Toward a Search Formula Creation Support for the Exhaustive Search of an Academic Paper

FUKUDA, Satoshi
Graduate School of Information Science and Electrical Engineering, Kyushu University

TOMIURA, Yoichi
Graduate School of Information Science and Electrical Engineering, Kyushu University

https://doi.org/10.5109/2230666

出版情報：Proceedings of Toward Effective Support for Academic Information Search Workshop. 1, pp.1-7, 2018-11-22. 九州大学大学院システム情報科学研究院
バージョン：
権利関係：

Toward a Search Formula Creation Support for the Exhaustive Search of an Academic Paper

Satoshi FUKUDA † Yoichi TOMIURA †

† Graduate School of Information Science and Electrical Engineering, Kyushu University
{s.fukuda, tom}@inf.kyushu-u.ac.jp

Keywords: Exhaustive search, Boolean search formula, Research paper

1 Introduction

In an academic paper search, particularly a search to confirm the originality of a user's research or a search to collect articles and write a survey, it is important that the search returns comprehensive results related to the user's information need. To achieve such a search, the user must define a search formula that collects research papers broadly related to his/her information need. In many cases, however, it is rare that relevant papers can be collected comprehensively with the search formula that was firstly created, and the user repeatedly conducts a burdensome work such as creating a new search formula or modifying the search formula by verifying the search result. In addition, in recent years, fusion research with interdisciplinary fields and interdisciplinary research have been actively studied. For the user majoring in a specific research area, the cognitive burden on the research paper's search such as investigating research trends in the different field is more increase.

To improve the performance of the search based on the search formula firstly created by the user, many query term expansion techniques have been studied; e.g., using a thesaurus [10, 11], document set retrieved by the user [12, 17], and query log [2, 14]. Moreover, to potentially anticipate all possible terms related to the search word that the authors might have used, several approaches of using a query as a potential topic by employing a topic model through latent dirichlet allocation (LDA) [4] on the document set have been studied. For example, a query likelihood model that incorporates topic analysis results into a language model has been proposed [16], and a method of searching for papers using multiple topic analysis results and a search formula created by a user has been proposed [3].

However, unless the structure of the search formula can accurately represent the user's information need, it is not possible to improve the search performance by expanding the search formula applying the above approaches. It is difficult to create a search formula that correctly expresses the information need. In particular, the cognitive burden is huge for the user who does not have background knowledge in the target field. To solve this issue, we construct a framework that supports the creation of a Boolean search formula reflecting the information need more accurately for the user. As a first step to realize such system, we investigate the effectiveness of modifying the original search formula by adding new words to investigate whether it is really beneficial to modify the original search formula.

2 Experimental Design for Investigating the Effectiveness of Adding New Search Words to the Search Formula

This section describes an experimental design to demonstrate whether the search performance is improved by adding new search words to the search formula created by the user. The procedure is as follows. The user first creates a search formula that represents his/her information need. The system then ranks the research papers in the database using the abstract of each research paper by a ranking method presented in [16]. The system next extracts candidate words to be added to the original search formula from the top 100 abstracts (papers) of the ranking result and creates a candidate word list. The system finally connects candidate words to the search words in the original search formula with the operator AND, and ranks the research papers in the database again.

The approach of connecting candidate word to the original search formula is as follows. If the original search formula is

((AAAA OR AAAA') AND (BBBB OR BBBB'))
OR ((AAAA OR AAAA') AND (CCCC OR CCCC'))

and the added word is DDDD, the search formula is modified as

((AAAA OR AAAA') AND (BBBB OR BBBB') AND DDDD)
OR ((AAAA OR AAAA') AND (CCCC OR CCCC') AND DDDD).

In adding words to the original search formula, it is highly possible that words unrelated to the user's information need and prepositions are not candidate. We therefore set three constraints as the conditions of candidate words to be added to the original search formula to re-rank the research papers in the database: (1) candidate words are nouns, verbs, and adjectives; (2) a candidate word appears more than three times in the top 100 papers ranked by the original search formula; and (3) a research paper searched by the modified search formula are contained in the database.

As an approach of adding search words to the original search formula, we select one or two types of words from the candidate word list. When adding a candidate word to the original search formula, we investigate all words in the candidate word set. When adding two types of candidate words to the search formula, we investigate all combinations for the candidate words. For example, if XXXX, YYYY, and ZZZZ are included in the candidate word list, the three candidate patterns (XXXX AND YYYY), (XXXX AND ZZZZ), and (YYYY AND ZZZZ) are created.

3 Experiment

3.1 Experimental Settings

We used test collections of information retrieval tasks from the NTCIR-1 and 2 datasets [6, 7]. The NTCIR dataset contains 132 search tasks that describe the conditions of research papers satisfying the information need, and approximately 1000 to 4000 research papers that are rated as relevant, partially relevant, or irrelevant for each task. In this experiment, we considered papers that were labeled as relevant or partially relevant

as satisfying a particular information need, and tested our system using the annotated research paper set for each search task. We also considered two types of search task with different research areas. Table 1 gives examples of the identification number, summary of the information need, original search formula, and number of annotated papers for each search task on the NTCIR datasets tested in the experiment. The search formulas in each search task were manually created by another subject majoring in the field of natural language processing (NLP) by reading the contents of the search tasks.

In the evaluation, we measured the cumulative recall for the paper set of the top 1% to 100% (in 1% increments) of the ranking results and evaluated candidate words to be added to the original search formula according to the size of the graphical area of the cumulative recall. The recall was calculated as (the number of relevant papers included in the paper set of the top *n*%) / (the number of papers judged as relevant papers in the search task). As parameters of the ranking model [16], we set $\alpha = 0.5$, $\beta = 0.1$, the number of topics as 10, $\lambda = 0.5$, and $\mu = 10$. In LDA analysis, we used original formed words that were nouns, verbs, and adverbs as a part-of-speech in the abstracts. We used TreeTagger [13] to distinguish the part-of-speech of words.

Table 1. Example of detailed information on search tasks tested in the experiment.

Task num.	Summary of the information need	Manually created search formula	Num. of annotated papers
0059	Research papers on automatic construction of a thesaurus.	(thesaurus OR ontology) AND (construct OR create OR construction OR creation)	2,608
0138	Research papers on the synthesis of stable triplet carbine.	carbine AND (synthesis OR synthesize)	834

3.2 Experimental Results and Discussion

Figs. 1 and 2 show experimental results obtained for the original search formula and the search formula with the addition of candidate words having the largest graph area of the cumulative recall. Fig. 1 shows results for search task 0059 while Fig. 2 shows results for search task 0138. Fig. 1 reveals that even if the subject majoring in the field of NLP creates a formula for searching research papers in the NLP field, the search performance is improved by adding queries such as "(pair AND word)" and "(relationship AND word)" to the search formula. Fig. 2 shows that the formula for searching chemistry papers created by the subject with no background knowledge in the field is not elaborate. However, the search performance was dramatically improved by adding a combination of search words, such as "(interaction AND poly)" and (comparison AND interaction), to the search formula. These results confirm the effectiveness of adding a search word to the search formula created by the user and it can be considered necessary to modify the search formula to realize a better search. Moreover, these results suggest that it is better to use two types of search word than one type of word.

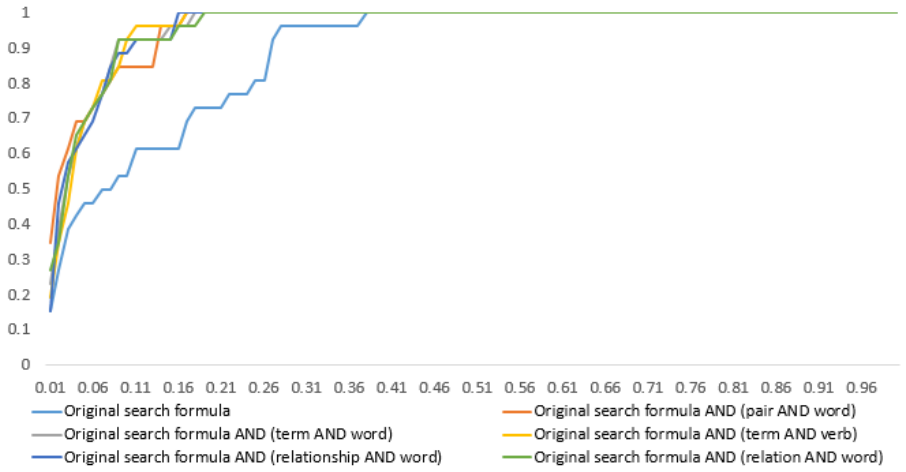

Fig. 1. Cumulative recall ratio curves of the ranking results obtained using search formulas for search task 0059.

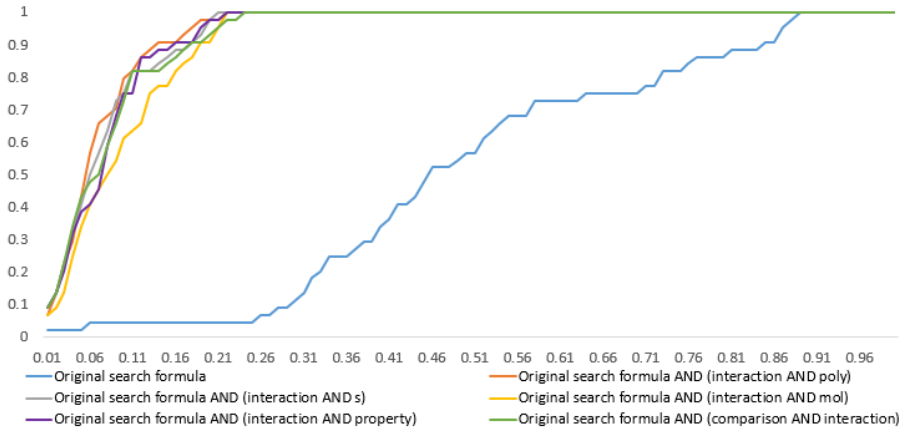

Fig. 2. Cumulative recall ratio curves of the ranking results obtained using search formulas for search task 0138.

In future work, we aim to develop an approach that allows the user to discover words useful for improving search performance as shown in Figs. 1 and 2. As a simple approach, it is conceivable that the system asks the user to judge some of top papers ranked by the original search formula and then estimates candidate words by feeding back the annotation results. By using the annotated paper set, for example, the system can find words that related to the information need from various viewpoints. Table 2 shows some candidate words that appear high frequently in the relevant paper set in the top 100 papers ranked using the original search formula in each search task. Table 3 shows top candidate words sorted according to (frequency of appearance in the relevant paper set) / (frequency of appearance in the paper set) for the top 100 papers ranked by the original search formula in each search task. It is seen that some candidate words

shown in Figs. 1 and 2 appear in Tables 2 and 3 for each search task; in particular, "pair" and "interaction", which do not appear in Table 2, appear in Table 3. These results show that it is possible to obtain indicators for the selection of candidate words by receiving feedback for the search result from the user. However, to appropriately select words used for modification of the original search formula, it is necessary to judge not only using the indices given in Tables 2 and 3 but also, for example, considering how the word is related to the information need, and whether to treat the word as a synonym in a concept unit in the original search formula or as new concept unit for the original search formula. Furthermore, as shown in Figs. 1 and 2, when adding new search words to the original search formula, the search performance is improved by adding two types of unit rather than one type, however, it is difficult to judge whether between words in the candidate word list have a relationship of AND, OR or originally no relation for the user. For example, "(word AND term)" in Fig. 1 is better to treat as "(word OR term)" because these words are synonymous. The relationship between these words may be understood by the user majoring in the NLP field, however, it is a burden to grasp the relationship manually for the user majoring in the different field. It can therefore be said that a framework that supports how to combine candidate words is also necessary.

Table 2. Examples of candidate words appearing high frequently in the relevant paper set in the top 100 papers ranked using the original search formula.

Task	Candidate words
0059	paper, method, word, dictionary, base, information, machine, use, language, relation, semantic, knowledge, classification, important, problem, Japanese, result, concept
0138	1, property, polymer, other, work

Table 3. Examples of candidate words having the highest ratio calculated as (frequency of the appearance in the relevant paper set) / (frequency of appearance in the paper set) in the top 100 papers ranked by the original search formula.

Task	Candidate words
0059	suitable, attempt, readable, translation, similarity, pair, high, algorithm, value, classification, verb, corpus, point, important, sentence
0138	equation, al., applicable, H., interaction, recent, et, diamine, orientation, work, attention, mol, organometallic, atom, open, table, s, 60

4 Related Work

A query suggestion is the most relevant task in our research. This task is intended to recommend semantically different queries from a query input by the user and is different from query expansion by acquiring synonyms. For the query suggestion, in addition to studies using query logs and session information [1, 5, 9], several studies have recommended queries based on document sets retrieved from a database using relevant feedback [8, 15]. Verberne et al. proposed an approach that recommends queries related to the user's information need from the initial query by an interactive operation based on the relevance feedback [15]. Kim et al. proposed a method of automatically estimating Boolean queries using pseudo-relevant feedback and a decision tree [8]. Pseudo-relevant feedback is a relevant feedback assuming that the highest-ranked documents

are relevant. We consider that good results can be obtained when applying this model to the ranking result obtained using the original search formula as shown in Fig. 1. However, it is possible that pseudo-relevant feedback produces a poor result when applied to the result ranked by the original search formula as in Fig. 2. Relevant feedback provides relatively stable performance because the user manually judges documents to be relevant or irrelevant in the search result. Although reference [15] used only documents judged as relevant, we consider that estimates the candidate words by using relevant and non-relevant information judged by the user.

For the annotation work in our system, we will also investigate a method to mitigate the cognitive burden on the user. In Tables 2 and 3, we estimated the candidate words from the top 100 research papers. But actually, it may be costly for the user to read all sentences in the abstract and judge whether each paper is related to the information need. However, if the same effect as when judging top 100 papers may be obtained by judging only the top 30 or 50 papers, the cost for the user will be less than half. We also consider that the same effect of reading all sentences in the abstract can be obtained by reading only sentences related to the user's information need when judging whether the research paper is related to the information need. This is based on the knowledge that if a user searches the papers related to the information need, he/she will mainly judge whether a paper is relevant from sentences related to the information need in the abstract. If such effect is confirmed, the user's cognitive burden will more decrease by presenting only sentences related to the user's information need in the abstract.

5 Conclusion

Our research task is to construct a framework for supporting the creation of a search formula more accurately reflecting the user's information need in an academic paper search. As the first step, we verified the effectiveness of adding new search words to the original search formula using the operator AND. In future work, we will develop methods of effectively estimating search words that will improve the search performance and decreasing the user's cognitive burden for the modification of the original search formula and annotation work whether the research paper is related to the information need.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP15H01721. We thank Glenn Pennycook, MSc, from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

References

1. P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna.: Query suggestions using query-flow graphs. In: WSCD, pp. 56–63 (2009).

2. B. M. Fonseca, P. Golgher, B. Pôssas, B. Ribeiro-Neto, and N. Ziviani.: Concept-based interactive query expansion. In: CIKM, pp. 696–703 (2005).
3. S. Fukuda and Y. Tomiura.: Using topic analysis techniques to support comprehensive research paper searches. In: IALP (2017).
4. T.L. Griffiths and M. Steyvers.: Finding scientific topics. In: National Academy of Sciences, pp. 5228–5253 (2004).
5. C.-K. Huang, L.-F. Chien, and Y.-J. Oyang.: Relevant term suggestion in interactive web search based on contextual information in query session logs. American Society for Information Science and Technology 54(7), 638–649 (2003).
6. N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, S. Hidaka, and J. Adachi.: The NTCIR workshop: The first evaluation workshop on Japanese text retrieval and cross-lingual information retrieval. In: Information Retrieval with Asian Languages Workshop, pp. 1–7 (1999).
7. N. Kando.: Overview of the second NTCIR workshop. In: NTCIR Workshop, pp. 35–43 (2001).
8. Y. Kim, J. Seo, and W.B. Croft.: Automatic Boolean query suggestion for professional search. In: SIGIR, pp. 825–834 (2011).
9. Q. Mei, D. Zhou, and K. Church.: Query suggestion using hitting time. In: CIKM, pp. 469–478 (2008).
10. Y. Qiu and H.-P. Frei.: Concept based query expansion. In: SIGIR, pp. 160–169 (1993).
11. H. Schütze and J.O. Pedersen.: A cooccurrence-based thesaurus and two applications to information retrieval. Information Processing and Management 33(3), 307–318 (1997).
12. E. Terra and C.L.A. Clarke.: Scoring missing terms in information retrieval tasks. In: CIKM, pp. 50–58 (2004).
13. TreeTagger homepage, http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
14. K. Uchiumi, M. Komachi, K. Machinaga, T. Maezawa, T. Satou, and Y. Kobayashi.: Japanese abbreviation expansion with query and clickthrough logs. In: IJCNLP, pp. 410–419 (2011).
15. S. Verberne, M. Sappelli, and W. Kraaij.: Query term suggestion in academic search. In: ECIR, pp. 560–566 (2014).
16. X. Wei and W.B. Croft.: LDA-based document models for ad-hoc retrieval. In: SIGIR, pp. 178–185 (2006).
17. J. Xu and W.B. Croft.: Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems 18(1), 79–112 (2000).