Computers in Industry 65 (2014) 937–951

Contents lists available at ScienceDirect

Computers in Industry

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

Enhancing passage retrieval in log files by query expansion based on explicit and pseudo relevance feedback

Hassan Saneifar[a,b], Stéphane Bonniol[b], Pascal Poncelet[a], Mathieu Roche[a,c,*]

[a] LIRMM – Univ. Montpellier 2, CNRS, Montpellier, France
[b] Satin Technologies, Montpellier, France
[c] UMR TETIS – Cirad, Irstea, AgroParisTech, Montpellier, France

ARTICLE INFO

Article history:
Received 21 March 2013
Received in revised form 3 February 2014
Accepted 11 February 2014
Available online 27 March 2014

Keywords:
Information retrieval
Passage retrieval
Question answering
Query enrichment
Context learning
Log files

ABSTRACT

Passage retrieval is usually defined as the task of searching for passages which may contain the answer for a given query. While these approaches are very efficient when dealing with texts, applied to log files (*i.e.* semi-structured data containing both numerical and symbolic information) they usually provide irrelevant or useless results. Nevertheless one appealing way for improving the results could be to consider query expansions that aim at adding automatically or semi-automatically additional information in the query to improve the reliability and accuracy of the returned results. In this paper, we present a new approach for enhancing the relevancy of queries during a passage retrieval in log files. It is based on two relevance feedback steps. In the first one, we determine the explicit relevance feedback by identifying the *context of the requested information* within a learning process. The second step is a new kind of pseudo relevance feedback. Based on a novel term weighting measure it aims at assigning a weight to terms according to their relatedness to queries. This measure, called TRQ (TERM RELATEDNESS TO QUERY), is used to identify the most relevant expansion terms.

The main advantage of our approach is that is can be applied both on log files and documents from general domains. Experiments conducted on *real data* from logs and documents show that our query expansion protocol enables retrieval of relevant passages.

1. Introduction

Passage retrieval, representing an important phase in question answering and information locating methods, is the task of searching for passages which may contain the answer for a given question. As an accurate and reliable definition, a passage is a fixed-length sequence of words which can begin and end anywhere in a document [36]. Passage retrieval has been extensively investigated since late 1980 and early 1990s [35,54]. Information retrieval methods are typically designed to identify whole documents that are relevant to a query. In these approaches a query is evaluated by using the words and phrases in each document (*i.e.*, the index terms) to compute the similarity between a document and a query [19]. This means that by using information retrieval methods, we are not able to locate the seeking information in documents, but rather find only the relevant documents (*i.e.*, documents containing the seeking information). As an alternative, we have passage retrieval which makes it possible to locate the requested information within a document. In this context, each document is seen as a set of passages, where a passage is a contiguous block of text. Thus, in order to retrieve passages containing the sought information, the similarity of each passage to a query is calculated. Then, the retrieved passages can be used by Information Extraction methods, in order to extract piece of information. Therefore, the main objective of passage retrieval is to locate sought information within documents and thus *reduce* the search space wherein we will look to extract the exact information. Hence passage retrieval can be considered as an intermediate between document retrieval and information extraction.

In this work, we investigate a study on *passage retrieval* and *boosting its performance* within a *question answering* system on a

[*] Corresponding author at: MTD – UMR TETIS, 500, rue J.F. Breton, 34093 Montpellier Cedex 5, France. Tel.: +33 0467548754.
E-mail addresses: hassan.saneifar@gmail.com (H. Saneifar), stephane.bonniol@satin-tech.com (S. Bonniol), poncelet@lirmm.fr (P. Poncelet), mathieu.roche@cirad.fr (M. Roche).
URL: http://hsaneifar.info/, http://www.lirmm.fr/~poncelet, http://www.lirmm.fr/~mroche

http://dx.doi.org/10.1016/j.compind.2014.02.010
0166-3615/© 2014 Elsevier B.V. All rights reserved.

specific domain [49,21,28]. Here we deal with a particular type of complex textual data, *i.e.*, log files generated by Integrated Circuit (IC) design tools. Since IC design tools run a long time in batch mode in Industry, the generated log files are often the user's sole feedback. Users constantly need to check progress by listing these logs. These log files are digital reports on configurations, conditions, and states of systems. In this domain, to ensure the design quality, there are some quality check rules which should be verified. These quality check rules are usually formulated in the form of *natural language questions* (eg., "*Capture the total fixed cell STD*" or "*Capture the maximum Resistance value*"). Verification of these rules is principally performed by analyzing the generated log files. In the case of large designs that the design tools may generate megabytes or gigabytes of log files each day, the problem is to wade through all of this data *to locate the critical information* that we need in order to verify the quality check rules. To this end, we aim at finding the new passage retrieval solutions which are relevant to this domain as the particularities of such textual data (*i.e. log files*) significantly impact the accuracy and performance of passage retrieval in this context.

Due to the fact that log files are multi-source and multi-vocabulary data, the main challenge is the existing gap between vocabulary of queries and those of log files. We call this problem mismatch vocabularies. The problem of mismatch vocabularies is a well-known problem known as the lexical chasm [3]. This issue is also noted in some other work. For example, the authors of [6] note that the answer to a question may be unrelated to the terms used in the question itself, making classical term-based search methods useless [6]. Because the user's formulation of the question is only one of the many possible ways to state the seeking information, there is often a discrepancy between the terminology used by the user and the terminology used in the document collection to describe the same concept [52]. This issue is highlighted in the case of log files which are by default multi-vocabulary data. Also, we have to deal with other challenges in passage retrieval from log files. We can briefly note the lack of data redundancy and thus lack of answer (information) repetitions, lack of paraphrasing or surface patterns in log files, and the lack of semantic resources. We discuss and develop all these issues as well as the problem of mismatch vocabularies in Section 2.

Taking all these difficulties into account, we choose Query Expansion in order to improve the performance of passage retrieval in log files and overcome this domain problems notably mismatch vocabularies. Query expansion (or query enrichment[1]) attempts to improve retrieval performance by reformulating and adding new correlated terms to queries. In general the idea is to add more terms to an initial query in order to disambiguate the query and solve the possible term mismatch problem between the query and the relevant document [20]. As a solution for query expansion, the relevance feedback, introduced in the mid-1960s, is a controlled, automatic process for query reformulation, that has been proved to be unusually effective [42]. More precisely relevance feedback is a powerful technique whereby a user can instruct an information retrieval system to find additional relevant documents by providing relevance information on certain documents or query terms [33]. The basic idea behind relevance feedback is to take the results initially returned from a given query and to use information about whether or not those results are relevant to reformulate a new query. All these methods will be detailed in Section 3.

Here we present a query expansion approach using an adapted relevance feedback process. Our approach enables us to improve the relevance of queries and thus the passage retrieval performance in despite of the studied corpus' difficulties. Our approach of query expansion, based on relevance feedback, involves two levels. In the first one, we implement an explicit relevance feedback system. The feedback is obtained from a training corpus within a *supervised learning approach*. We propose a new method for learning the *context* of questions (queries), based on the "*lexical world*" notion [16,2]. Then, the contexts of questions are used as relevant documents wherein we look for expansion terms. The second phase consists of a novel kind of pseudo relevance feedback [25]. Contrary to most pseudo relevance feedback methods considering the initial top-ranked documents as relevant, our method is based on a *new term weighting function*, called TRQ,[2] which gives a score to terms of corpus according to their *relatedness* to the *query*. Indeed, we present the TRQ measure as an original term weighting function which aims at giving a high score to terms of the corpus which have a significant probability of existing in the relevant passages.[3] Terms having the highest TRQ scores are selected as expansion terms.

We also evaluate the application of our approach in general domains. We thus use the documents used in TREC[4] evaluation campaigns. We study the difference between the application of our approach in specific and general domains. We show that our approach gives *satisfactory* results on *real data* from the industrial field as well as general domains.

In Section 2, we present the main characteristics of log files which raise some challenges in passage retrieval. Existing work about passage retrieval systems and the application of relevance feedback in the query expansion are presented in Section 3. Section 4 presents some notions used in our query expansion approach and also the first level of query enrichment. In Section 5, we develop our query expansion approach by presenting our novel term weighting function. Section 6 is devoted to developing the application of our approach in open domains. Finally, the experiments on real data are presented in Section 7.

2. Difficulties in passage retrieval in log files

The particular characteristics of logs, described below, give rise to some challenges in passage retrieval and information extraction in log files. Here, by presenting these challenges and difficulties, we explain how they led us to query expansion as a solution.

2.1. Vocabulary mismatch

First, we focus on problems arising from the multi-source aspect of log files. In the design of Integrated Circuits, different design tools can be used at the same time, while each tool generates its own log files. Therefore, although the logs of the same design level contain the *same* information, their *structure* and *vocabulary* can vary significantly *depending* on the *design tool used*. In other words, each design tool has its own vocabulary to report the same information. This implies that queries which are expressed using a vocabulary could not necessarily correspond to the vocabulary of all tools, or the query terms do not necessarily exist in the corresponding answers. We explain this issue, called here mismatch vocabularies, by the help of an example.

Consider the sentence "`Capture the total fixed STD cell`" as a given query and the two log files, log_A and log_B generated by two different tools, as the data resources wherein we look for answers. The answer to this question, in log_A, is expressed in the third line of its following segment.

[1] We use query expansion and query enrichment interchangeably in this paper.

[2] Term Relatedness to Queries.

[3] Passages containing answers to questions.

[4] http://trec.nist.gov/.