Query expansion using Language Models

Flavio Forenza

August 13, 2021

Abstract

The use of different methods of language modeling, within the field of information retrieval, is finding a wide diffusion in the state of the art. Based on the accuracy of the language model, the problem related to the information retrieval, in a large corpus of documents, can be solved. In order to do this, the basic idea of these approaches is to estimate a probabilistic linguistic model, for each document in the collection, which is able to generate a ranking of relevant documents given a query. One of the problems that afflicts this family of methods is due to the lack of data present. From this, it is necessary to apply smoothing techniques capable of adjusting the maximum likelihood estimator in order to correct the generated imprecision. This paper shows how their application outperforms the performance of classic methods, such as *tf-idf*, useful for generating rankings of documents ordered by relevance. Using them, we will look at some concepts that are useful for query expansion.

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. 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) [7]. 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 [5]. This method has been further developed within the concept of the Language Model [6]. Statistical language models are widely used within Information Retrieval as they have a solid theoretical background and good empirical performance. At the state of the art, two well known probabilistic approaches in infromation retrieval are the Robertson and Sparck Jones model [8] and the Croft and Harper model [11]. Both models estimate the probability of relevance of each document to the query. 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 $P(q, D)$ [4]. 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 [2] [9]. 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*.

Research question and Methodology

Language models can be used for a variety of purposes, such as: speech recognition, spelling correction, grammar correction and automatic translation. All these applications have the task of assigning a probability to a sequence of words, based on the number of times they appear in one or more documents. As briefly mentioned in the introductory chapter, the purpose of the following project is to verify the existence of a further method, which makes use of the concept of language model, specifically an *n-grams* model, capable of expanding a query. Achieving this goal means solving the problem of mismatch between the terms present in the query and those present in a corpus of documents. The probability of generating a new q query given the estimate of a Language Model for a D document can only occur through a ranking of relevant documents. If the corpus of documents is large, thinking of generating n Language models, with n equal to the number of documents, turns out to be a computationally expensive operation. This paper has used a useful approach to be able to generate a first ranking of documents ordered by relevance with the query q. The technique applied is the *tf-idf* recovery model. The adoption of this method of weighing the terms, as well as being widely used in the state of the art, produces excellent results. The introduction of the *LM*, and of other semantic analysis techniques, made it possible to outperform the performance of the *tf-idf* baseline, generating ranking, starting from the latter, of documents more pertinent to the query q [10]. In the following paper, tests are carried out which certify the veracity of this thesis. It should be noted that the generation of the ranking, obtained from the weights of the *tf-idf* method, is obtained using the well-known *cosine similarity* metric between the weight vectors. To comply with the set objective, the position of the relevant target document, that is the document that the user is searching for, was kept track. This was possible because a query was chosen, among those available, that was close to the title of this document. The score assigned to the target document will represent the minimum *threshold* to be able to form the new ranking of documents. This step is fundamental as, by setting a higher threshold, the target document would be lost in subsequent calculations. By setting a lower threshold, however, those documents that represent noise will be taken into account. From this ranking, the *LMs* for each document will be calculated [1], generating n LMs, with n equal to the number of relevant documents, based on the terms in the query. It should be noted that, in order to carry out this step, the concept of *skip-gram* has been applied to the query, with step s equal to two. The creation of each LM is possible only after using one of the existing smoothing techniques. To prevent a linguistic model from assigning zero probability to an invisible event, i.e. when a term present in the query is not present in an LM, we should eliminate some probability mass from some more frequent events and give it to events that we haven't never seen. There are a variety of methods for smoothing, some of these are: *Laplace (add- one) smoothing*, *Linear Interpolation smoothing*, *add-k smoothing*, *back-off smoothing* and *Kneser-Ney smoothing*. Among these, the following paper has experimented the use of the first two smoothing methods, each of which will produce different results useful for achieving the final goal. The core of the algorithm lies in being able to derive the best ranking of relevant documents, using the initial query, through an iterative process, as s changes. This variation will lead to the generation of several LMs, each with step s, with $s = \{2, 3, \ldots, 10\}$. At each iteration, a new ranking of documents will be generated, thanks to the calculation of the Maximum Likelihood Estimation *MLE*, between the query q and each document d(1).

$$\begin{aligned} P(q|d) &\approx P(q|M_d) \\ &\approx \prod_{i}^{n} P(w_i|w_{i-1}) \\ &\approx \frac{count(w_i, w_{i-1})}{\sum_{j=1}^{n} count(w_j, w_{i-1})} \\ &= \frac{count(w_i, w_{i-1})}{count(w_{i-1})} \end{aligned} \quad (1)$$

Among all these nine rankings, after being sorted, the one that will have the target document in the highest position, close to the first, will be chosen. Also thanks to this, it will be possible to determine the lambda parameters present in the interpolation smoothing technique. By focusing on the latter technique, it was necessary to implement the concept suggested by [3] as regards the calculation of linear interpolation. Instead of adding one to the probability calculation, as is well known in Laplace's smoothing (2), linear interpolation, recursively, calculates the MLE of order two (*bi-grams*) (3), up to the order zero (*zero-grams*)(4).

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1}) + 1}{count(w_{i-1}) + |V|} \quad (2)$$

$$P(w_i|w_{i-1}) = \lambda P(q|M_d) + (1 - \lambda)P(q|M_c) \tag{3}$$

$$P(w_i) = \lambda \frac{1}{|V|} + (1 - \lambda)P(w_i) \tag{4}$$

where:

- $\sum_i \lambda_i = 1$
- M_d: represents the language model of the single document;
- M_c: represents the language model of the entire collectionof documents;
- $|V|$: represents the number of unique words within the corpus of documents.

References

[1] C.D.Manning, P.Raghavan, and H.Schütze. *Introduction to Information Retrieval*, chapter 12. Cambridge University Press, 2008.

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

[3] A. Gutkin. *Log-Linear Interpolation of Language Models*. PhD thesis, 11 2000.

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

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

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

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

[8] S.E.Robertson and K.Spärck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, page 129—46, 1976.

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

[10] V.Lavrenko and W.B.Croft. *Relevance Models in Information Retrieval*, pages 11–56. Springer Netherlands, 2003.

[11] W.B.Croft and D.J.Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, pages 285–295, 1979.