

Words Temporality for Improving Query Expansion

Olga Craveiro^{1,2}, Joaquim Macedo³, and Henrique Madeira²

¹ School of Technology and Management, Polytechnic Institute of Leiria, Portugal

² CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
{marine, henrique}@dei.uc.pt

³ Algoritmi, Department of Informatics, University of Minho, Portugal
macedo@di.uminho.pt

Abstract. There is a lot of recent work aimed at improving the effectiveness in Information Retrieval results based on temporal information extracted from texts. Some works use all dates but others use only document creation or modification timestamps. However, no previous work explicitly focuses on the use of dates within in the document content to establish temporal relationships between words in the document. This work estimates these relationships through a temporal segmentation of the texts, exploring them to expand queries. It was achieved very promising results (13% improvement in Precision@15), especially for temporal aware queries. To the best of our knowledge, this is the first work using temporal text segmentation to improve retrieval results.

Keywords: Query Expansion, Temporal Information Retrieval, Temporal Text Segmentation.

1 Introduction

A large range of Information Retrieval (IR) tasks can benefit by incorporating temporal information [2]. The novelty or temporal incidence of information is obviously an important aspect in the assessment of its quality, importance and relevance.

In Web IR, time was proven to be valuable for crawling [8], clickstream analysis [11], and determine temporal snippets for summarize documents [1]. Even being a low percentage of total search engine queries, there is a huge number of temporal aware information needs [13]. So, there is several research work exploring temporal information to improve the effectiveness of IR results. There are time-aware language models [12] and similarity formulas with temporal components in vector space models [10]. There is work concerned with searching future events [5]. In query expansion (QE), one idea is to consider that if exists a large set of query results documents with same date, they are more likely to be relevant. The work of Amodeo et al. [4] relates the publication date of documents and their relevance for a given query to select the pseudo-relevant documents. The proposal of Whiting et al. [16] is to select better the

terms, considering their temporal profile correlation from the pseudo-relevant documents and temporally significance.

This work establishes relationships between words based on temporal segmentation of texts. Terms co-occurring with original query terms in the same segment of pseudo-relevant documents are positively discriminated in QE process.

Section 2 describes the used temporal segmentation tool for Portuguese texts. Section 3 presents the three QE methods based on temporal segmentation. Section 4 presents the evaluation of these methods using the CHAVE collection [15]. Then, we discuss the obtained results and possible future research issues.

2 Temporal Segmentation of Texts

Since temporal segmentation of texts has a crucial importance to this work, we present a brief description of the followed approach, detailed in our prior work [9].

Prior to segmentation, temporal expressions found in the content of documents are recognized, annotated and normalized, as far as possible, linked to *chronons*. *Chronons* are dates with a given granularity (year, month, day or hour) and a mark on a timeline. Too vague temporal expressions cannot be transformed into a *chronon*.

As one can model a text as a narrative of a sequence of different topics, it makes sense to think that it describes a chronology of events that occur at different instants of time. So, an alternative to topical segmentation [7] is temporal segmentation [6], which the underlying idea is the detection of temporal discontinuities in the text. Instead of subject change, it is amended the date or time of the narrative.

Our segmentation algorithm uses temporal information extracted from the text to divide it into temporally coherent segments. These segments must be tagged with *chronons* in their sentences (including document timestamp), to obtain an association between time and document terms. The length of a segment ranges from a single sentence to a multi-paragraph text. Thus, adjacent sentences with the same *chronons* must belong to the same segment.

3 Time-Aware Query Expansion

Our proposal incorporates the word temporality in the automated QE models. Fig. 1 shows a general algorithm of QE, considering a broad view of the QE process. The two important parameters of the algorithm are: k , the number of relevant documents for the original query q_0 , and n , the maximum number of terms to in the expanded query qE . The formulae used in the computation of the scores (steps 3 and 5) are based on the formula of Rocchio [27], which need the definition of the parameter α . Roughly speaking, the algorithm can be described as follows: first, it is necessary to obtain a set of terms T_I to be included in the expanded query qE ; and then, the original query q_0 is reformulated considering these terms and their scores.

The three proposed methods are based on the assumption that the terms that co-occur in the same temporal segment with original query terms (T_0) must assume a