

Query Expansion with Temporal Segmented Texts

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. The use of temporal data extracted from text, to improve the effectiveness of Information Retrieval systems, has recently been the focus of important research work. Our research hypothesis is that the usage of the temporal relationship between words improves the Information Retrieval results. For this purpose, the texts are temporally segmented to establish a relationship between words and dates found in texts. This approach was applied in Query Expansion systems, using a collection with Portuguese newspaper texts. The results showed that the use of the temporality of words can enhance retrieval effectiveness. In particular for time-sensitive queries, we achieved 9.5% improvement in Precision@10. To 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

The importance of time in Information Retrieval has been studied from a various perspectives including document clustering, searching in time, Web archiving, etc. However, there is still much to do to achieve its full integration in the most popular retrieval models [1]. Some approaches only use the documents' metadata, such as document's timestamp, while some others also consider the temporal information found in the documents' content. However, none of them establishes an association between the words and the time, as it is proposed in our approach.

Establishing a relationship between words and time, we can temporally relate the words and make our interpretation of texts richer, allowing the identification of relations between entities, facts or events described by documents.

One of the possible applications for which words temporality could be of benefit is automated query expansion, henceforth QE. Traditionally, in QE, the relevant retrieved documents of an initial query are the top-ranked documents. Some approaches also incorporate the time dimension to improve traditional QE, although they only consider the documents' timestamps, unlike our approach which also uses the temporality of each word found in documents.

The work of Amodio et al. [3] 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. [10] is to get the terms selection better, considering the temporal profile correlation between terms from the pseudo-relevant documents and temporally significant terms.

In the following, we present a brief description of the temporal segmentation, which is the base of this work. Section 3 describes three methods to incorporate temporal information into QE systems. Experiments and results are reported in Section 4 and we conclude in Section 5.

2 Temporal Segmentation of Texts

The temporal segmentation of texts has a crucial importance in this work. So, we present a brief description of the followed approach, detailed on our prior work [7].

The used segmentation is based on temporal discontinuities found in texts. For a given segment, dates and words establish relationships, later used on our QE methods.

The temporal segmentation partitions the text into coherent segments as in topic segmentation, but these segments are tagged with timestamps. In this way, the words of the segment become associated to the timestamps of the same segment.

In fact, the timestamps are the *chronons* collected from the text which belongs to the segment. A *chronon* is a normalized date which is anchored in a calendar/clock system by timelines defined by points. The timeline can have different levels of granules, such as year, month, week, day, and hour. To obtain the *chronons*, we developed a system which extracts the temporal information found in Portuguese texts, putting it into the *chronon* format. A detailed description of this system can be found in [5, 6].

A temporal segment is defined as a set of adjacent sentences that shares the same temporal focus. The segment length ranges from a single sentence to a multi-paragraph text. Thus, adjacent sentences with the same *chronons* must belong to the same segment.

The boundaries of a segment are firstly defined, by the temporal information found in the segment, unless this information does not exist in the text. In that case, the boundaries are identified by the topic changes using the cosine similarity formula.

Fig.1 presents two examples of temporal segmentation. In the first example, the segment “A *tempestade de Domingo* causou alguns problemas nas redes de energia eléctrica. A empresa de energia eléctrica recebeu cerca de 31 mil chamadas.”¹ is composed of two sentences. The first one has a date (*Domingo*) which was resolved by our system of temporal information extraction. In the next sentence, the topic stays the same, so that sentence will also belong to this segment.

The other segment “*Choveu na sexta-feira e no sábado.*”² is tagged with two *chronons*, because these two normalized dates are in the same sentence.

¹ English version: **Sunday**'s storm caused some problems in electricity networks. The company of electricity received about 31,000 calls.

² English version: It rained on **Friday** and **Saturday**.