Proceedings of TOGO 2010, pp. 1 – 4.

Optimization Algorithm for Improving the Efficacy of an Information Retrieval Model

Alberto Costa [1]

[1] *LIX, École Polytechnique, 91128 Palaiseau, France,* costa@lix.polytechnique.fr

Abstract The aim of Information Retrieval (IR) is to find the documents that are more relevant for a query, usually given by a user. This problem is very difficult, and in the last four decades a lot of different models were proposed, the most famous being the logical models, the vector space models, and the probabilistic models. In this paper is proposed a greedy algorithm for maximizing the efficacy of an Information Retrieval model based on Discrete Fourier Transform (DFT), which has shown a good efficacy level in the first tests. Even if the mathematical programming model used to increase the efficacy is a Mixed-Integer Nonlinear Program (MINLP), with nonlinear objective function and binary variables, its structure is very simple and a greedy algorithm can find the optimal solution.

Keywords: Information Retrieval, Efficacy, Greedy Algorithm, Discrete Fourier Transform

1. Introduction

Information Retrieval (IR) tries to solve this problem: given a query and a set of documents (collection), which are the relevant documents for the query?

The efficacy of an IR model depends on the number of relevant and non-relevant documents retrieved: the "perfect" IR model (that is the model with the maximum efficacy) should be able to retrieve all and only the relevant documents. Each time a non-relevant document is retrieved, or a relevant document is not retrieved, the efficacy decreases.

Several models were proposed in the last four decades, such as the logical models, the vector space models and the probabilistic models. The different techniques proposed by these models produced a significant increase of retrieval effecteveness in the last fifteen years, as experimentally observed within the Text REtrieval Conference (TREC) [8]. However, the current technology is far from being optimal, and the quest for new theoretical frameworks has been intense [3–5, 7].

Recently, a new IR model based on Discrete Fourier Transform (DFT) called Least Spectral Power Ranking (LSPR) was proposed, and it has shown good efficacy level in the first tests [2]. In this model the input is a collection of a document and a query, while the output is the ranking list, that is the list of the retrieved documents ordered from the most to the least relevant. The important thing to remark is that each document is associated with a score (called power in the LSPR model), such that if a document has a low power, it is considered highly relevant by the system, hence, the documents are ordered by increasing power. This is why the model is called Least Spectral Power Ranking.

In this paper an algorithm is proposed, which tries to increase the efficacy of LSPR: starting from the ranking list, this algorithm removes the documents that are not relevant with high probability. Basically, the problem of choosing the documents that maximize the efficacy can be described as a Mixed-Integer Nonlinear Program (MINLP), with a quadratic objective function

and binary variables. However, due to the structure of the problem, a simple greedy algorithm can find the optimal solution.

The remainder of the paper is organized as follows. In section 2 there is a more formally explanation of the concept of efficacy. After that, in section 3 is presented the algoritm. Finally in section 4 there are the conclusions.

2. Evaluation of an IR system

The most important parameters for evaluating an IR system are:

- efficiency, that refers to the time complexity and the memory occupation;
- efficacy, that refers to the quality of the results.

In order to evaluate the efficacy, the so called "experimental collections" were introduced. An experimental collection is composed of a collection of documents, a set of queries and the relevance judgements; the latter is the list of the relevant documents for each query. In this way, comparing the documents retrieved by the system with the relevance judgements, it is possible to have an indication about the efficacy.

Among the most used parameters there are:

- precision: ratio between the number of relevant documents retrieved and the number of retrieved documents; is a measure of accuracy of search,
- recall: ratio between the number of relevant documents retrieved and the number of relevant documents; is a measure of completeness of search.

It is easy to see that if the number of document retrieved increases, the precison decreases and the recall increases.

In recent years, other measures have become more common, such as the Mean Average Precision (MAP), that is the average of the precision value obtained for the top k documents, each time a relevant document is retrieved, or graphically it is roughly the average area under the precision-recall curve for a set of queries. The MAP varies from 0 to 1; in the tests performed in [2], using the CACM experimental collection,[1] the MAP of the vector-space model was 0.242, while the MAP of DFR was 0.329. The MAP of LSPR was 0.348, thus indicating a good performance comparable to the state-of-the-art.

3. Greedy algorithm

In this section an algorithm for increasing the MAP of the LSPR model is described.

Suppose there are a query Q and a collection C as input for LSPR, and the output is ranked list R, whose each document i is associated with a power Pw_i: the less the power, the more the relevance.

The first step is to compute, for each document i in the ranking list, a probability p_i to be relevant. The informations given by LSPR can be very useful for this scope. Let Pw_m and Pw_M be respectively the power associated with the first and the last document in the ranking list R; a simple way to compute p_i can be the following:

$$p_i = \frac{Pw_i - Pw_M}{Pw_m - Pw_M}. \tag{1}$$

It is easy to see that the probability is from 0 to 1: 0 if the power of the document is Pw_M (that is the last document retrieved), 1 if the power of the document is Pw_m (that is the first document retrieved.

[1] The collection can be found on http://www.search-engines-book.com/collections.

In order to maximize the MAP, we should maximize both precision and recall, leading to a multiobjective model. Furthermore, the recall depends on the number of relevant documents for a query, but usually this information is not available, so we have to semplify the problem.

Let $x_i \in \{0,1\}$ be a variable that is 1 if the document i is selected, 0 otherwise. Precision and recall are rounded respectively as:

- precision= $\frac{\sum_{i=1}^{|R|} p_i x_i}{\sum_{i=1}^{|R|} x_i}$
- recall= $\frac{\sum_{i=1}^{|R|} p_i x_i}{N(Q)}$

where $N(Q)$ is the unknown number of relevant documents for the query, and the sums at the numerators play the role of the number of relevant documents retrieved.

At this point it is possible to semplify both the problem of the multiobjective function and the unknown value of $N(Q)$ by maximizing the product of recall and precision. Thus, the objective function to maximize is the following:

$$f(x) = \frac{\left(\sum_{i=1}^{|R|} p_i x_i\right)^2}{N(Q) \cdot \sum_{i=1}^{|R|} x_i}. \tag{2}$$

Since $N(Q)$ is a constant number, even if unknown, we can remove it from the objective function. The final MINLP model is

$$\begin{aligned} \max \quad & \frac{\left(\sum_{i=1}^{|R|} p_i x_i\right)^2}{\sum_{i=1}^{|R|} x_i} \\ \texttt{s.t.} \quad & x_i \in \{0,1\} \quad \forall i \in \{1,2,\ldots,|R|\} \end{aligned}$$

The greedy algorithm that solves this problem is very simple: first, the documents are ordered by decreasing probability to be relevant (i.e. $p_i \geq p_{i+1}$, $\forall i \in \{1,2,\ldots,|R|-1\}$). After that, we try to add the documents, from the first to the last in the ordered list, until the objective function increases. As soon as the objective function decreases, the algorithm stops; this is summarized in the following pseudo-code.

greedy_select {
  ⋆ call LSPR, to get the ranking list R of the documents and the powers ⋆
  ⋆ starting from the powers, compute the probability, for example using Eq. (1) ⋆
  ⋆ order the documents by decreasing probability to be relevant ⋆
  $f \leftarrow 0$
  $x_i \leftarrow 0$, $\forall i \in \{1,2,\ldots,|R|\}$
  for $d \leftarrow 1$ to $|R|$ do
  {
    $x_d \leftarrow 1$
    $f_d \leftarrow \frac{\left(\sum_{i=1}^{d} p_i x_i\right)^2}{\sum_{i=1}^{d} x_i}$
    if $(f_d < f)$
      $x_d \leftarrow 0$
      break
    else
      $f \leftarrow f_d$
  }
  ⋆ return the documents i for which $x_i = 1$, ordered by increasing p_i ⋆
}

4. Conclusion and future work

This paper presents a possible way to increase the efficacy of an IR model. Actually, this technique could also be the base of a stand-alone IR model.

The important thing is, for each document, the computation of the probabilities to be relevant. In this paper a simple idea is proposed (see Eq. (1)), but more sophisticated techniques, which take account of the distribution of the powers, should led to better results.

Removing the documents from the ranking list has other advantages: the user has less document to check, and also the efficiency increases, because the memory occupation for the list of documents decreases.

Future work has 2 main objectives: First, this idea needs to be tested with some experimental collections. Second, a more precise mathematical description of the precision and the recall, and consequentely of the MAP, should be found. In this way the efficacy should increase, even if the model probably will be solved by some Nonlinear Global Optimization solver, such as BARON [6] or Couenne [1], instead of a simple greedy algorithm.

Acknowledgments

The author is grateful to Digiteo Project 2009-55D "ARM" for financial support.

References

[1] P. Belotti, J. Lee, L. Liberti, F. Margot, and A. Wächter. Branching and bounds tightening techniques for non-convex MINLP. Optimization Methods and Software, 24(4):597–634, 2009.

[2] A. Costa and M. Melucci. An information retrieval model based on discrete fourier transform. In H. Cunningham, A. Hanbury, and S. Rüger, editors, Proceedings of the 1st Information Retrieval Facility Conference, volume 6107 of Lecture Notes in Computer Science, pages 84–99, Vienna, 2010. Springer, Heidelberg.

[3] W.B. Croft and J. Lafferty, editors. Language Modeling for Information Retrieval. Springer, 2003.

[4] N. Fuhr. A probability ranking principle for interactive information retrieval. Journal of Information Retrieval, 11(3):251–265, 2008.

[5] S. Robertson. Salton award lecture: On theoretical argument in information retrieval. SIGIR Forum, 34(1):1–10, 2000.

[6] N.V. Sahinidis and M. Tawarmalani. BARON 7.2.5: Global Optimization of Mixed-Integer Nonlinear Programs, User's Manual, 2005.

[7] C.J. van Rijsbergen. The geometry of information retrieval. Cambridge University Press, 2004.

[8] E.M. Voorhees and D.K. Harman, editors. TREC: Experiment and evaluation in information retrieval. The MIT Press, Cambridge, MA, USA, 2005.