paper [2]. One of the closest studies to this study supports appropriate article retrieval for young researchers who have published only a few articles [10]. These studies recommend articles based on previously published papers; however, different information is required to provide recommendations to users who have not yet published academic articles, such as high school and university students, for the initial stages of inquiry activities.

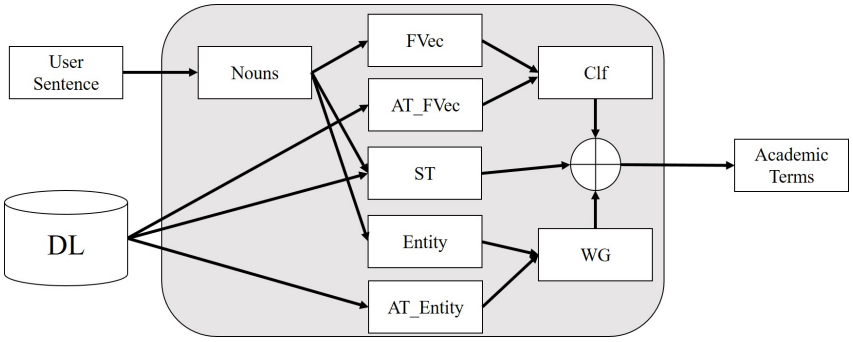

Fig. 1. System overview

3 Ranking Algorithm

Fig. 1 shows an overview of the proposed algorithm[3]. This search engine uses a database (DB) including journals downloaded in advance from digital libraries (DL). The current version uses J-STAGE[4] because it is one of the most widely used digital libraries in Japan. We collected all 5,293,761 articles stored in the DL and extracted all the nouns from the titles. We also extract nouns from the input text of the user. The nouns are used 1) to create the feature vectors *FVec* for user text and *AT_FVec* for academic terms that are used in the classification algorithm *Clf*, 2) to collect named entities *Entity* for user text and *AT_Entity* for academic terms and their corresponding Wikipedia articles to measure similarity using a Wikipedia-graph-based algorithm *WG*, and 3) to perform a simple text analysis *ST* that counts the number of words that have the same characters. We describe the two algorithms, *Clf* and *WG*, in the remainder of this section.

3.1 Document Classification Based Ranking Algorithm

The *Clf* algorithm calculates the probability of the learner's question text being related to each academic term by applying a classifier. To train the classifier, we first created a labelled dataset that included combinations of questions and academic terms. We asked 1,000 high school and university students without research experience or publication history to propose questions as an initial step in inquiry-based history learning. We then annotated the questions by assigning academic terms stored in the encyclopedia of historiography [7]. Note that all annotators had a Ph.D. degree in machine learning or education. Before annotation, we filtered the academic terms for which corresponding academic articles could not be found in the DL database. Consequently, we decided to use 943

[3] The current search engine supports only Japanese. We will support other languages such as English.

[4] https://www.jstage.jst.go.jp/browse/-char/ja/

sentences for constructing our algorithms and 51 academic terms to present the output results of the algorithm.

As the current version uses texts written in Japanese, it first applies morphological analysis. The algorithm uses JUMAN++[13] for user text and MeCab[5] for journal text. After removing the stop words, we create feature vectors. Because this algorithm classifies sentences entered by high school students, it is assumed that sentences are relatively short. To create effective feature vectors for short sentence classification, we use the following seven that are validated in the previous studies [11, 12] that classify events written in short sentences: TF-IDF to use words from the user and academic term texts (F_1), LSA (F_2), LDA (F_3), and Doc2Vec (F_4) to capture latent semantic structures, noun context (F_5) to capture the semantic meanings of nouns by replacing the top k closest words on Word2Vec space, Wikipedia title (F_6) and Wikipedia category (F_7) to use a knowledge collection by mapping the short sentence by explicit semantic analysis (ESA) [6]. We use the top 5 Wikipedia articles as the results of ESA for the last two features. Finally, we trained the classifier on the feature vectors.

3.2 Wikipedia Reference Graph Based Ranking Algorithm

This algorithm analyzes the *term context*, indicating a term set that is often used together to determine the rankings. If two different academic terms are used together with the same term or if the same term is used to explain them, the two terms are considered to be similar. For this algorithm, we used Wikipedia because we need the large number of academic terms. Wikipedia not only provides a detailed description of each article in the text but also provides references to other related articles.

Wikipedia article collection. To apply this algorithm, we manually collected Wikipedia articles corresponding to the academic terms output by the algorithm in advance. We apply ESA to collect Wikipedia articles that corresponded to the learner texts. The ESA outputs concepts if it analyzes them to be the most appropriate for the input sentence. Because ESA considers Wikipedia articles as concepts, it is possible to retrieve the relevant Wikipedia articles from sentences.

Scoring based on the similarity of reference graphs. Under the assumption that articles with matching reference relations are similar, we define the similarity WG between the user's sentence L_t and the academic term sentence W_t on the Wikipedia reference graph using the following equation:

$$WG(L_t, W_t) = \frac{Ref(L_t, W_t) \cap Refed(L_t, W_t)}{Ref(L_t, W_t) \cup Refed(L_t, W_t)} \tag{1}$$

Function *Ref* returns the number of Wikipedia articles cited in both Wikipedia articles provided in the argument. Function *Refed* returns the number of Wikipedia articles that cite both Wikipedia articles provided in the argument. In the above

equation, the numerator calculates the similarity between the user text and text of the academic term. Because both the number of references and non-references to the Wikipedia articles vary from article to article, the denominator normalizes this number such that the number of references does not affect the result.

3.3 Integrated Algorithm

To integrate all results of the three algorithms mentioned above, we use the following equation to calculate ranking scores of academic terms.

$$rank(L_t, W_t) = \alpha * Clf(L_t, W_t) + \beta * WG(L_t, W_t) + \gamma * ST(L_t, W_t) \quad (2)$$

α, β, and γ are hyper-parameters. We assume that the sum of the three hyper-parameters is 1.0. In our search engine, we set the hyper-parameters as 0.333, 0.333, and 0.334, respectively.

4 Experimental Evaluations

4.1 Research Questions

In this study, we performed experimental evaluations according to the following two research questions:

RQ 1 Can we discover learners' text that can present appropriate academic terms?

RQ 2 How accurately can we find appropriate academic terms to text?

For the first research question, we evaluated the classifiers trained on texts as inquiry-based history-learning topics. First, we defined each academic term as a category. We also defined a category *None* that indicates that there is no academic term that can be assigned. We then evaluated the accuracy of the classifier in classifying each category. As 51 academic terms were used in this evaluation, we trained classifier algorithms to classify them into 52 categories, including *None*.

For the second research question, we evaluated the accuracy of our algorithm using only test data to which academic terms could be assigned.

4.2 Experimental setting

Dataset As described in Section 3.1, this study uses 943 crowdsourced texts that summarize the topics in which high school and college students currently have an interest to conduct inquiry learning that they actually filled out for this experiment.

Algorithms The classifiers used for the evaluation of RQ1 were: random forests (RFs), SVM with linear kernel (SVM-Lin), and SVM with RBF kernel (SVM-RDF).

The following algorithms were used in the evaluation of RQ2:

- Jaccard: This uses only nouns in all sentences to measure similarity scores based on letter agreement alone.
- *Clf*: This finds the similarity using only the classifier described in Section 3.1.
- *WG*: This algorithm uses the similarity of Wikipedia's reference graphs described in Section 3.2.
- *Proposed*: The algorithm uses *Clf*, *WG*, and *ST*, as described in Section 3.3.

This evaluation requires the use of Wikipedia, which corresponds to an academic term. In this evaluation, 77 learner data met this condition.

Evaluation criteria As the evaluation of RQ1 can be regarded as a multi-class classification problem, we used three types of evaluation criteria for this problem: precision, recall, and F1-score.

As the evaluation of RQ2 can be regarded as a problem of information retrieval, we used the mean average precision (MAP), which is widely used for the evaluation of search engines.

All evaluation results in this study were the average of the results obtained from a 10-split cross-validation test.

4.3 Results

Q. Which feature vectors achieved the highest accuracy for RQ1?
A. The best results were obtained by combining five feature vectors: Wikipedia title, TF-IDF, Doc2Vec, Wikipedia category, and noun context.

First, we analyzed the features that should be used for classifier training. Tab.1 shows the results of training an RF using each of the features listed in Section3.1. The overall trend shows that the precision value is high; however, the recall is relatively small. Next, we checked whether the combination of these features would improve the accuracy, as previous studies of short sentence classification on events found that combining several features improve accuracy [11, 12]. In this study, we combined each feature individually in the order of their F1 scores. After combining the features, we used random forests to extract only the top 500 elements that were important for classification and used them to train the classifier. As shown in Tab. 1, combining them provided a better result than using them individually. Especially, we can see that the combination of F_6, F_1, F_4, F_7, and F_5 is the best. Therefore, in the following sections, the classifier is trained using only a combination of the F_6, F_1, F_4, F_7, and F_5 features.

Table 1. Scores in feature selection with random forests

	Precision	Recall	F1
F_1	77.2%	*68.9%*	72.8%
F_2	80.3%	63.1%	70.7%
F_3	70.6%	62.8%	66.3%
F_4	78.1%	67.9%	72.6%
F_5	77.9%	67.1%	72.1%
F_6	78.1%	68.6%	73.0%
F_7	78.3%	67.4%	72.4%
$F_6 + F_1$	79.1%	68.4%	73.4%
$F_6 + F_1 + F_4$	78.8%	67.7%	72.8%
$F_6 + F_1 + F_4 + F_7$	78.7%	67.7%	72.7%
$F_6 + F_1 + F_4 + F_7 + F_5$	80.2%	68.5%	*73.8%*
$F_6 + F_1 + F_4 + F_7 + F_5 + F_2$	*80.4%*	67.5%	73.3%
$F_6 + F_1 + F_4 + F_7 + F_5 + F_2 + F_3$	80.1%	67.4%	73.2%

Table 2. Scores in classifier selection

	Precision	Recall	F1
RFs	*80.2%*	68.5%	*73.8%*
SVM-lin.	78.2%	*69.9%*	*73.8%*
SVM-rbf	76.9%	68.5%	72.4%
NB	53.4%	68.6%	59.7%

Table 3. Comparison of unsupervised learning and classifiers

	Precision	Recall	F1
Jaccard	20.4%	33.1%	22.4%
WG	6.5%	36.7%	10.4%
Clf	*80.2%*	*68.5%*	*73.8%*

Q. Which classification algorithm achieved the highest accuracy for RQ1?
A. Random forests provided the best accuracy.

Tab. 2 shows the precision, recall, and F1 scores obtained using the four classifiers. This result also shows that the precision was high for all algorithms, whereas recall was low. Because the best F1 score was obtained by RFs, we used RFs in the following evaluations as a classifier.

As our algorithm is a search engine system, Tab. 3 shows the results of the comparison with the algorithms we planned to use in RQ2, which uses pure word matching (Jaccard) and the Wikipedia graph structure (*WG*) with classification. As a characteristic of the results, we can confirm that unsupervised learning has low precision but relatively high recall, whereas the classifier tends to have high precision but low recall.

Table 4. Comparison of the top 10 search results

	MAP	No result in TOP10
WG	26.6%	54.7%
Jaccard	12.7%	54.7%
Clf	3.5%	75.3%
Jaccard + *WG* + *Clf*	*36.2%*	*39.3%*

Table 5. Distribution of ranking where correct answer data existed in the top 10 of the proposed algorithm.

Rank	1	2	3	4	5	6	7	8	9	10
Num.	17	10	1	4	2	3	2	1	2	2

Q. Which is the best search algorithm among the 4 algorithms?
A. The proposed algorithm is the best.

Next, we evaluated the accuracy of the academic terms listed as the top 10 results for each algorithm to verify its effectiveness as a retrieval system. As mentioned above, we used 77 data points in this evaluation because we used only those academic terms from Wikipedia that could be assigned to learners' sentences in the test data.

Tab. 4 shows the results. Focusing on the MAP score, using only the classifier obtained the worst result. However, the best result was obtained when we combined all the algorithms. As shown in the evaluation in Tab. 3, the classifier provided a high precision score, whereas the other algorithms provided high recall scores. It is likely that combining these factors would have improved the overall results.

The number of correct answers that were not in the top 10 is shown in the third column of Tab. 4. This result indicates that the proposed algorithm is the best. Our algorithm did not provide the correct answer for only 40% of the test data, whereas the other algorithms failed to obtain the correct answers for more than half of the test data.

4.4 Error Analysis

Q. What was the distribution regarding the location of the correct answer among the top 10 results output by the proposed algorithm?
A. Rank 1 was the highest.
A. Most data provided the correct answer using the top two numbers.

Tab. 5 shows where the correct answer was obtained in the top 10 cases when the proposed algorithm was applied. First, 44 (57%) data had correct answers in the top 10. The location with the highest number of correct answers was ranked

Table 6. Academic terms for which correct data were not available and their frequencies

Academic term	English name	Num.
辺境	Periphery	4
世界システム論	Theory of the world-system	3
社会進化論／社会ダーウィニズム	Social evolutionism/social Darwinism	3
昭和史論争	Showa history controversy	3
文化圏	Kulturkreis	2
オーラルヒストリー	Oral history	2
社会集団	Social group	2
発展段階	Stages in development	2
（解釈／）解釈学	Hermeneutics	2
社会変動	Social change	1
産業革命論争	Industrial revolution	1
オリエンタリズム	Orientalism	1
文化伝播	Trans-cultural diffusion	1
文化財保護	Law for the protection of cultural properties	1
明治維新論争	Meiji restoration	1
実証主義	Positivism	1
日本文化論	Culture of Japan	1
マルクス主義（歴史学）	Marxism	1
共同体	Community	1

1. Rank 2 also had a high number of correct answers. Therefore, the majority of correct answers were found in the top two cases.

Q. What were the academic terms that the proposed algorithm tends to get wrong most often?
A. "Periphery" was the most difficult word to assign correctly.

Tab. 6 shows the academic terms that were not ranked in the top 10. The reason why "periphery" and "theory of the world-system" were listed at the top of the error list would be that these academic terms could be regarded as theories that explain various historical phenomena from a broad spatio-temporal perspective; thus, the accuracy of the search results was lower than that of terms related to specific periods or phenomena. In contrast, "social evolutionism/social Darwinism" and "Showa history controversy" were also listed as the top error terms because they were both academic terms with a narrow range of applicability although the words included in these terms have a wide range (for example, "Showa" is the name of a period in Japanese history).

5 Conclusions

In this study, we proposed an academic term search engine to assist high school and university students in conducting literature research. In particular, our algo-

rithm aims to support the selection of appropriate keywords to search for articles in the literature.

In future work, we aim to make high school students use the search engine with our algorithm and analyze how beneficial it is for inquiry-based history learning.

Acknowledgments. This work was supported in part by MEXT Grant-in-Aids (#20H01717).

References

1. Azad, H.K., Deepak, A.: Query expansion techniques for information retrieval: A survey. Information Processing & Management **56**(5), 1698–1735 (2019). https://doi.org/https://doi.org/10.1016/j.ipm.2019.05.009, https://www.sciencedirect.com/science/article/pii/S0306457318305466
2. Färber, M., Jatowt, A.: Citation recommendation: approaches and datasets. Int J Digit Libr (2020)
3. Head, A., Eisenberg, M.: Truth be told: How college students evaluate and use information in the digital age. SSRN Electronic Journal (01 2010)
4. Keselman, A.: Supporting inquiry learning by promoting normative understanding of multivariable causality. Journal of Research in Science Teaching **40**(9), 898–921 (2003)
5. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to japanese morphological analysis. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain. pp. 230–237. ACL (2004)
6. M. Chang, M.W., L. Ratinov, L., Roth, D., Srikumar, V.: Importance of semantic representation: Dataless classification. In: AAAI (7 2008)
7. Ogata, I., Kato, T., Kabayama, K., Kawakita, M., Kishimoto, M., Kuroda, H., Sato, T., Minamizuka, S., Yamamoto, H.: Encyclopedia of historiography. koubundou (1994), http://ci.nii.ac.jp/ncid/BN10236869
8. Pedaste, M., Mäeots, M., Siiman, L.A., de Jong, T., van Riesen, S.A., Kamp, E.T., Manoli, C.C., Zacharia, Z.C., Tsourlidaki, E.: Phases of inquiry-based learning: Definitions and the inquiry cycle. Educational Research Review **14**, 47–61 (2015)
9. Scharf, D., Dera, J.: Question formulation for information literacy: Theory and practice. The Journal of Academic Librarianship **47**(4), 102365 (2021). https://doi.org/https://doi.org/10.1016/j.acalib.2021.102365, https://www.sciencedirect.com/science/article/pii/S0099133321000562
10. Sugiyama, K., Kan, M.Y.: Scholarly paper recommendation via user's recent research interests. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries. pp. 29–38. JCDL '10, Association for Computing Machinery, New York, NY, USA (2010)
11. Sumikawa, Y., Jatowt, A.: Classifying short descriptions of past events. pp. 729–736. ECIR'18 (2018)
12. Sumikawa, Y., Ikejiri, R.: Feature selection for classifying multi-labeled past events. Int. J. Digit. Libr. **22**(1), 63–83 (2021)
13. Tolmachev, A., Kawahara, D., Kurohashi, S.: Juman++: A morphological analysis toolkit for scriptio continua. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 54–59. Association for Computational Linguistics, Brussels, Belgium (2018)