Query expansion using Language Models

Flavio Forenza

August 7, 2021

Abstract

Un abstract

INTRODUCTION

Over the years, query expansion techniques have been proposed as a solution to the problem of term mismatches between a query and its relevant documents. Their goal is to enrich the number of words present in the user's dictionary, so that they match those present in the corpus of documents in which to search. There are typically two types of query expansion method families; Local (based on Pseudo / Relevance / Indirect Feedback) and Global (based on the generation and use of a thesaurus) [5]. This paper focuses on the first category. Given the difficulty in gathering the users' feedback, only the first documents recovered will be considered relevant. Pseudo-relevant documents are used to find possible candidate terms to help expand the query [3]. This method has been further developed within the concept of

the Language Model [4]. Statistical language models are widely used within Information Retrieval as they have a solid theoretical background and good empirical performance. Clearly, the two main problems relate to correctly estimating both the query model and the document model. A Language Model calculates the relevance of a document d to a query q by estimating a factored form of the distribution $P(q, D)$ [2]. The construction of a good Language model must necessarily make use of smoothing models when one or more terms do not appear in a document. In the latter case, the maximum likelihood estimator would produce a probability equal to zero, invalidating the creation of the model itself [1] [6]. Another concept, useful for expanding the query and widely used within the project, is that of *Word Embeddings*. The latter is obtained precisely from the use of Language models, or rather thanks to the co-occurrence of the terms made available. This method is based on being able to map every single word into a vector of real numbers, within a vector space. The idea is to be able to compare the distance of these in order to understand their similarity relationship. If one word is similar to another, then these will be considered as synonyms. The remainder of the paper is organized as follows. In the next section, *Research question and Methodology*, the objectives of the project will be introduced followed by an overview of the proposed approach. The third section, *Experimental result*, describes the whole system and the results obtained. Finally, the conclusions are presented in section four, *Concluding remarks*.

0.1 INTRODUCTION

Qualcosa [**?**].

References

[1] G.Alexander. *Log-Linear Interpolation of Language Models.* PhD thesis, 2000.

[2] H.Zaragoza, D.Hiemstra, and M.Tipping. Bayesian extension to the language model for ad hoc information retrieval. page 4–9, 2003.

[3] J.Lafferty and C.Zhai. Document language models, query models, and risk minimization for information retrieval. page 111–119, 2001.

[4] J.M.Ponte and W.B.Croft. A language modeling approach to information retrieval. page 275–281, 1998.

[5] R.Cummins and C.O'Riordan. Evolving co-occurrence based query expansion schemes in information retrieval using genetic programming. pages 137–146, 2005.

[6] S.F.Chen and J.Goodman. An empirical study of smoothing techniques for language modeling. page 310–318, 1996.