Information Processing and Management 45 (2009) 664–682

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Retrieval parameter optimization using genetic algorithms ☆

Sumio Fujita *

Yahoo Japan Corporation, Midtown-tower, 9-7-1 Akasaka, Minato-ku, Tokyo 107-6211, Japan

ARTICLE INFO

Article history:
Received 14 March 2008
Received in revised form 6 February 2009
Accepted 28 April 2009
Available online 16 June 2009

Keywords:
Information retrieval
Test collections
Parameter optimization
Genetic algorithm

ABSTRACT

This paper describes our experiments on automatic parameter optimization for the Japanese monolingual retrieval task. Unlike regression approaches, we optimized parameters completely independently of retrieval models enabling the optimized parameter set to illustrate the characteristics of the target test collections. We adopted genetic algorithms as optimization tools and cross-validated with four test collections, namely the CLIR-J-J collections for NTCIR-3 to NTCIR-6. The most difficult retrieval parameters to optimize are the feedback parameters, because there are no principles for calibrating them. Our approach optimized feedback parameters and basic scoring parameters at the same time. Using test sets and validation sets, we achieved effectiveness levels comparable with very strong baselines, i.e., the best-performing NTCIR official runs.

1. Introduction

The choice of scoring function is crucial for improving search effectiveness in retrieval evaluation experiments such as the Text Retrieval Conference (TREC) (Harman, 1995) or NACSIS-NII Test Collection for Information Retrieval Systems (NTCIR) (Kando, 2004) workshops, where search experiments are conducted against publicly available test collections. For each topic representing an independent search request, effectiveness is measured by well-known metrics such as the average precision against the top 1000 documents retrieved by the system. For a set of about 50 topics, the mean of each average precision (MAP) is utilized to measure the effectiveness of the set of search results. Such an execution unit of search tasks against a set of search topics is called a "run", and the MAP values for runs with different systems or different parameters are compared for experimental purposes. Typically, the retrieval procedure against a search topic in such experiments comprises the following three stages.

(1) Query construction from the topic descriptions provided

Recently, automatic query construction has usually been adopted, in which terms are extracted from natural language texts following morphosyntactic analysis. For Japanese, in particular, word-based indexing and *n*-gram-based indexing achieve comparable effectiveness, but word-based indexing is usually adopted because word-indexed terms make sense for both the machine and human users. This stage is not critical even in Japanese because simple short unit words in Japanese achieve the best effectiveness in many cases (Fujita, 1999).

(2) Relevance computation by a scoring function

A scoring function computes a relevance score between the query representation and the indexed document representation. The function is defined via a retrieval model such as a similarity function in term space, as in the vector space model (Salton, 1988), or via the log odds ratio of the probability of observing a document given its relevance in the probabilistic

☆ Part of this work was presented at the NTCIR-6 workshop.

* Tel.: +81 3 6440 6000.
E-mail address: sufujita@yahoo-corp.jp

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.04.008

model (Robertson & Walker, 1994). The choice of function is crucial to effectiveness, although further improvement by further elaboration of the scoring function is unlikely, given that performance improvement seems to have plateaued in ad hoc search tasks against static text document collections. A new scoring paradigm, namely the language model-based approach recently proposed by Lafferty and Zhai (2001), does not greatly outperform more orthodox approaches. Even in Japanese retrieval tasks, in spite of intensive efforts to further optimize the ranking, it does not perform better than well-known probabilistic models when properly calibrated for the task (Fujita, 2005b, 2007).

(3) (Pseudo-)relevance feedback

For automatic search tasks, pseudo-relevance feedback is a frequently used technique, achieving as much as a 20% improvement in MAP. In practice, only systems using this strategy achieve the best positions in official runs of ad hoc search tasks in recent TRECs and NTCIRs.

In our previous NTCIR-4 and NTCIR-5 cross-language information retrieval: Japanese to Japanese (CLIR-J-J) experiments, choosing the BM25TF∗IDF retrieval method for official submissions was found to be successful, in comparison with other participating groups. Decisions are taken based on presubmission experiments using the test collections of the previous NTCIR. The scoring functions have coefficient parameters that are determined during the presubmission experiments. Failing to calibrate these parameters properly results in poor effectiveness in the official evaluation, even for good scoring functions. Therefore, calibration of parameters becomes a major element of presubmission experiments for the NTCIR tasks. These coefficient parameters make the scoring function adaptable to diverse environments, which compensates for the effort of recalibration. When using a pseudo-relevance feedback strategy, effectiveness is more sensitive to feedback-specific parameters than to basic scoring parameters. In particular, feedback parameters are subject to collection characteristics, and it is difficult to find a theoretically sound approach to estimating the best feedback parameters. The sensitivity of feedback parameters and feedback effectiveness to test collections was studied in Fujita (2005a).

Given the four available Japanese test collections in the NTCIR-3 to NTCIR-6 CLIR tasks, we aim to investigate whether automatic calibration can achieve the same effectiveness as human calibration with a limited number of training examples.

We consider the calibration process as a general optimization problem and adopt a problem-independent approach, using genetic algorithms (GAs) to optimize the parameters for the given test collections. The following difficulties may arise:

- The optimization process may terminate at local maximum points and fail to find the global maximum.
- The optimized parameters might be overfitted to the training collection and therefore might not perform well for other collections.

In adopting GAs, we face mainly the second of these issues. In this paper, we present our experiments on a Japanese monolingual retrieval task, focusing on the possibility of automatic calibration of search parameters.

The remainder of the paper is organized as follows. Section 2 introduces prior research on related issues. Section 3 describes our experimental environment and retrieval system, with Section 4 describing the NTCIR test collections. Section 5 briefly explains the GA. Section 6 reports our baseline experiments, and Section 7 presents the GA optimization in runs of CLIR-J-J tasks for NTCIR-3 to NTCIR-6. Section 8 concludes the paper.

2. Related work

Optimization of scoring functions has been studied as part of several regression approaches to information retrieval (IR) (Cooper, Gey, & Dabney, 1992; Fuhr & Buckley, 1991; Fuhr & Pfeifer, 1994; Gey, 1994). Fuhr and Buckley use a least-squares polynomial function, whereas others use a logistic regression function. In both cases, these regression models represent the probability of the document relevance given term attributes such as document term frequency or inverse document frequency. Optimum coefficients of each term attribute are estimated by a least-squares method or a maximum likelihood estimate from training collections. While the regression models approximate the probability of the relevance of a document given a term, we directly optimize the MAP for a set of topics against the training collection by tuning the model parameters using GAs.

The application of GAs to IR has a long history: Raghavan and Birchard (1979) applied GAs to the clustering problem in IR, and Gordon (1988) applied them to document index modification. Yang and Korfhage (1994) optimized query weights in relevance feedback by using GAs. Vrajitoru (1998) adopted Gordon's model and introduced improved crossover operations.

Our approach is close to that of Fan et al. (Fan, Gordon, and Pathak (2004); Fan, Luo, Wang, Xi, and Fox (2004)), but we optimize the feedback parameters at the same time, aiming to achieve a MAP comparable with the best official runs in NTCIRs. Some researchers have applied discriminative models to learn ranking functions, but the reported effectiveness is much poorer than the TREC best official runs (Nallapati, 2004). More recently, de Almeida, Goncalves, Cristo, and Calado (2007) reported on genetic programming (GP)-based ranking function discovery experiments and presented optimization results using the TREC-8 collection, but the results are much worse than the best official ad hoc runs of TREC-8 (Kwok, Grunfeld, & Chan, 1999). Yeh, Lin, Ke, and Yang (2007) used a GP approach in the "learning to rank" framework, where the learning algorithm trains a ranking function from query–document pairs provided with graded relevance judgments.