

Time-Aware Focused Web Crawling

Pedro Pereira¹, Joaquim Macedo¹, Olga Craveiro^{2,3}, and Henrique Madeira³

¹ Centro Algoritmi/Dep. of Informatics, University of Minho, Portugal

² ESTG, Polytechnic Institute of Leiria, Portugal

³ CISUC/Dep. of Informatics Engineering, University of Coimbra, Portugal

Abstract. There is a plethora of information inside the Web. Even the top commercial search engines can not download and index all the available information. So, in the recent years, there are several research works on the design and implementation of focused topic crawlers and also on geographic scope crawlers.

Despite other areas of information retrieval, research on Web crawling is not using the temporal information extracted from Web pages in the used crawling criteria. Therefore, our research challenge is the use of temporal data extracted from Web pages as the main crawling criteria to satisfy a given temporal focus. The importance of the time dimension is quite amplified when combined with topic or geography, but now we want to study it isolated. The used approach is based on temporal segmentation of Web pages text. It only follows links within segments tagged with dates in the scope of restriction. A precision around 75% was achieved in preliminary experimental results.

Keywords: Web Crawling, Temporal Text Segmentation, Temporal Information Extraction, Temporal Information Retrieval.

1 Introduction

Crawling the Web is an old problem that it was subject of extensive research [1] and is widely used today, mainly by search engines like Google, Yahoo! or Bing.

These crawlers try to cover a significant part of the Web, looking for information to build their indexes. Whenever users submit queries, these indexes are processed and the search engine returns responses that include URLs pointing to supposed relevant documents.

The huge size of the Web points to the need of the decentralization of the crawling process, based for instance in geographical partition of the Web [2]. Additional criteria are used to drive the crawling process like Web page rank, freshness, topic focus, geographic scope for overcome such scalability problems. This work introduces the temporal scope, based on dates extracted from Web pages content, as a new crawling criteria.

Temporal information is also be used for many different purposes, such as searching future events [3], clustering search results with timelines [4], defining snippets based on temporal information [5]. Nunes et al. [6] proposed an approach

of web pages timestamping, to determine the publication or modification date of a web page. Yu et al. [7] modified the Page Rank algorithm, using the date of citation to improve the quality of the search results. Dai and Davidson [8] proposed a link-based ranking method considering the freshness of the page content.

In section 2, a strategy of temporal segmentation of Web pages text is presented. Later, this segmentation is used by a crawler to download Web pages, using temporal constraints (section 3). Finally (sections 4 and 5) ,some promising preliminary results and directions for future work are presented.

2 Temporal Segmentation of Text

The main objective of the text segmentation is to divide text into smaller units according to many different criteria. The proposed approach follows a segmentation algorithm based on temporal discontinuities identified in the text [9]. A temporal segment is a set of contiguous sentences or paragraphs sharing the same temporal focus. The identification of the temporal boundaries is given by the temporal information found in the text which could be mapped in chronons with normalized dates anchored in a calendar/clock system [10]. If two adjacent sentences do not have the same chronons, then there is a temporal discontinuity. In this case, these sentences must belong to different segments. Thus, adjacent sentences with the same chronons must belong to the same segment. In the absence of this normalized temporal information, this algorithm follows the approach of traditional topic segmentation. Figure 1 shows an example of temporal segmentation. The thick rectangle shows the document timestamp.

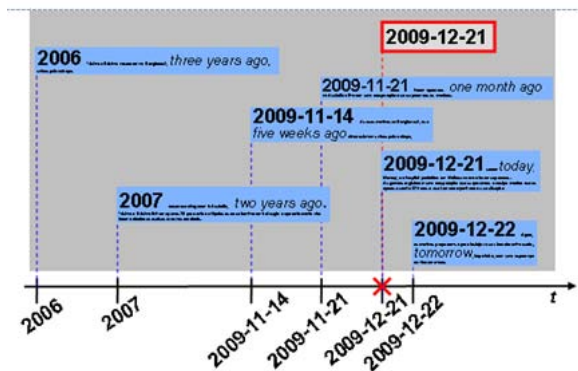


Fig. 1. Temporal Segmentation of a Text [9]