Time-Aware Focused Web Crawling

Pedro Pereira[1], Joaquim Macedo[1], Olga Craveiro[2,3], and Henrique Madeira[3]

[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

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 534–539, 2014.

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.

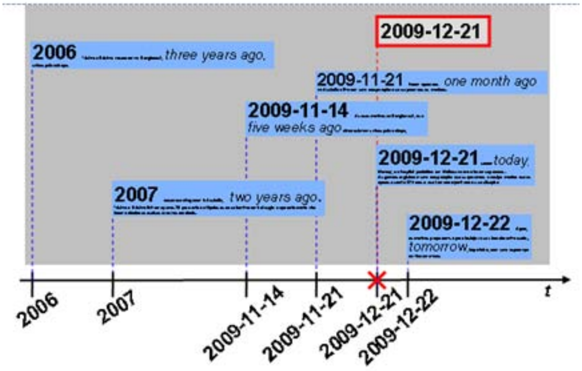

Fig. 1. Temporal Segmentation of a Text [9]

3 Temporal Crawler

A time-aware crawler must traverse the Web in search of informations which are within a given temporal scope. The downloaded pages are analyzed temporally to verify if they are within the temporal scope. The start hypothesis is that a Web page (or a segment of a page) within a given temporal scope has, with high probability, links to to Web pages with the same temporal scope.

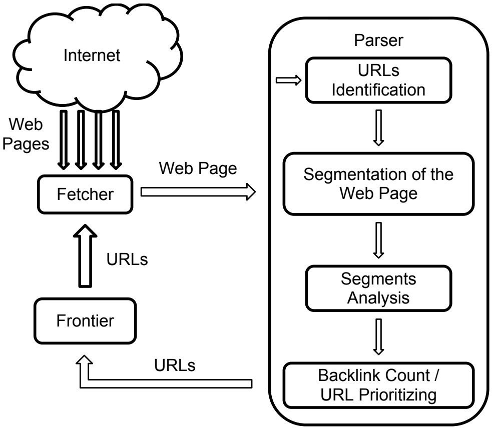

Fig. 2. Temporal Crawler Architecture

The developed time-aware crawler is no more than a general crawler with embedded temporal analysis capabilities in specific modules (see Figure 2). Its implementation was based on the *Crawler4j* [11], an easily modifiable generic crawler. The temporal analysis is done in two steps. First, the web page is temporally segmented. The Web documents are processed by the tool mentioned in section 2. This transforms a Web document into a XML document divided into temporal segments. Note that each segment is labelled with one or more dates[1].

Then, the existing links are classified in two categories: those inside the temporal focus and the remaining ones. The first set is eligible for download (valid URLs) and the others URLs are ignored.Furthermore, the valid URLs are added to a backlink list. This list has, for each page, the number of valid URLs pointing to it. This count is used to ordering the pages sent to the Frontier for download.

4 Experimental Results

To evaluate this crawler we will use precision and recall. In this preliminary evaluation we used only pages from Portuguese Wikipedia domain

[1] Most segments have a single date.

(http://pt.wikipedia.org/). This choice may be considered inappropriate[2], but we run the risk for this first proof of concept. Wikipedia is a domain which offers more guarantees in terms of quality of the pages with temporal data. Also, we selected two important events in our recent history including the Second World War (1939-1945) and the terrorist attacks at 9/11/2001.

We used Google to select 10 seeds for each scope. Along the topic, the queries include the word Wikipedia for only Portuguese language documents. Our segmmenter only supports Portuguese language. The used seeds are omitted due to space restrictions. This technique will be used normally combined with a topic or even a geographical scope. So, it is expected to avoid the consideration that such seeds cause a bias on the results.

As output the crawler generates a list with URLs, ordered by download time. For each event, we perform two runs. In the first one, the general crawler downloaded 5000 pages without any time restriction. In the second one, we used the time-aware crawler with the temporal scope (1939-1945 for WW2 and 2001 for 9/11/2001). At end, each run provides 5000 fully segment files and a URL crawl list. For general crawler the segmentation was a post-processing operation.

A file is considered valid, when at least one segment can be placed inside the temporal scope. To compute the total amount of positive files (required to calculate recall), we considered total valid files crawled by the temporal crawler plus the ones that were crawled by the general crawler. So, the recall@N is the quotient between the actual valid files and the total ones. For the precision@N are used the percent of valid files crawled.

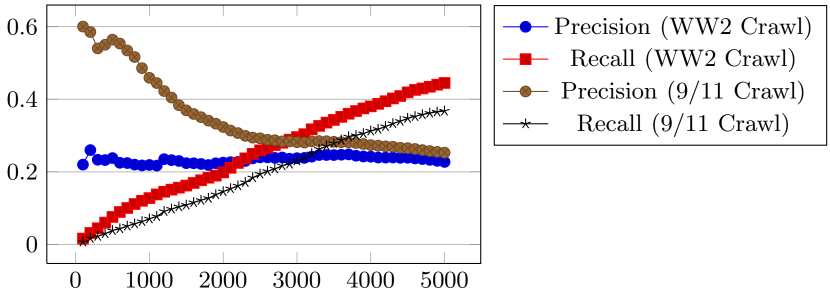

Fig. 3. Precision and Recall of the General Crawler

The Figure 3 shows that the precision for generic crawl using the WW2 seeds is very low throughout the entire crawl. The seeds do not give to the crawler a good start because, as we can see, the precision value at merely 100 files is very low. In the 9/11 crawl, for the first 800 pages the precision is above the 50%. The fairly good results in the first few hundred files do not mean anything. After there is a steep decline in the precision results. Both crawls finish around

[2] Wikipedia is a good representation for common web page?

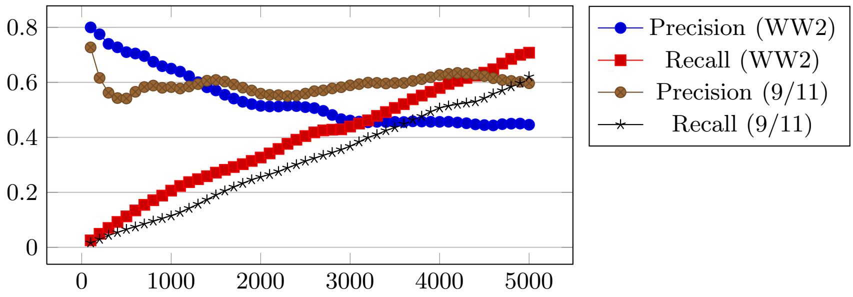

Fig. 4. Precision and Recall of the Temporal Crawler

the 25% precision. The recall value is similar for both crawls, finishing a around 40% for the 5000 files.

As we can see in the Figure 4, for the WW2 crawl, we begin with a precision of exactly 80% at the 100 files mark then it starts to descend. After that descent, we see just a little climb at the 2000 files stabilizing at the 50% mark but then the precision begins to fall and reaches the 50% precision at the 2600 files mark, stabilizing around the 45% precision mark.

If we look to the 9/11 crawl we do not see the same slow descent. The precision results at the 100 files mark is over 70% but then the precision begins to drop. At around 400 files the precision hits the lowest point and then climbs to around 60% and stays like that throughout the entire crawl. This climb and stabilization above the 50% may be due to the recent temporal scope (2001), making it easier to find recent information.

The bad precision results of the 9/11 crawl (comparing to the WW2 crawl) may be caused by the restrictive temporal scope imposed. In the WW2 scope we had a 6 year interval but in the 9/11 scope we have no interval, we only have one year that must be found. This restrictive scope is reflected in the results it produces. Analyzing recall, we see that both crawlers have the same curve, being the WW2 crawl slightly better.

The Table 1 presents some observed statistics namely processing times (collection and page), size (collection and average for page), average number of segments and chronons per page, and number of chronons in temporal scope (TS).

Table 1. Statistics of the Crawled Collections

	Total Proc.	Page Proc.	Collection	Page	Seg/	Seg.	Chronon
	Time (m)	Time (s)	size (MB)	Size(KB)	Page	in TS	/Page
WW2	85	61.2	441.0	90.0	82.6	32.6	192.3
9/11	90	66.2	430.0	88.0	74.9	71.8	112.0

5 Conclusions

The main objective of this work is the creation of a time-aware crawler architecture that analyzes temporally the documents, using time as crawling criteria.

Even using as evaluation Web pages from Portuguese Wikipedia, we get promising results. The results from the two different crawls exhibit some variance. This means that more exhaustive experiments are necessary and preferably with the most common web pages.

Either way, it was confirmed our hypothesis, albeit with still preliminary results. Using pages or portions of pages (segments) within a given temporal scope as a starting point, and using the URLs included in them, it is most likely to find new pages with the same temporal scope.

Acknowledgments. This work is partially supported by Algoritmi and CISUC, financed by FCT - Fundação para a Ciência e Tecnologia, within the scope of the projects PEst-OE/EE/UI0319/2014 and PEst-OE/EE/UI0326/2014.

References

1. Olston, C., Najork, M.: Web crawling. Foundations and Trends in Information Retrieval 4(3), 175–246 (2010)
2. Exposto, J., Macedo, J., Pina, A.: Geographical partition for distributed web crawling. In: GIR (2005)
3. Baeza-Yates, R.: Searching the future. In: SIGIR Workshop (2005)
4. Alonso, O., Gertz, M., Baeza-Yates, R.: Clustering and exploring search results using timeline constructions. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 97–106. ACM, New York (2009)
5. Alonso, O., Baeza-Yates, R., Gertz, M.: Effectiveness of temporal snippets. In: WSSP Workshop, WWW 2009 (2009)
6. Nunes, S., Ribeiro, C., David, G.: Using neighbors to date web documents. In: WIDM, pp. 129–136 (2007)
7. Yu, P.S., Li, X., Liu, B.: On the temporal dimension of search. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, WWW 2004, pp. 448–449 (2004)
8. Dai, N., Davison, B.D.: Freshness matters: in flowers, food, and web authority. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, pp. 114–121 (2010)
9. Craveiro, O., Macedo, J., Madeira, H.: It is the time for portuguese texts! In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012. LNCS, vol. 7243, pp. 106–112. Springer, Heidelberg (2012)
10. Craveiro, O., Macedo, J., Madeira, H.: Leveraging temporal expressions for segmented-based information retrieval. In: ISDA 2010, pp. 754–759 (2010)
11. Crawler4j website, `https://code.google.com/p/crawler4j/` Technical report