informationresearch

VOL. 18 NO. 3, SEPTEMBER, 2013

The impact of time in link-based Web ranking

Sérgio Nunes, Cristina Ribeiro and Gabriel David
INESC TEC & Department of Informatics Engineering, Faculty of Engineering, University of Porto, Portugal

Abstract

Introduction. *The strong dynamic nature of the Web is a well-known reality. Nonetheless, research on Web dynamics is still a minor part of mainstream Web research. This is largely the case in Web link analysis. In this paper we investigate and measure the impact of time in link-based ranking algorithms on a particular subset of the Web, specifically blogs.*

Method. *Using a large collection of blog posts that span more than three years, we compare a traditional link-based ranking algorithm with a time-biased alternative, providing some insights into the evolution of link data over time. We designed two experiments to evaluate the use of temporal features in authority estimation algorithms. In the first experiment we compare time-independent and time-sensitive ranking algorithms with a reference rank based on the total number of visits to each blog. In the second, we use feedback from communication media domain experts to contrast different rankings of Portuguese news Websites.*

Results. *The distribution of citations to a Web document over time contains valuable information. Based on several examples we show that time-independent algorithms are unable to capture the correct popularity of sites with high citation activity. Using a reference rank based on the number of visits to a site, we show that a time-biased approach has a better performance.*

Conclusions. *Although both time-independent and time-aware approaches are based on the same raw data, the experiments indicate that they can be treated as complementary signals for relevance assessment by information retrieval systems. We show that temporal information present in blogs can be used to derive stable time-dependent features, which can be successfully used in the context of Web document ranking.*

change font

Introduction

The World Wide Web is a particularly dynamic medium. Several studies have observed and documented this dynamic nature (Ntoulas *et al.* 2004; Fetterly *et al.* 2004; Adar *et al.* 2009). A recent study by Adar *et al.* (2009) has shown that popular pages exhibit a very high change rate, with approximately 40% of the pages in the sample changing nearly every hour. In a previous study, Ntoulas *et al.* (2004) found that, after one year, 50% of the content on the Web is new, reflecting a high degree of change. However, and despite this intrinsically dynamic aspect, research on Web dynamics has not been incorporated into mainstream Web research. As an example, consider the well-known link-based ranking algorithms, such as PageRank (Page *et al.* 1999) or HITS (Kleinberg 1999), where the Web is modelled as a directed graph without any temporal information attached. These algorithms have been used as an important signal to determine the authority and relevance of documents on the Web or, more precisely, on snapshots of the Web.

We believe that the temporal dimension of the Web is a potentially rich source of data for information retrieval tasks. In this context, we measure and investigate the impact of time in link-based authority estimation algorithms on a well-defined subset of the Web: blogs. This study is based on a large collection of blogs and spans more than three years. We start by observing the dynamics of global ranks as older and newer information is removed. Next, we observe how citations to Web documents evolve over time. In this process we identify several cases where standard, time-independent, algorithms fail to identify emerging trends. To overcome these limitations, we study the distribution of citations over time as a possible signal for information retrieval tasks. We design two experiments to evaluate the use of temporal features in authority estimation algorithms. In the first experiment, we compare time-independent and time-sensitive ranking algorithms with a reference rank based on the total number of visits to each blog. In the second experiment, we use feedback from communication media domain experts to contrast different rankings of Portuguese news Websites. Both experiments confirm that time-dependent features obtained from link-analysis in blogs are able to capture valuable information that can be used to improve ranking tasks.

Related work

The use of time-dependent features combined with link-authority measures for information retrieval is still relatively rare. In this section we present and discuss previous studies in this area. In this paper, the authors present a model for representing hyperlinks over time based on the concept of TimeLinks — '[...] a triple (s,d, $[t_i:t_j]$), where s is the source URL, pointing to the destination URL d during the time interval $[t_i:t_j]$'. This model was tested by

conducting a few measurement experiments using data obtained from the Internet Archive. This constitutes one of the earliest attempts to model the dynamic character of the Web.

A very interesting example of the value of temporal information for Web information retrieval tasks is the work of Amitay *et al.* (2004). In this work, the last-modified field available in the HTTP header of Web documents is used to estimate the content's age. Based on this information, it is shown that real life events can be exposed because of what the authors call *fossilized* content. While exploring the notion of time-stamped Web resources, the authors introduce the concept of *timely authorities*, opposed to simple link-based authorities. This idea is illustrated with the adaptation of the HITS (Kleinberg 1999) and SALSA (Lempel and Moran 2001) algorithms by adjusting vertices' weights to include a time-dependent bias. To demonstrate the different results obtained with this approach, two authority rankings are computed for two queries: a basic authority ranking and a timely authority ranking. In the second ranking, results with a declining number of citations over time tend to be ranked in lower positions. On the contrary, resources with a large percentage of recent citations are ranked higher. Although the experiments conducted were limited to only two selected queries, the results suggested that the use of temporal information could positively influence authority rankings.

Baeza-Yates *et al.* (2004) conducted a detailed characterization of the Chilean Web in 2000 and 2001, focusing on the relationship between site quality and site age. The authors found a strong relationship between the macro-structure of the Chilean Web and age and quality characteristics. Results show that PageRank is biased against new pages. The authors discuss the need to incorporate age in document ranking functions, particularly in very dynamic environments such as the Web.

There are several proposals on how to incorporate temporal features in classic link-based formulae. Yu *et al.* (2004) adapt the PageRank algorithm by weighting each citation according to its date using an exponential decay function. The proposed algorithm, named TimedPageRank, was empirically evaluated using a bibliographic collection and a set of twenty-five selected queries. Results show that by including a simple time-dependent measure (i.e., citation age), retrieval performance is consistently improved. The authors suggest that the proposed methods can be conveniently adapted to Web search in general.

In a similar work, Berberich *et al.* (2004) and Berberich, Vazirgiannis *et al.* (2006) also adapt the original PageRank algorithm to favour certain nodes according to their freshness and the freshness of incoming and outgoing links. The impact of this approach, named T-Rank, was evaluated with user studies using the DBLP (a computer science bibliography Website) dataset and a selected number of Amazon's product pages. Another interesting approach was proposed by Berberich, Bedathur *et al.* (2006) to

complement time-independent ranking algorithms. By analysing time series of importance scores (e.g., PageRank scores), the proposed BuzzRank algorithm identifies trends using generic growth models and curve fitting techniques. The presented version of BuzzRank is very expensive both in terms of computing power and storage requirements since it needs to store the entire graph and perform PageRank computations for each time interval. The authors conducted a simple experiment, also using the DBLP database, to evaluate this approach. They found that BuzzRank tends to highlight topics that were popular during the specified periods.

More recently, Yang *et al.* (2007) proposed TemporalRank, a new algorithm that combines the current PageRank score with an historic score for each Web document. The historic score is obtained by combining previous PageRank scores computed using preceding snapshots of the Web graph. The proposal is evaluated using five large sub-graphs based on snapshots of the Web and user feedback collected from a browser toolbar plugin used by a commercial search engine. Results show that historic information directly improves the quality of the ranking. Moreover, the quality of the ranking improves as more snapshots are included. This is one of the few studies in this area that was tested over a collection of real Web documents and evaluated with user judgments. The main limitations of this study are the use of a collection covering only a total of six months, and the use of indirect user judgments, based on data collected through a browser toolbar.

In a recent work, Dai and Davison (2010) presented a new time-dependent, link-based algorithm to estimate the authority of Web pages. The algorithm, named T-Fresh, produces an authority score for each page and incorporates time by considering two aspects, freshness and the multiple Web snapshots at which a page exists. The freshness of a Web page includes both the freshness of incoming links and the freshness of the page itself. The authors adapt the original PageRank model by giving preference to fresher Web resources. This approach is experimentally evaluated using a corpus of archival Web pages in the .ie domain (Ireland). This corpus spans more than seven years and contains 158 million Web pages and approximately 12 billion temporal links. A collection of relevance judgments for a set of ninety queries is prepared using Amazon's Mechanical Turk. Based on these data, the authors report significant improvements over PageRank on both relevance and freshness.

In our work, we conduct several experiments and measure the impact of time in link-based rankings over a long time and a large collection of Web documents. Moreover, we study the impact of citation age in a standard link-based rank, making it more sensitive to recent events. We do not introduce new ranking algorithms; instead we concentrate on the measurement and characterization of the impact of time in Web link analysis. In this

context, we decided to use a simple time-aware scoring function to minimize the number of dependent parameters at play. The most distinctive aspect of our work is that it is based on genuine data. First, we have access to a collection of real Web documents. Contrary to most of the aforementioned contributions, which are focused on *structured collections* (e.g., DBLP), our emphasis is on Web documents. This introduces several challenges, most notably for the handling of a large-scale dataset and the parsing of real, unfiltered data. Second, we support the evaluation using a set of direct user judgments, based on the real number of visits to each blog. In addition, we conduct tests with communication media domain experts in a realistic scenario designed to evaluate the importance of a time-sensitive classification.

Materials and methods

As a result of a cooperation agreement established between the University of Porto and Portugal Telecom (PT), we have access to a large collection of Portuguese blogs. The criteria originally used to determine if a blog is Portuguese or not are unknown to us. One important aspect of this collection is the fact that it includes all the blogs registered at SAPO Blogs, a service operated by SAPO, a subsidiary of Portugal Telecom. Since this is a complete collection of a single and large blog service provider, independent of crawling policies or problems, it can be seen as a good sample of Portuguese blogs. The blogosphere (or blogspace) is very rich in temporal information given that, by default, all posts have an attached timestamp. This provides a context especially well suited for time-dependent analysis. The remainder of this paper is based exclusively on the blogs from this service. No other Web sources were crawled.

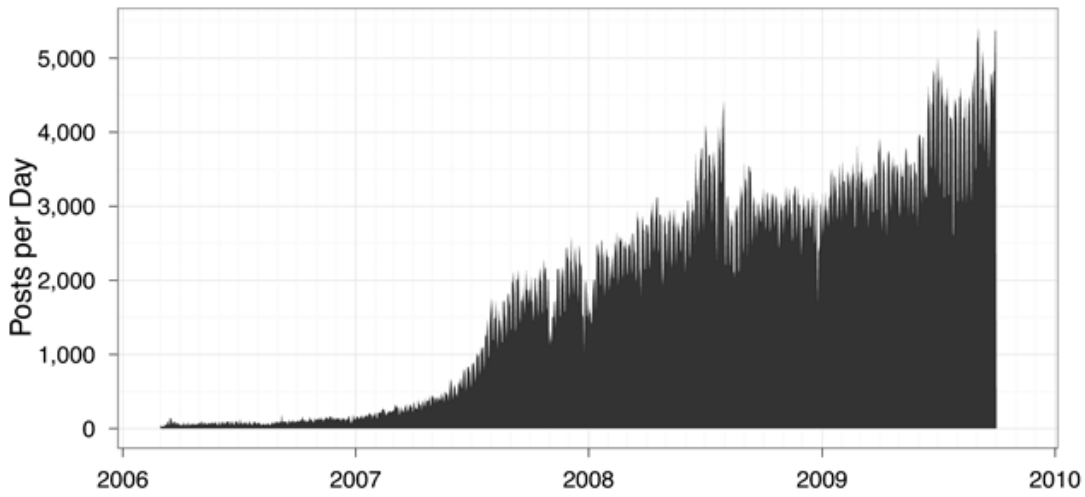

Figure 1: Posts published over time.

The current version of the SAPO Blogs service was launched in March 2006 and, as of September 2009, had more than 100,000 registered blogs, and approximately 2.4 million posts. Figure 1 depicts the evolution in the number of posts published per day

over time. To obtain an overall picture of the linking activity in this collection, we extract all HTML links from each post, together with its publication date. Figure 2 shows the total number of links found each month (labelled Original). This Figure reveals that an anomalous pattern in link activity occurred in mid-2008 and mid-2009. After manual observation of the atypical periods, we found that these peaks can be attributed to fake blogs, commonly used as *link farms* to artificially promote selected sites. To address this problem we implemented a simple filter based on three signals: the ratio between the total number of links in a blog and the total number of posts, the average number of posts per day, and the average number of links per day. We removed all blogs according to the following criteria: an average number of links per post higher than ten, an average number of posts per day higher than fifty, and an average number of links per day higher than 500. Additionally, we also removed all blogs with less than three posts and all blogs with an overall temporal span smaller than three days. We identified and removed 62,346 blogs (49.4% of the initial set), resulting in a drastic reduction of the original collection. The revised link activity over time is included in Figure 2 (labelled *Filtered*). It is clear from this figure that the unexpected peaks were most likely due to artificial links.

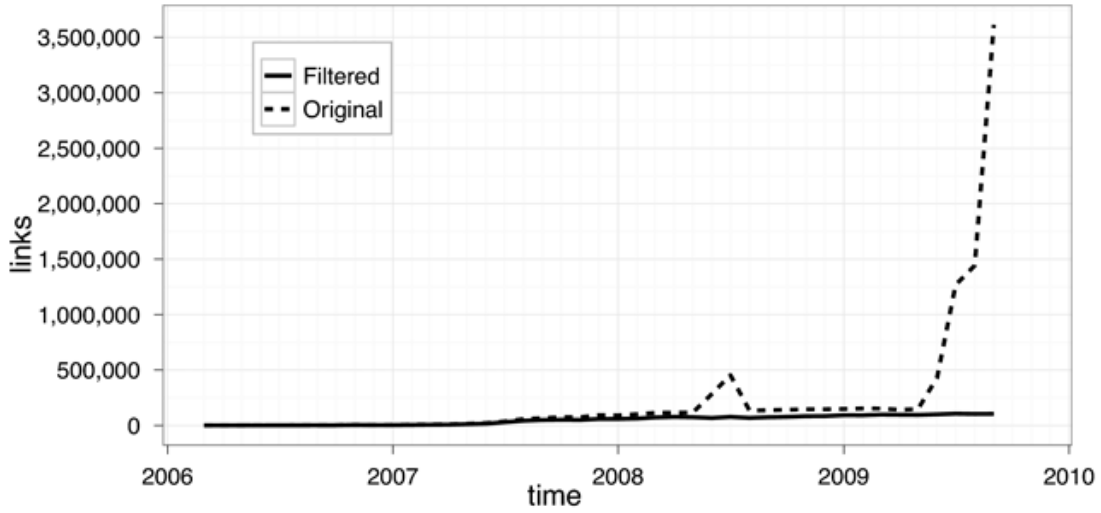

Figure 2: Links found over time.

Experiments and results

Link authority dynamics

In this section we present the results of our investigations on the dynamics of link-based algorithms over time. We use a simple measure based on the total number of incoming links (i.e., citations) to determine a Web document's link authority. This is commonly referred to as a document's *in-degree* when modelling the Web as a directed graph. Previous work has shown that basic in-degree counts outperform PageRank using several information retrieval measures, i.e., normalized discounted cumulative gain (nDCG), mean reciprocal rank (MRR) and mean average precision

(MAP) (Najork *et al.* 2007). Additionally, computing a document's in-degree is a simple procedure when compared to PageRank or other sophisticated measures. Also, we focus our study on host-based rankings rather than document-based rankings because, contrary to isolated Web documents, host-based rankings are more stable over time, and therefore more suitable to our goal. Moreover, host-based rankings are easier to evaluate by human assessors. In addition, the number of citations for each host is much higher, resulting in richer sets of temporal information. To obtain a host-based ranking, we simply coalesced individual Web documents by host.

A ranking based solely on the total number of citations is vulnerable to manufactured scenarios in which a small number of hosts originate the large majority of the citations. We found several obvious examples in our dataset. For example, one host had an in-degree of 1,056 but all citations were from only two different sites. To account for these situations we limited to one the number of citations from one host to another, over the time window, instead of using the raw number of citations. In other words, several citations from one host to another host (typically from different pages) are combined and only count as one, more specifically the first one. In the end, we obtain a weightless directed host-graph without multiplicity. It should be noted that this is an oversimplified approach to producing a link-based rank.

The global rank based on the entire collection of blog posts from March 2006 to September 2009 (forty-three months) is dominated by media hosting sites like YouTube, Flickr and Blogger's cache servers. The top positions of the rank also include reference sites such as the English and Portuguese versions of Wikipedia. The list is then filled with Portuguese news sites, reflecting the attention given by blogs to mainstream media.

To investigate the impact of the temporal dimension in link authority measures, we trimmed the initial collection and computed new ranks. First, we trimmed the data by removing one month from the beginning of the collection and only counting the links found within the remaining forty-two months. The collection was further trimmed one month at a time, until only one month of data was left (September 2009). This approach was intended to simulate the common situation where a search engine only has access to a limited period of historical data. We also trimmed the newest months from the collection using a similar method. This was intended to simulate the situation where a search engine does not have access to the freshest information (such as where there is a temporal lag between the published data and the crawled copy of these data). Finally, we also produced forty-three independent ranks each based on the data from a single month.

To evaluate the impact of removing data from the collection, we compared the percentage of common items between the global

rank (the baseline) and each one of the other ranks based on partial information. The adopted baseline represents an ideal scenario, where an algorithm has complete knowledge of the reality. We limited each rank to a maximum of 100 items. Figure 3 shows the rank intersection for each scenario. We also tested Kendall's *tau* distance (Kendall 1938) to measure the correlation between ranks and obtained very similar results. Given that the idea of *common items* is easier to visualize and understand, we opted for this measure. From Figure 3 we can see that, when trimming the newest months (end trimming), the percentage of common items increases from 20% (when only one month is considered, March 2006) to 100% (when the complete collection is considered). On the contrary, when the oldest months are trimmed (start trimming) the percentage of common items decreases from 100% (complete collection) to 55% (only September 2009 is considered). The monthly ranks, also depicted in the Figure, show the expected behaviour, with an interesting maximum of 75% in November 2008. In other words, if we only had access to one month worth of data, this month would produce the best approach with respect to a rank based on the entire collection (forty-three months).

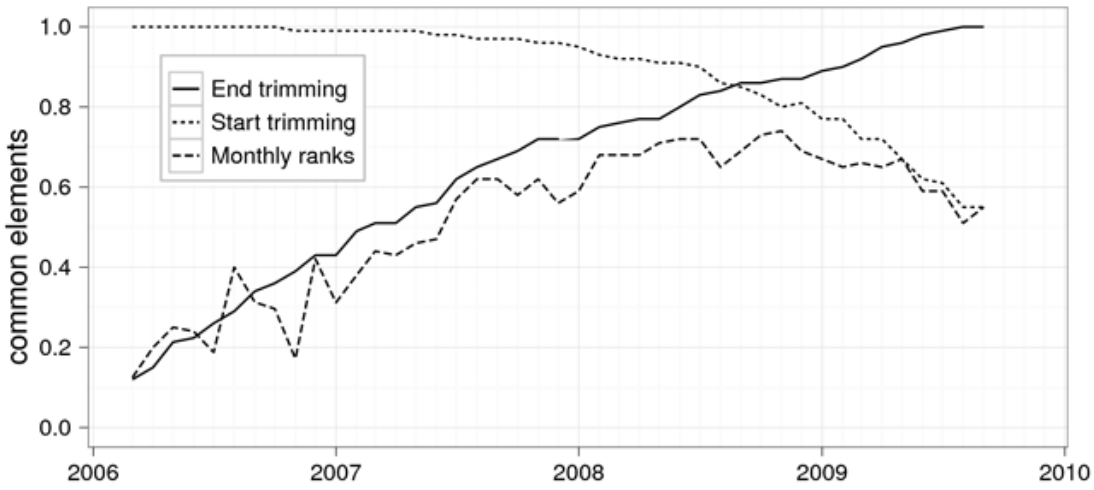

Figure 3: Common items with baseline in top 100 ranks for different trimmings.

Overall, the curves behave as expected: as more months are removed and less information becomes available, the number of common results decreases. At first glance one could think that historical information is less determinant given that removing older months from the complete dataset results in a slower decrease in the similarity when compared to removing newer months. As can be seen in Figure 3, the similarity between ranks has a steeper decay when new months are removed (top right area of the chart). However, if we recall that the SAPO Blogs service only gathered significant attention after mid-2007 (see Figure 1), we see that between this date and the final date of the collection, the similarity between ranks based on start and end trimming is

comparable. Until mid 2007, the correlation between the top 100 lists when older months are removed is very high and stable, a situation that can be attributable to this reduced overall activity.

To obtain a more fine-grained perspective on the internal dynamics of link-based ranks over time, we compared the rank obtained when using data from a single month with the rank obtained using data from the previous month. As depicted in Figure 4, the percentage of common items oscillates between 20% and 60%, with an apparent tendency to stabilize in more recent months. We add a smoother function to the figure to better illustrate this pattern. This shows that even when using data from a single month there is an important fraction of items that prevail.

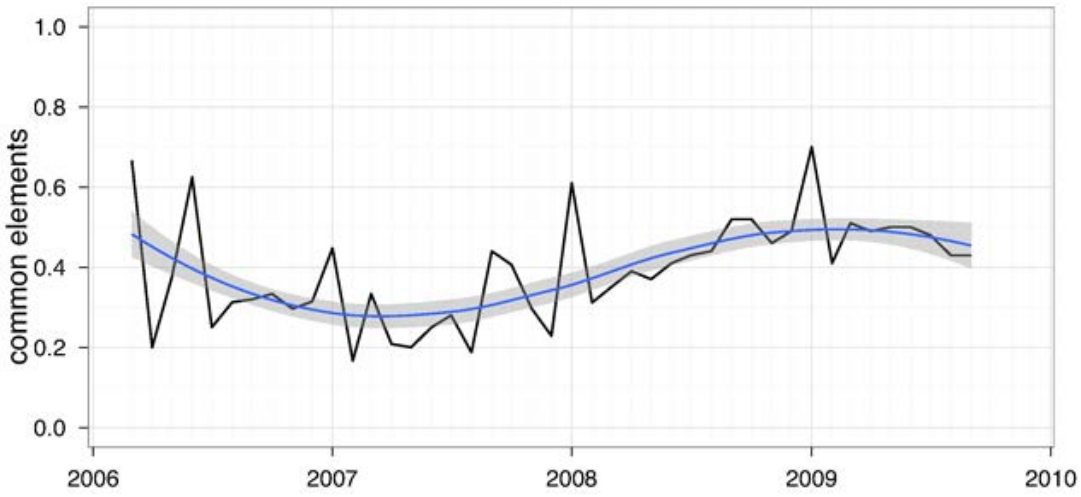

Figure 4: Common items in top 100 ranks for each month compared to the previous month.

Overall, this analysis shows that, for standard, cumulative, link-based ranking algorithms, access to historical information is important but not critical. We can see that even without six months of the latest data we can produce a rank that is very similar to the baseline rank obtained with complete knowledge of the collection (more than 90% common items). As mentioned before, similar plots were obtained when using Kendall's *tau*. This means that when using a time-independent algorithm, the freshness of the index is not decisive to obtaining accurate estimations of citation-based rankings. The presented data also seems to indicate that there is no evident advantage of historical data over contemporary data, nor the other way around.

Temporal profiles

As discussed in the previous section, time-independent ranks based on in-degree counts tend to be relatively stable over time. To better understand the evolution of citations for a given host, we extract *temporal profiles* using the dates of each individual citation. A temporal profile of a Web page or site corresponds to the distribution of citations to that page or site projected over time.

Figure 5 shows the monthly evolution in citations for two well-known social networking Websites: Hi5 and Facebook. The Hi5 service has historically been very popular in Portugal compared to other competitors. This is reflected in the host's temporal profile. Facebook was mostly unknown in Portugal until early 2009. As is very clear from Figure 5, Facebook had a very significant growth during 2009. However, when comparing the total number of citations, Hi5 (position 47) is placed ahead of Facebook (position 55). Given the fact that Facebook has consistently collected a higher number of citations in all months during 2009, this ordering can be seen as out-dated. A second example is presented in Figure 6. In this case, Twitter's temporal profile is presented side by side with two other sites: MySpace, a social networking Website, and Slide, a media-sharing service. This comparison illustrates that although Twitter has a significantly higher number of citations in the last months, it is still ranked below the other two sites when comparing the raw number of accumulated citations.

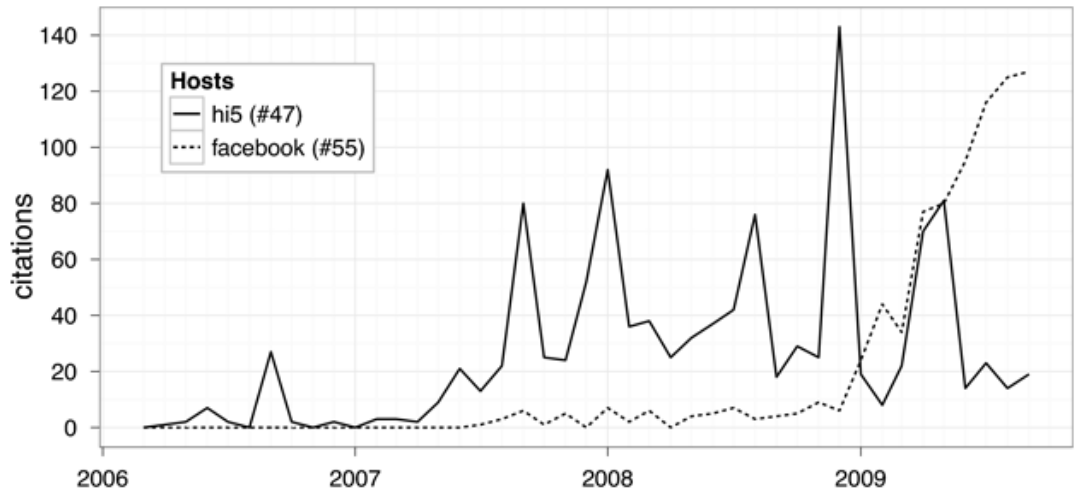

Figure 5: Hi5 and Facebook monthly citations over time.

Temporal profiles are a useful tool to unveil part of the history behind simple in-degree counts. They can reveal growing or fading citation activity over time. Thus, the temporal profile of a Web resource can be regarded as an original signal that captures time-sensitive information about these resources. It is worth noting that for the two examples presented here, the rise in citations occurs in the most recent months of the collection. This indicates that although access to the most recent information may not be determinant for overall ranks (as discussed in the previous section), it can be critical for particular sites such as these. To conclude: it is important to keep in mind that these charts suffer from a bias introduced when reducing the multiplicity of citations between two hosts. Since we are considering only the first citation when multiple citations occur between the same hosts, it is likely that these figures are under-evaluating recent data.

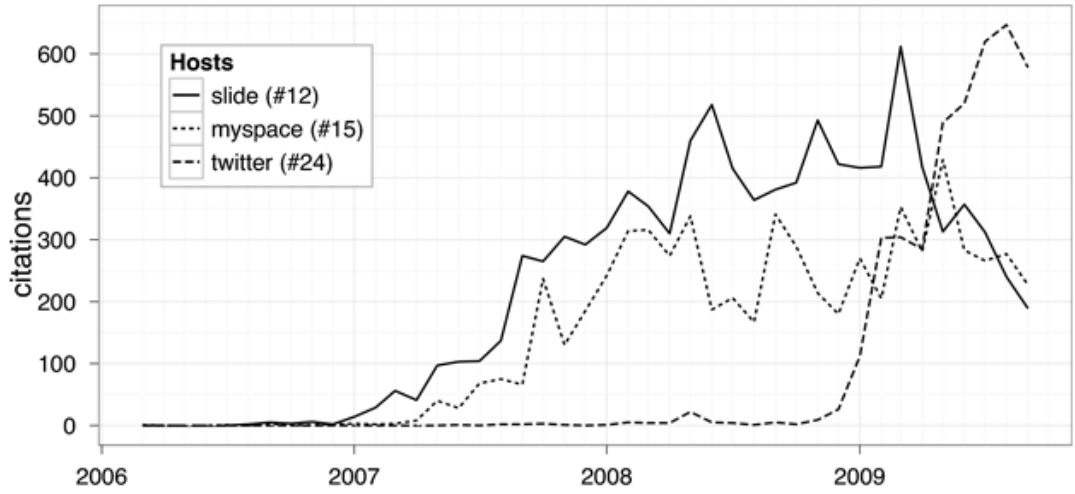

Figure 6: Twitter monthly citations over time.

A time-sensitive link authority

We explore an approach that takes into account temporal information together with citation counts to produce a time-sensitive ranking. In a nutshell, instead of considering that all citations have the same weight, we attribute a lower weight to older citations. In the previously presented time-independent approach, each citation counted as one vote with weight 1. Here, we use the formula defined in Equation 1 to compute the value of each citation as a function of its age (in months). A citation's weight decays as its age increases. We recall that, as defined in the Link authority dynamics section above, we limited to one the number of citations from one host to another.

$$\frac{1}{(age+1)^p} \quad , \quad p \geq 0$$

Equation 1: Formula used to compute the value of each citation as a function of its age.

The parameter *p* can be adjusted to define the rate of decay. For instance a higher *p* value means that older citations loose value more quickly. The age of a citation was measured as the number of months to September 2009. For example, if *p*=1, then each citation made during September 2009 would have a weight of 1, and each citation made during August 2009 would have a weight of 1/(1+1) = 0.5 . A citation made one year earlier (September 2008) would have a weight of 1/(12+1) = 0.077. This approach is similar to the previously discussed work of Yu et al. (2004). The previously defined original rank is obtained with *p*=0, i.e., all weights are equal independent of age.

We produced two ranks considering the complete dataset using *p*=0.5 (named *decaysoft*), and *p*=1.5 (*decayhard*). A comparison between these ranks and the previous time-independent rank (named *original*) is depicted in Figure 7. This figure presents the percentage of common items (*y-axis*) between each pair

considering a different number of top items (*x-axis*). There are a high number of common items between each rank's top items since the most cited sites tend to be consistently very cited over time. This means that top ten (or lower) ranks tend to be very similar, even when older citations are discarded. This figure also shows that by decreasing the value of older citations we obtain different ranks. Moreover, we can see that this difference increases, as expected, in proportion to the weight given to older citations. With p=0.5 the percentage of common items converges to 87%, while with p=1.5 this percentage decreases towards 60%.

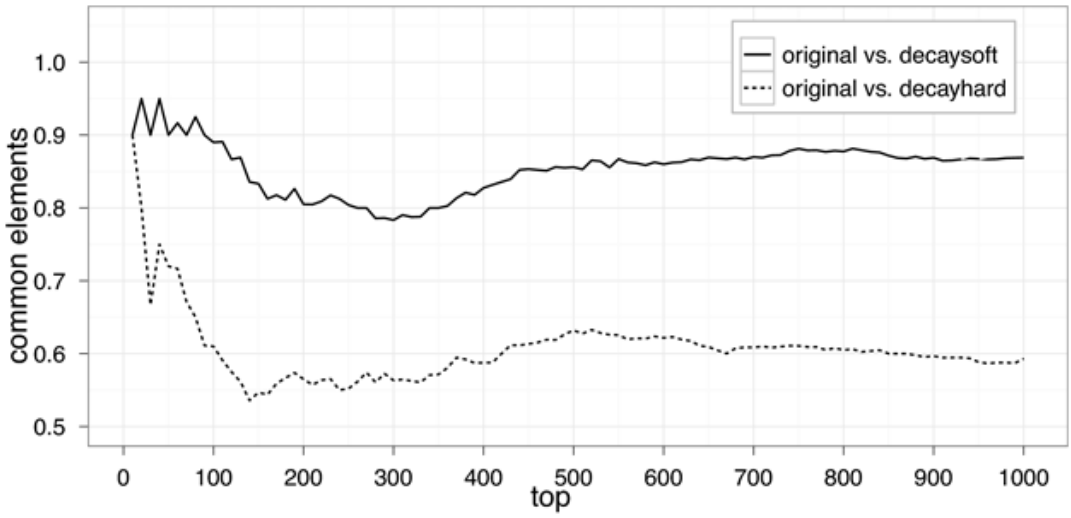

Figure 7: Top 1,000 rank intersections between original ranking and time-sensitive alternatives.

To better understand the impact of modifications to the parameter p, we compared the original rank with different time-sensitive ranks, at different p values. The results of this experiment are depicted in Figure 8, with p varying between 0.1 and 4. While in the previous analysis we compared different top ranks keeping fixed p values (i.e., *decayhard* and *decaysoft*), here we compare three fixed top ranks (i.e., 50, 500 and 1000) and change p in the *x-axis*. The percentage of common items between the original rank and the time-biased ranks reaches a plateau near p=3. This happens because the range of possible weights obtained with Equation 1 decreases as p increases. In other words, the weight attributed to posts with different ages becomes 0 for higher p values, resulting in almost identical ranks.

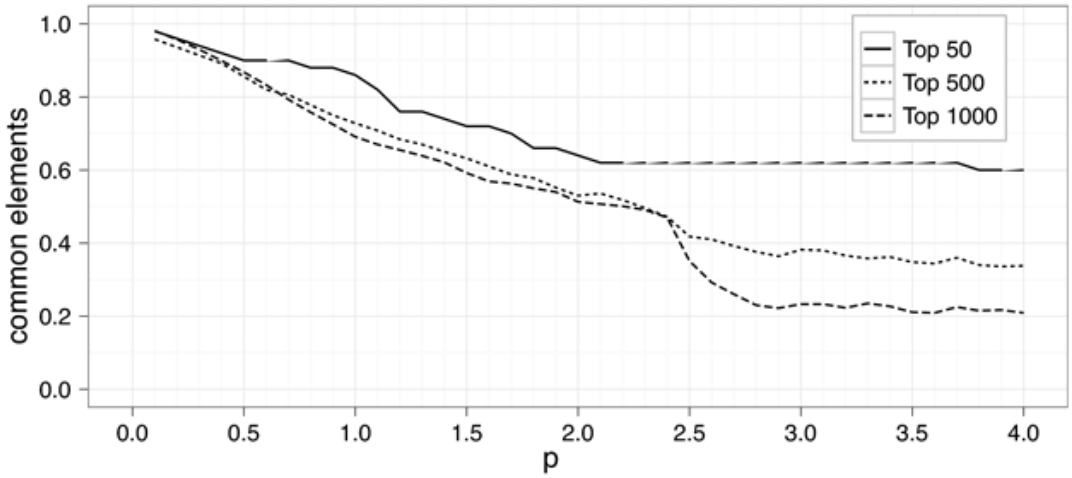

Figure 8: Common items between original ranking and a time-sensitive alternative with different values for the *p* parameter, for a distinct number of top items.

The above experiment shows that by altering the weight of links as a function of time, we can produce different rankings. We now evaluate the quality of these ranks by comparing them to the previously defined time-independent ordering. In the following, except where otherwise noted, we compare the baseline rank with the time-dependent rank where p=0.5. First, we revisit the cases discussed in the Temporal profiles section in the light of the time-sensitive rank. Given the higher value given to recent citations in this new approach, Facebook (no. 39) is ranked ahead of Hi5 (no. 54) as expected, and Twitter (no. 10) clearly jumps to the top, in front of Slide (no. 11) and MySpace (no. 12). This result is a better reflection of the current Portuguese Web. In a nutshell, sites that exhibit a quick and unexpected increase in recent popularity are discernible when using a time-sensitive approach. Another point worth noting is the fact that the temporal ranking is able to quickly detect changes in a site's address. One of the major television networks in Portugal (SIC) has changed its URL twice in recent years, from sic.sapo.pt to sic.aeiou.pt (in use from April 2008 to August 2009) and then back to its original sic.sapo.pt. This evolution is explicit in the temporal profile shown in Figure 9. While the baseline rank still puts sic.aeiou.pt (no. 31) ahead of sic.sapo.pt (no. 41), the time-sensitive rank with p=1.5 inverts this order (no. 23 *vs.* no. 53). We have found several other cases where the same happens.

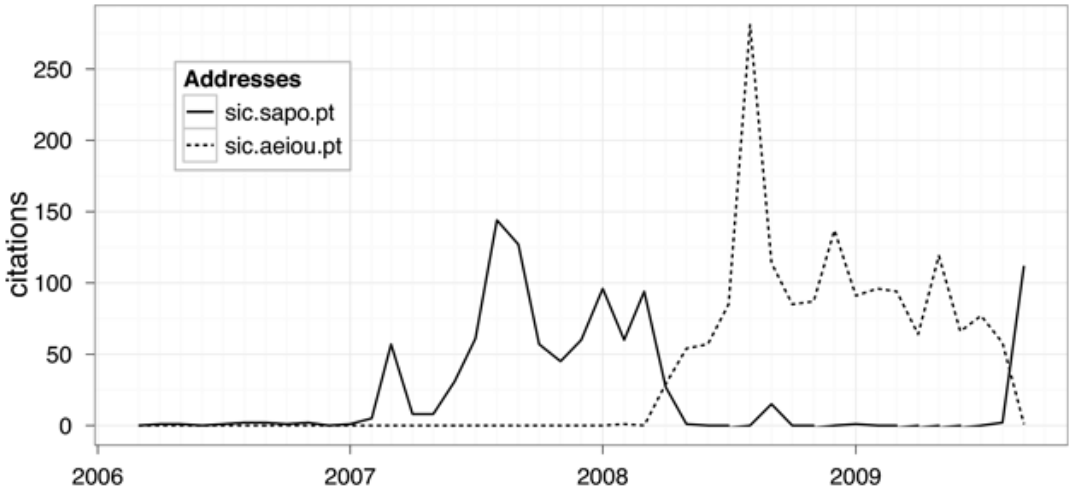

Figure 9: sic domains and citations over time.

Ranking blogs

In the context of the previously mentioned protocol between the University of Porto and Portugal Telecom, we were also given access to a sample of SAPO Blogs' Web access logs. This is an extremely relevant and unique resource that can be used to estimate a global abstract ranking for SAPO's blogs. These log files span over twenty-seven days between late March and mid April 2010. Consequently, there is a lag of approximately six months between the collection of blog posts studied and the corresponding Web access logs. To assess the quality of this dataset, we first estimate the consistency of these Web access logs by extracting ranks for each of the days found in the log files. In this process we only include HTTP GET requests, POST requests are excluded because they represent post or comment submissions, not directly related to page visits. In addition, we also remove all requests made by the top Web crawlers (namely Google, Yahoo and Bing). These crawlers account for a significant number of the total hits (approx. 10% in our case), introducing a bias that we try to minimize. We observe that the total number of hits per day for all blogs exhibits a stable value across all days of our sample (in the order of millions). Next, for each day of access logs, we extract a list of blogs ordered by the total number of hits in that day filtering crawlers' accesses. We produce a global rank that combines all twenty-seven days, aggregating all hits per blog. The intercorrelation between the rank in each day and this global rank shows high stability over time: for the top 10, 100 and 1,000 ranks, the percentage of common elements is stable and close to 85%.

To evaluate the impact of time in authority estimation algorithms, we compare the baseline, *decaysoft* and *decayhard* ranks defined in the section A time-sensitive link authority with this global rank based on the total number of visitors per blog. First, to obtain comparable lists, we filter our ranks to include only SAPO's blogs. The rank of blogs ordered by the total number of visitors has more

than 85,000 elements, while the three alternative ranks we have created have less than 20,000 elements. Table 1 summarizes this information and also includes the overlap between the reference rank (based on the number of visitors) and each one of the other ranks. In this table we see that the percentage of common elements between the reference rank and the ranks to be evaluated is considerably low. Taking this into account, we adopt the approach proposed by Bar-Ilan (2005), i.e., we intersect and re-rank the two lists being compared and use Spearman's *rho* as a measure of rank correlation. Spearman's *rho* ranges between -1 and 1, where 0 stands for no correlation, -1 for complete disagreement and 1 for complete agreement between the two ranks. The coefficients of correlation, for different number of top items, between the reference rank and the time-independent and time-aware alternatives are outlined in Table 2. The highest correlation value in each row is highlighted in bold. Also included in this table is the significance value of each correlation test (p-value). When comparing ranks of size equal or greater than 500, all results are statistically significant at the 0.01 level. Examining this table we see a consistently higher correlation value in the *decaysoft* rank, except when comparing the top 10 elements (not statistically significant). Moreover, for most of the top values observed, the *decayhard* rank also exhibits a higher correlation compared to the *original* rank. These results show that, when compared to a time-independent alternative, a rank including temporal information has a higher correlation with an independent reference rank based on the number of visitors.

Rank	Size	Overlap (%)			
		@100	@500	@1000	full set
Visits	87,914	*100*	*100*	*100*	*100*
Original	17,132	24	23	25	38
Decaysoft	14,874	22	21	25	34
Decayhard	14,874	15	16	20	34

Table 1: Size and overlap of SAPO's blogs ranks.

Top	Original		Decaysoft		Decayhard	
	rho	p	rho	p	rho	p
10	**0.22**	0.54	-0.10	0.78	0.09	0.81
20	0.20	0.38	**0.28**	0.24	0.22	0.35
100	0.12	0.25	**0.15**	0.12	0.14	0.16
500	0.16	0.00	**0.22**	0.00	0.18	0.00
1,000	0.17	0.00	**0.22**	0.00	0.21	0.00
5,000	0.25	0.00	**0.30**	0.00	0.27	0.00
10,000	0.32	0.00	**0.38**	0.00	0.30	0.00

Table 2: Correlations, for different number of top items, between the reference visits-based rank and the time-independent plus time-sensitive ranks.

Ranking Portuguese news Websites

Obtaining relevance judgments for abstract top rankings, such as the ones we are producing, is very challenging. Without a context (e.g., an information need) we cannot select users, nor ask them to

simply rank the theoretical importance of sites. Thus, to address this problem and reduce the uncertainty of an evaluation, we design an experiment that narrows down the context to a particular topic, specifically *Portuguese news Websites*. First, for each one of the three ranks being evaluated, we handpick all the news sites and produce new ranks keeping the information about the original positions. Then, for each set of two ranks being compared, we identify all pairs of sites where the ordering in each rank was reversed. Next, for each pair, we calculate the distance between the positions in each rank. Finally, we add the two distances (one for each rank) and order these values, obtaining a list representing the pairs where the difference between the ranks is more noticeable. To illustrate this procedure, consider the following case: in the time-independent rank site A is ranked in position 10 and site B is ranked in position 22. In the time-dependent rank the ordering is reversed and site A is ranked in position 8 and B is in position 6. The final value for this pair is 14, obtained by adding the differences found in each rank (12+2).

After this initial procedure, we select nine pairs of popular news sites from the top of this ordered list and conduct two experiments to assess the relative value of each ranking. Given that most media outlets manage various Web hosts (e.g., news.example.com, www.example.com), we decided to use the principal URL for each site (typically the www). In a nutshell, we have singled out pairs of popular news sites where the ordering found in the two ranks exhibits a more striking difference. It is worth restating that we try to reduce the ambiguity of the evaluation by focusing on the most extreme cases.

We conduct a first assessment of this data using the information published in Netscope's monthly rank. Netscope is an opt-in service that measures Websites' audiences. This is the most popular service in Portugal, frequently cited when discussing the typical profile of the Portuguese Internet user. All sites included in our pairs could be found in Netscope's top 100 rankings. These rankings are published monthly and, to match the end of our dataset, we base this evaluation on the September 2009 rank. The Netscope ranking supports the baseline rank in seven out of nine cases, while it endorses the temporal-sensitive ranking in two of the pairs. While this experiment clearly attests the importance of a global, time-independent rank, it also suggests that the temporal profile of citations contains valuable information.

To complement this first experiment we organized an additional evaluation using human experts. We contacted eleven experts in the area of communication media, mostly teachers and editors in reference publications, and asked them to express their views about the set of nine pairs. More specifically, for each pair, we asked them to indicate which of the sites they viewed as more popular for the Portuguese bloggers in general. We asked them to skip the pair if they had doubts about the relative popularity of the

sites. From the answers collected, the experts agreed with the baseline rank in five cases and with the temporal rank in the four remaining cases. This experiment reinforces the idea that a rank biased towards newer citations contains valuable information. Overall, our experiments show that, while historical data is indispensable to determine the value of resources, contemporary data contains important information that reflects current trends familiar to users.

Discussion and conclusions

Research in the area of Web information retrieval has been mostly focused on static, snapshot-like representations of the Web. Nevertheless, it is well known that the World Wide Web is a communication medium rich in temporal information. Research has shown that Web documents change both frequently and substantially over time. Despite that, previous work in these areas has devoted little attention to study the importance of time-dependent features for relevance estimation. This might be partially explained by the difficulties in obtaining good Web collections containing historical data for research.

This work is based on a large sample of blogs from a single service provider, spanning a period of forty-three months. Although the collection is limited to a single provider, it corresponds to a complete copy of the data from a large service without any sampling. This results in an organic collection without the biases typically introduced during the crawling phase. Using this data we study the temporal properties of a ranking obtained with a simple algorithm based on the total number of citations. The evidence collected suggests there is no significant difference between the value of historical information and contemporary information in time-independent rankings, i.e., the impact of removing either older data or newer data from the collection is alike. Using several informal examples we also observe that the standard time-independent ranking is unable to capture the correct popularity of sites with very high citation activity in recent periods. This is a problem likely to occur more frequently as the size of a collection grows and more value is accumulated in past citations, i.e., ranks will tend to crystallize and be less vulnerable to subsequent changes. We compare these results with an alternative link-based ranking algorithm where newer links are given a higher weight. In other words, a citation's value exponentially decays, as it gets older. As expected, we observe that time-sensitive and time-independent ranks become more divergent as the value attributed to older links decays more rapidly.

To evaluate and compare both time-independent and time-dependent approaches in ranking we designed several experiments. From these, we can draw the following conclusions. First, based on various handpicked examples, we see that an algorithm that favours recent citations more quickly discards abandoned Websites and more rapidly identifies popular present

trends. Also, we find that a time-sensitive rank can be used as a good indicator of the correct address of Websites that had multiple URLs over time. Next, we designed a detailed experiment to assess the value of the temporal rank versus the classic temporal-independent rank. We establish a reference rank based on the number of visits to each individual blog and measure the correlation between this rank and the alternatives being evaluated. The time-biased alternatives exhibit the highest correlation with the reference rank. This experiment clearly confirms the advantages of combining temporal information in a link-based authority estimation algorithm.

As a final experiment, we asked a group of experts to give feedback on a selected number of pairwise comparisons. While the standard baseline rank was preferred in five out of nine cases, in the remaining four cases the time-dependent rank was favoured. This suggests that valuable information can be found in both ranks. While the time-independent rank is still a vital source of quality data, the time-sensitive rank captures important information that is otherwise invisible.

It is clear that access to fresh data is essential, but not sufficient by itself, to identify current trends and algorithms are of major importance. Temporal profiles are a valuable tool and offer a richer picture of a Web resource's authority when compared to raw citation counts. Although both time-independent and time-aware approaches are based on the same raw data, the experiments conducted indicate that they can be treated as complementary signals for relevance assessment by information retrieval systems. We show that temporal information present in blogs can be used to derive stable time-dependent features, which can be successfully used in the context of Web document ranking. We conjecture that the use of time-dependent features will be conditioned by the retrieval task being addressed. In summary, the major contributions of this paper include a detailed analysis and characterization of a realist, large-scale collection of Web documents, specifically blogs, from a time-aware point of view, and the evaluation of the impact of a time-dependent scoring function in a standard citation-based rank.

Acknowledgements

This work would not be possible without the collaboration of SAPO, a subsidiary of Portugal Telecom. We are especially grateful to Benjamin Júnior for his attention and dedication to this project. We would also like to thank the reviewers whose suggestions helped to improve and clarify this manuscript. Sérgio Nunes was financially supported by Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE - III Quadro Comunitário de Apoio), under grant SFRH/BD/31043/2006.

About the authors

Sérgio Nunes is an Assistant Professor at the Department of Informatics Engineering, Faculty of Engineering, University of Porto and a researcher at INESC Porto. He holds a Ph.D. in Informatics Engineering from the University of Porto. His main research interests are information access and retrieval, Web information systems and databases. He can be contacted at: sergio.nunes@fe.up.pt

Cristina Ribeiro holds a PhD in Informatics from Universidade Nova de Lisboa. She is an Assistant Professor at the Department of Informatics Engineering, Faculty of Engineering, University of Porto, a researcher at INESC Porto, and teaches undergraduate and graduate courses in information retrieval, markup languages and knowledge representation. Her research interests include information retrieval, multimedia databases and digital repositories. She can be contacted at: mcr@fe.up.pt

Gabriel David holds a Ph.D. in Informatics, Artificial Intelligence branch, at Universidade Nova de Lisboa, 1994. He is currently an Associate Professor at the Department of Informatics Engineering, Faculty of Engineering, University of Porto and a researcher at INESC Porto. His main research interests are in Information Management, Databases, and Digital Preservation. He has been the leader of projects (funded by Portuguese FCT) MetaMedia on multimedia archives and DBPreserve on preservation of databases. He can be contacted at: gtd@fe.up.pt

References

Adar, E., Teevan, J., Dumais, S. T. & Elsas, J. L. (2009). The Web changes everything: understanding the dynamics of Web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (ACM WSDM'09)*, (pp. 282–291) New York, NY: ACM Press. Retrieved 31 July, 2013 from http://www.cond.org/wsdm09-change-camready.pdf (Archived by WebCite® at http://www.webcitation.org/6IWMNI0Wx)

Amitay, E., Carmel, D., Herscovici, M., Lempel, R. & Soffer, A. (2004). Trend detection through temporal link analysis. *Journal of the American Society for Information Science and Technology*, **55**(14), 1270–1281. Retrieved 31 July, 2013 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.3573&rep=rep1&type=pdf (Archived by WebCite® at http://www.webcitation.org/6IWkckRIw)

Baeza-Yates, R., Castillo, C. & Saint-Jean, F. (2004). Web dynamics, structure and page quality. In M. Levene & A. Poulovassilis, (Eds.). *Web Dynamics: adapting to change in content, size, topology and use*, (pp. 93–109). Berlin: Springer Verlag. Retrieved 31 July, 2013 from http://chato.cl/papers/baeza04_web_dynamics_structure_page_quality.pdf (Archived by WebCite® at http://www.webcitation.org/6IWMXe5MD)

Bar-Ilan, J. (2005). Comparing rankings of search results on the Web. *Information Processing & Management*, **41**(6), 1511–1519. Retrieved 31 July, 2013 from http://www.jasonmorrison.net/iakm/cited/Bar-Ilan_Judit_comparing_rankings.pdf (Archived by WebCite®

at http://www.webcitation.org/6IWkFUvJT)

Berberich, K., Bedathur, S., Vazirgiannis, M. & Weikum, G. (2006). BuzzRank... and the trend is your friend. In *Proceedings of the 15th international conference on World Wide Web (WWW'06), University of Southampton, United Kingdom*, New York, NY: ACM Press. Retrieved 31 July, 2013 from http://bit.ly/1cntVsd (Archived by WebCite® at http://www.webcitation.org/6IWLqc9Ew)

Berberich, K., Vazirgiannis, M. & Weikum, G. (2004). T-Rank: time-aware authority ranking. In *Algorithms and Models for the Web-graph: Third International Workshop, WAW 2004*, (pp. 131–142). Berlin: Springer. (Lecture Notes in Computer Science Vol. 3243). Retrieved 31 July, 2013 from http://202.38.64.11/~jpq/paper/stier/2004-WAW-T-Rank-%20Time-Aware%20Authority%20Ranking.pdf (Archived by WebCite® at http://www.webcitation.org/6IWMgiEIp)

Berberich, K., Vazirgiannis, M. & Weikum, G. (2006). Time-aware authority ranking. *Internet Mathematics*, **2**(3), 301–332. Retrieved 31 July, 2013 from https://domino.mpi-inf.mpg.de/intranet/ag5/ag5publ.nsf/0/764352253d013b71c125718c003847a4/$FILE/Berberich.pdf (Archived by WebCite® at http://www.webcitation.org/6IWk4SkiP)

Dai, N. & Davison, B. D. (2010). Freshness matters: in flowers, food, and Web authority. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR'10),* (pp. 114–121). New York, NY: ACM Press. Retrieved 31 July, 2013 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.180.6421&rep=rep1&type=pdf (Archived by WebCite® at http://www.webcitation.org/6IWeslgj0)

Fetterly, D., Manasse, M., Najork, M. & Wiener, J. L. (2004). A large-scale study of the evolution of Web pages. *Software: Practice and Experience*, **34**(2). 213–237. Retrieved 31 July, 2013 from http://research.microsoft.com/en-us/projects/pageturner/pageturner-spe2004.pdf (Archived by WebCite® at http://www.webcitation.org/6IWjw8Ho9)

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**(1/2), 81–93.

Kleinberg, J. M. K. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**(5). 604–632. Retrieved 31 July, 2013 from http://www.cs.cornell.edu/home/kleinber/auth.pdf (Archived by WebCite® at http://www.webcitation.org/6IWjiPgHo)

Kraft, R., Hastor, E. & Stata, R. (2003). Timelinks: exploring the evolving link structure of the Web. In, *Proceedings of the Second Workshop on Algorithms and Models for the Web-Graph (WAW 2003)*. Retrieved 3 June, 2013 from http://cis.poly.edu/~qq_gan/papers/timelink.pdf (Archived by WebCite® at http://www.webcitation.org/6Gyi6B2dS).

Lempel, R. & Moran, S. (2001). SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, **19**(2), 131–160. Retrieved 31 July, 2013 from http://delab.csd.auth.gr/~manolopo/oikonomiko/salsa.pdf (Archived by WebCite® at http://www.webcitation.org/6IWfrl94g)

Najork, M. A., Zaragoza, H. & Taylor, M. J. (2007). HITS on the Web: how does it compare? In *Proceedings of the 30th annual*

international ACM SIGIR conference on Research and development in information retrieval (ACM SIGIR'07), (pp. 471–478). New York, NY: ACM Press. Retrieved 31 July, 2013 from http://research.microsoft.com/pubs/65139/sigir2007.pdf (Archived by WebCite® at http://www.webcitation.org/6IWeycHsd)

Ntoulas, A., Cho, J. & Olston, C. (2004). What's new on the Web? The evolution of the Web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web (WWW'04)*, (pp. 1–12). New York, NY: ACM Press. Retrieved 31 July, 2013 from http://cs.brown.edu/courses/csci2531/papers/www04-ntoulas.pdf (Archived by WebCite® at http://www.webcitation.org/6IWf640d6)

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank citation ranking: bringing order to the Web (Technical report). Stanford InfoLab. Retrieved 3 June, 2013 from http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

Yang, L., Qi, L., Zhao, Y. P., Gao, B. & Liu, T. Y. (2007). Link analysis using time series of Web graphs. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (ACM CIKM'07)*, (pp. 1011–1014). New York, NY: ACM Press. Retrieved 31 July, 2013 from http://research.microsoft.com/pubs/131496/CIKM2007.pdf (Archived by WebCite® at http://www.webcitation.org/6IWfFGnfU)

Yu, P. S., Li, X. & Liu, B. (2004). On the temporal dimension of search. In *Proceedings of the 13th international World Wide Web conference (WWW 2004): Alternate track papers & posters (WWW Alt. '04)*, (pp. 448–449). New York, NY: ACM Press. Retrieved 31 July, 2013 from http://www.cs.uic.edu/~xli3/www04.pdf (Archived by WebCite® at http://www.webcitation.org/6IWfMuyaY)

How to cite this paper

Nunes, S., Ribeiro, C. & David, G. (2013). The impact of time in link-based Web ranking. *Information Research*, **18**(3) paper 586. [Available at http://InformationR.net/ir/18-3/paper586.html]

Find other papers on this subject

Scholar Search | Google Search | Bing

Check for citations, using Google Scholar

Tweet

Share