A Time-aware Random Walk Model for Finding Important Documents in Web Archives

Tu Ngoc Nguyen, Nattiya Kanhabua, Claudia Niederée and Xiaofei Zhu

L3S Research Center / Leibniz Universität Hannover, Germany
{tunguyen, kanhabua, niederee, zhu}@L3S.de

ABSTRACT

Due to their first-hand, diverse and evolution-aware reflection of nearly all areas of life, web archives are emerging as gold-mines for content analytics of many sorts. However, supporting search, which goes beyond navigational search via URLs, is a very challenging task in these unique structures with huge, redundant and noisy temporal content. In this paper, we address the search needs of expert users such as journalists, economists or historians for discovering a *topic in time*: Given a query, the top-k returned results should give the best representative documents that cover most interesting time-periods for the topic. For this purpose, we propose a novel random walk-based model that integrates relevance, temporal authority, diversity and time in a unified framework. Preliminary experimental results on a large-scale, real-world web archival collection shows that our method significantly improves the state-of-the-art algorithms (i.e., PageRank) in ranking temporal web pages.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms

Algorithms, Experimentation, Performance

Keywords

Temporal Ranking, Web Archive, Authority, Diversity

1. INTRODUCTION AND BACKGROUND

Web archives reflect nearly all types of social cultural, societal and everyday processes of our lives in the web as well as the exponential growth and continuous change in content and structure of the world wide web. Therefore, web archives from organizations such as the Internet Archive have the potential of becoming invaluable gold-mines for temporal content analytics of many kinds (e.g., politics and social issues, economics or media). First hand evidences about such processes are of great benefit for expert users such as journalists, economists, or historians. However, support for navigational search as it is, for example, offered by the Wayback machine[1], is not sufficient for tapping the full potential of web archives. Instead, search results should provide a good coverage of the query topic over time for enabling exploration of the topic and its evolution. Therefore, content relevance is not the only driver: time relevance and impact are other key factors. Further aspects, which make web archive search very different from web search are the high redundancy (pages of near-identical content are crawled all over again) and the special role that time/crawling time is playing in the web archive structure.

In this paper, we tackle the problem of discovering important documents along the time-span of the web archives by a ranking approach. The intuition is that the impact/authority of a document in the web archives with regards to a query is strongly influenced by time. Hence, the temporal authority of a document should be accumulated over a surrounding time window (instead of considering only one or all temporal snapshots).

Temporal link analysis for web search improvement has been studied in previous work [1, 4, 9, 10]. Their common goal is to improve state-of-the-art link-based algorithms (i.e., PageRank [8]) in favoring new web pages, instead of old, stale pages (that often ranked high due to its accumulated in-links). The most advanced approach in this direction is described in [4], where they track the authority of a page over multiple historical (past) web snapshots. This allows the incorporation of web freshness into authority propagation, and hence, boosts the authority of fresh (new) page at the querying time. In estimating the temporal authority along the time dimension, we adapt their strategy by propagating the authority of *past* and *future* snapshots. This backward propagation (from future snapshots) accounts for the 'lagging' time a new document needs to gather in-links. For example, a document about health care reform issued in March, 2010 is more relevant with the time point than April, 2010, where it has more number of in-links.

Graph-based diversity for ranking based on random walk are addressed in [2, 5, 11]. The first two utilized a greedy algorithm in transforming a picked node into an absorbing state. This punishes neighboring nodes but the random walk still lingers at away nodes, hence increases diversity. Mei et al. [5] introduce a vertex- reinforcement base on the intuition that nodes are visited many times tend to be more likely to be re-visited. This reinforcement on the transition

[1] http://archive.org/web/

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766462.2767832.

probability has a strong theoretical foundation, brings less complexity and can be improved towards scalability in large graphs.

For temporal ranking in the web archive, we are the first ones to combine the problems of relevance, temporal authority, diversity and time in a unified framework. In more detail, we construct a temporal graph over the web archive. Using time preference and relevance as priors, in Section 2.1, we propose a novel random walk model on the temporal graph. The model accounts for the in-link and natural time lagging in the web archive in mining the temporal authority. We present two ways of injecting time preference in Section 2.2. Further more, we introduce a novel diversity mechanism that penalizes both neighbors in the same web snapshots and *across* snapshots in Section 2.3. Our experiments are conducted on a large-scale, real-world web archival dataset, which is further explained in Section 3.

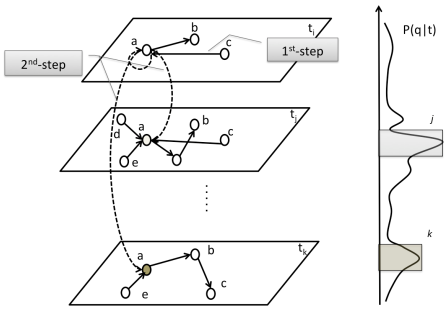

Figure 1: Time-travel web archive surfer. Solid lines indicate *within*-snapshot transitions, and dashed lines indicate the *across*-snapshot teleporting

2. METHODOLOGY

2.1 Temporal Ranking Model

In this work, we re-design the traditional PageRank (at document/page level) to more precisely measure the temporal authority of the documents. We propose a new time-aware random surfer model, with the intuition that instead of jumping to a random node with equal probability, this traveler favors jumping to a node at a time period of interest.

Temporal Graph Model A temporal graph $\mathcal{G}$ consists of multiple graphs at different time points, called graph snapshots (snapshots for short). A snapshot G is a directed graph with time annotation, $G_t = (V_t, E_t, t)$, with $t \in T$, $V_t \in V$ and $E_t \in E$. A vertex $v \in V$ can belong to multiple snapshots $\{V_t\}$. A vertex $v \in V_{t_i}$ is connected with $v \in V_{t_j}$ by an inter-link $\{v_{t_i}, v_{t_j}\} \in \mathcal{I}$. In the web archive context, each vertex v in a graph snapshot is a revision of a document d, identified by an unique URL. The vertex is time-stamped by the crawling time. The edge between two vertices is the hyper-link between two revisions. It is time-stamped as the time of the source vertex.

Temporal Random Surfer Model We describe a 'time travel random surfer model', which redefines how an web archive searcher (so-called *time traveler*) surfs in the web archive. The 'time travel random surfer model' is initiated by the 'random surfer model', which explains underlying PageRank [8]. The original model describes the surfing behavior of a web surfer that after following the link structure starting from a page for several steps and then jump to a random page. However, this surfer model does not well-captured the temporal nature of a longitudinal web archive. Surfing in the web archive, we assume that a user prefers search results from interesting time points. Hence, the surfing behaviour of a time traveler should be adapted to incorporate this temporally important aspect.

In this work, we model a time-travel surfing as in moves that consist of two distinct steps (i.e., non-temporal and temporal, as illustrated in Figure 1). At each move, a traveler starts with a *non-temporal* step. A searcher chooses to either follow the out-links of a page or jump to a remote page. In order to achieve the precise temporal authority of a page at a time point, we employ the fresh favoring mechanism as in [4], so that old pages are degraded. A traveler in our model prefers a new/fresh page. As contrast to [4], we do not consider the editorial behaviour of the page because we focus on *informational* queries, where the content of a document is nearly *static* over time. In order that, we take into account the freshness linked with the age of a document a at time t_i, $\mathcal{F}(a, t_i)$, which is quantified as:

$$\mathcal{F}(a, t_i) = e^{-\beta_1 (t_i - \mathcal{T}(a))} \quad (1)$$

where $\mathcal{T}$ denotes the first time a page appears in the collection. In the second (temporal) step, a traveler jumps to a snapshot of the page at different time points. To capture the temporal authority, we model the authority propagation flow among nodes at nearby time points. Here, we introduce two kinds of propagations: *forward-* (authority is propagated by past snapshots) and *backward-* (authority is propagated by future snapshots). The propagation is modeled in a decay fashion, and hence, in a time window with length controlled by a decay parameter. This authority propagation is different from [4] in the sense that, it helps capturing the precise temporal authority of a document at a given time point (instead of using for smoothing purpose). We model the length of the temporal propagation as controlled by the decay parameter β_2 (further explained in Section 2.2), that we model as a global parameter in this work.

2.2 Time-sensitive PageRank

In this section, we explain two novel methods to inject time preference into the PageRank: (1) the jumping probability at the 1^{st} step, so that the jumping scope is not restricted to within current snapshot but other snapshots and (2) via the transition probability between snapshots, at the 2^{nd} step.

In the normal case where no preferences are defined, the vector $\vec{v}$ which presents the jumping probability from node a to all the nodes N in the temporal graph is uniformly distributed ($= [\frac{1}{N}]N \times 1$). However, different from this ordinary behavior, we present a *query-dependent*, time-aware vector $\vec{v}_{temp}$ over the temporal graph as follows:

Time-aware Teleportation Instead of limiting the jumping scope in within a web snapshot, the traveler in this case can jump to any snapshot with a time preference. The probability of jumping from q to p at time t_i, $P_{t_i}(p|q, Jump)$, is dependent on the preference score of time t_i, $I(t_i)$. The

probability that a traveler reaches the page p at snapshot t_i can be written as[2]:

$$\pi_{p,i} = \sum_{t_j \in T_i} P_{t_i|t_j}(p) \sum_{q:q\to p|t_j} P_{t_j}(Follow|q) P_{t_j}(p|q, Follow) + \sum_{\forall q} P(Jump|q) I(t_i) \quad (2)$$

Time-aware Transition Probability For this second type of embedding time preference into the model, we modify the transition probability across time snapshots. Intuitively, a snapshot at time t_i with high time preference will have higher transition probability. In this case, the jumping scope is restricted within the time snapshot. The probability that a traveler reaches the page p at snapshot t_i can be written as:

$$\begin{aligned}\pi_{p,i} = & \sum_{t_j \in S_i} P_{t_i|t_j}(p) \sum_{q:q\to p|t_j} P_{t_j}(Follow|q) P_{t_j}(p|q, Follow) \\ & + \sum_{t_j \in S_i} P_{t_i|t_j}(p) \sum_{q|t_j} P_{t_j}(Jump|q) P_{t_j}(p|q, Jump) \\ = & \sum_{t_j \in S_i} P_{t_i|t_j}(p) \cdot \left[(1-\alpha) \sum_{q:q\to p|t_j} F_{t_j}(p,q) \cdot \pi_{q,j} + \alpha \sum_{q|t_j} \frac{\pi_{q,j}}{N_{t_j}} \right]\end{aligned}$$

where S_i is the set of snapshots which can directly distribute authority to t_i within one step. Even though presenting a similar generalization of the propagation model to us, the results in [4] indicate that the decay propagation is not most suitable for their task (normal web ranking). In our case, instead this transition probability (propagation) is strongly time-influenced. A node most propagates its authority to the time of interest (to help the time-aware ranking) that most near its time snapshot. The transition probability $P_{t_i|t_j}(p)$ is derived from the interestingness measure of the two time points and is calculated as:

$$P_{t_i|t_j}(p) = \frac{I(t_j)}{\sum_{\forall t_k} I(t_k)} \cdot w(t_i, t_j) \quad (4)$$

where $w(t_i, t_j) = e^{\beta_2 |t_i - t_j|}$. Hence the propagation scope is restricted to a time window $\mathcal{W}$ with the size (also size of S) is controlled by β_2. Within the time window $\mathcal{W}$, one with high time preference will be more likely to be propagated.

2.3 Time-based Diversity in Temporal Graph

In this section, we target another issue of the ranking problem, the time-based diversification of the top-k results.

Reinforcement in Random Walk

Mei et al. [5] introduce the integration of the vertex-reinforced random walk (VRRW) into the conventional PageRank to address the diversity ranking in graphs. Their intuition follows the 'rich gets richer' phenomenon, which specifically, the node that has been visited many times will have higher probability to be revisited again. Hence, the transition probability in the Markov random walk (to a state from others) is reinforced by the number of previous visits to that state. Our time-based diversity model follows the same intuition. The vertex reinforcement is applied within each snapshot, so that the *within*-snapshot neighbors of a popular node (visited many times) are penalized. For the authority propagation across time snapshots, however, this mechanism cannot be integrated directly. Instead, we follow a *voting* propagation mechanism. For every step, we check for the node snapshot with maximum number of visits over the propagation time window, and only this node snapshot got propagated from others. The other nodes receive no propagation from other nodes. Hence, this approach allows a partial *across*-snapshot penalty and helps the time-based diversity.

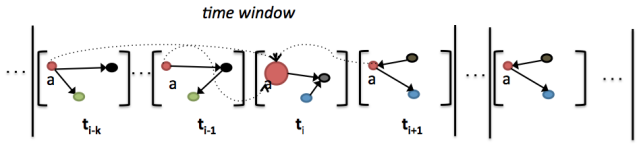

Figure 2: Reinforced/dynamic propagation over a time window. There is only one node gets all the propagations.

In a similar fashion to [5], the time-variant transition probability from q to p at step T (of the random walk) within a time snapshot t_i is defined as:

$$P^T_{t_i}(p,q) = (1-\alpha) \cdot s(p_{t_j}) + \alpha \cdot \frac{P^0_{t_i}(p,q) \cdot N^T_{t_i}(q)}{D^T_{t_i}(p)} \quad (5)$$

where $N^T_{t_i}(q)$ is the number of visits to q at step T. $D^T_{t_i}(p) = \sum_{q \in t_i} P^0_{t_i}(p,q) N^T_{t_i}(q)$. The cross snapshot- transition probability $P_{t_i|t_j}(p)$ at step T is:

$$P_{t_i|t_j}(p) = \begin{cases} \text{calculated as Equation 4} & if \\ p = \text{argmax}_{\forall t_i \in \mathcal{W}_{t_j}} N^T_{t_i}(p), & \\ 0 & otherwise. \end{cases} \quad (6)$$

where $\mathcal{W}_{t_j}$ is the time window of t_j, such that $\forall t_i \in \mathcal{W}_{t_j}$, $w(t_i, t_j) > 0$.

3. EXPERIMENTS

In this section, we evaluate the performances of 5 different methods: temp-BM25, temp-PageRank[3](baseline), our approach with time-aware teleportation (ours), our approach with vertex-reinforcement random walk (ours+div) and our approach with time-aware vertex-reinforcement random walk (ours+tempdiv).

3.1 Experiment settings

Dataset We utilize a corpus of archival web pages in .gov domain collected by the Internet Archive from January 1995 to September 2013. The corpus contains over 900 million of text captures and over 58.8 billion temporal links. In order to shrink down the huge collection to extract a subset/subgraph of interest, we follow the idea of recent work that exploiting the value of anchor text. First, we achieve over 60 related long-term controversial political topics from the debate website[4]. We then look into the document linking graph and extract the links with anchor text reflecting any of the topics (both lexicographically and semantically). We

[2]One can introduce a parameter α in the formula so that $\sum_p \pi_{p,i} = 1$. However, to simplify the problem, we omit the parameter.

[3]To incorporate temporal prior, we apply PageRank and BM25 at each time snapshot and then multiply them with the corresponding time preference score.

[4]http://www.debate.org/big-issues/

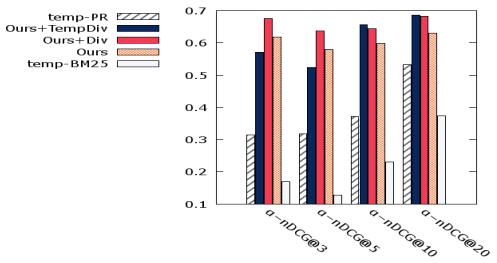

Figure 3: Performance of time-based subtopic diversity.

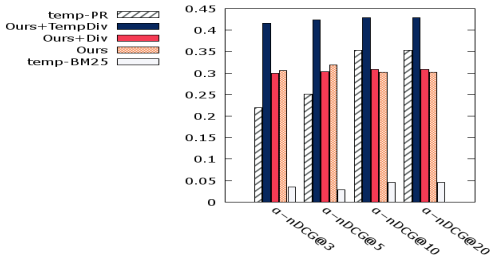

Figure 4: Performance of content-based subtopic diversity.

then captured the source and destination web pages of these links and treat them as the document seeds. We further capture the in-pages pointing to source pages and out-pages (which the destinations point to) to achieve a substantially large collection of over 100 million document revisions (approx. 40 million unique documents). We then picked 20 controversial/informational queries to conduct the experiment.

Ground Truth and Metrics Since there is no publicly available gold standard for our work, we rely on manual annotation for the 20 queries. For each document, we asked human experts, whether it is relevant to (1) one of the time-points and (2) any of the content-based subtopics. The scale for time is binary, whereas to account for authority, we use a scale from 0 (not related) to 4 (highly relevant) for the subtopic relevance judgment. A document with score larger than 2 is considered as relevant. For evaluation, follow the adopted setting of recent TREC web tracks (diversity task) as presented in [6], we use a generalization of a well-known result diversification metric (that accounts for both diversity and relevance), α-$nDCG$ [3] (so that it takes into account the subtopic weights). We consider two different subtopic dimensions, (1) *time* - with associated weight mined from the anchor-text distribution and (2) subtopics mined from *content* - equally weighted.

Priors and Parameter Tuning For the temporal prior, we mined from the anchor text frequency distribution (to mimic the query logs, following [7]). For the relevance prior, we utilize the scores from the BM25 retrieval model. For the random walk parameters, we set the jumping probability to be the default 0.15. All decay parameters are set to 0.4. All algorithms are run over Apache Giraph[5].

3.2 Experiment results

Figures 3 and 4 show the *time* and *subtopic* diversification results of the compared models respectively. For the *time* diversity, our models outperform the baseline significantly ($p < 0.05$) at $k = 3$ and $k = 5$. It is empirically found that, our propagation method helps identifying the temporal authority with regards to the relevant time more precisely. For example, given the query electoral college and a time period February 2009, a document issued in September 2008 has a high score for the traditional PageRank. However, our propagation accounts for both *freshness* and *lagging* ranks another document issued in February 2009 higher (that is more time relevant). The good performance of our approach shows that we capture better the temporal authority of documents with regards to the time preference. Our temporal diversity method also shows that it diversifies time effectively. The results for *content*-based subtopic diversity measurement also indicate a good performance of our method. Ours+tempdiv best performs (significant with $p < 0.05$) for all cases. This rather shows the effectiveness of our time-aware diversity approach.

[5]http://giraph.apache.org/

4. CONCLUSIONS

In this paper, we have studied the problem of finding important document in the web archive and address it in a unified framework that integrates relevance, temporal authority, diversity and time together. In detail, we proposed a novel random walk model incorporating time and the link structure of the web archive. Our model is shown to outperform PageRank for both relevance and diversity tasks. For future work, we would like to investigate the application of the novel model on different open challenges of web archive, i.e., time-aware summarization. Scalability issues will also be another target.

Acknowledgments The work was partially funded by the European Commission for the ERC Advanced Grant ALEXANDRIA under grant No. 339233 and the FP7 project ForgetIT under grant No. 600826. We thank the Internet Archive and Altiscale for providing access to the archived data.

References

[1] K. Berberich, M. Vazirgiannis, and G. Weikum. Time-aware authority ranking. *Internet Mathematics*, 2(3):301–332, 2005.
[2] X.-Q. Cheng, P. Du, J. Guo, X. Zhu, and Y. Chen. Ranking on data manifold with sink points. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1):177–191, 2013.
[3] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR'2008*, pages 659–666.
[4] N. Dai and B. D. Davison. Freshness matters: in flowers, food, and web authority. In *Proceedings SIGIR'2010*, pages 114–121.
[5] Q. Mei, J. Guo, and D. Radev. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of SIGKDD'2010*, pages 1009–1018.
[6] T. N. Nguyen and N. Kanhabua. Leveraging dynamic query subtopics for time-aware search result diversification. In *Proceedings of ECIR'2014*, pages 222–234.
[7] T. N. Nguyen, N. Kanhabua, W. Nejdl, and C. Niederée. Mining relevant time for query subtopics in web archives. In *Proceedings of WWW'2015 companion*, pages 1357–1362.
[8] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.
[9] L. Yang, L. Qi, Y.-P. Zhao, B. Gao, and T.-Y. Liu. Link analysis using time series of web graphs. In *Proceedings of CIKM'2007*, pages 1011–1014.
[10] P. S. Yu, X. Li, and B. Liu. On the temporal dimension of search. In *Proceedings of WWW'2004*, pages 448–449.
[11] X. Zhu, A. B. Goldberg, J. Van Gael, and D. Andrzejewski. Improving diversity in ranking using absorbing random walks. In *Proceedings of HLT-NAACL'2007*, pages 97–104.