Temporal Wikipedia search by edits and linkage

Julianna Göbölös-Szabó András A. Benczúr
Computer and Automation Research Institute, Hungarian Academy of Sciences (MTA SZTAKI)
{gszj, benczur}@ilab.sztaki.hu

ABSTRACT

In this paper we exploit the connectivity structure of *edits* in Wikipedia to identify recent events that happened at a given time via identifying bursty changes in linked articles around a specified date. Our key results include algorithms for node relevance ranking in temporal subgraph and neighborhood selection based on measurements for structural changes in time over the Wikipedia link graph. We measure our algorithms over manually annotated queries with relevant events in September and October 2011; we make the assessment publicly available. While our methods were tested over clean Wikipedia metadata, we believe the methods are applicable to general temporal Web collections as well.

1. INTRODUCTION

Considering a chain of events, we are often interested in the causes and effects, naturally represented by citations and links. If we want to understand what and why did happen, what other stories had influences on the event, it is worth discovering connected articles. The problem is even more interesting if we want to know how a story evolved in time. In this case we also need the information about the time of appearance of pages and links, and this can help understanding the temporal causality of the analyzed event.

In this paper we develop methods to automatically discover temporal events along important connections. For our experiments we selected Wikipedia as a clean corpus where measurements are not biased for example by date identification, yet the methods can directly be applied for any hyperlinked collection. Wikipedia is certainly the most used and best-known online encyclopedia and knowledge-base of the past decade. Almost every action or event, be it tiny or slightly remarkable, immediately appears in blog posts, news articles or sometimes even in Wikipedia articles.

Certainly not all Wikipedia modifications are triggered by headline news. People contribute to pages for various reasons. Sometimes a mistake is found and gets corrected or the editor has a special field of interest without a satisfying of coverage so she starts to add new pieces of information to the encyclopedia. In this case we expect isolated edits of a low number of editors and hence less accumulated change in the neighborhood. In addition, these pages, even being just created, will link old, stable pages and will collect incoming links with moderate speed.

In our experiments we use three monthly snapshots of Wikipedia from September to November 2011. We selected 20 queries for September and 15 for October and manually assessed the articles for relevant events at that time. We made the queries and the set of relevant articles public[1].

Our results complement recent results on temporal information retrieval where the main goal was to correctly date the events mentioned in Web pages. In Wikipedia we may rely on exact metadata to date addition and deletion and our main goal is to distinguish bursty, time relevant modifications from sporadic edits of past events and general, time independent information.

For ranking we use both the link structure and the content. The user can specify a query and should get a "temporally changing" subgraph of relevant articles. First we try to find the relevant articles respective to the query by a text search engine. As the next step, based on these articles, we try to find those nodes that have not only important changes according to the definition above but also their content is related to the original query. In a recursive definition reminiscent of PageRank [12] and HITS [8], we will consider the change of a page relevant if relevant changes can be observed in the neighborhood of the page as well. Finally we retrieve and present the top ranking articles and their linkage.

2. RANKING BY CHANGE AND LINKAGE

In our temporal information retrieval task, the user specifies a broad topic (e.g. Arab spring) and a date. Relevant documents should describe events that happened around the specific date involving the broad topic in question. A query with sample relevant Wikipedia articles is in Table 1.

Our ranking model combines text relevance with scores for change at the specified date that we boost by bursty changes of interlinked articles. First we identify a seed set scored by classical ranking techniques, e.g. Okapi BM25.We extend this seed set by neighboring articles that changed. These steps yield a candidate subgraph that is small enough to run subgraph scoring at query time, yet sufficiently large to contain most information relevant to the user query.

2.1 Measure of change in time

In order to discover recent events and trends, we consider changes both in linkage and content. We expect pages related to a certain event increase in content as well as connectivity of both in and out-edges after the specific event. Consequently, we measure change

As illustrated in Table 1, we measure change as the sum of the change of the logarithm of the in and out-degree as well as the absolute difference between the number of words in the article between two fixed dates t_1 and t_2 as

$$change(u) = |\log \frac{deg_{in,t_1}(u)}{deg_{in,t_2}(u)}|+|\log \frac{deg_{out,t_1}(u)}{deg_{out,t_2}(u)}|+|\log \frac{words_{t_1}(u)}{words_{t_2}(u)}|.$$

This work was supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956), by the EU FP7 Project LAWA—Longitudinal Analytics of Web Archive Data, and by the grant OTKA NK 105645. The work of the first author reported in the paper has been developed in the framework of the project "Talent care and cultivation in the scientific workshops of BME" project. This project is supported by the grant TÁMOP–4.2.2.B-10/1–2010-0009.

TAIA'13 August 1, 2013, Dublin, Ireland.
.

[1] http://dms.sztaki.hu/en/download/wimmut-searching-and-navigating-wikipedia

		Sep-Oct	Oct-Nov	Sep-Nov
Muammar Gaddafi	content	0.044	0.18	0.23
	inlink	0.55	0.12	0.68
	outlink	0.033	0.04	0.074
Death of Muammar Gaddafi	content	0	7.71	7.71
	inlink	0	4.21	4.21
	outlink	0	4.64	4.64
Battle of Sirte (2011)	content	7.78	0.79	8.56
	inlink	4.78	0.21	4.99
	outlink	4.9	0.14	5.06

Table 1: Change of articles related to Muammar Gaddafi.

2.2 Expanding the seed set

Seed expansion requires a score over the nodes that measure their relevance and freshness. In a naive solution one would specify a given number of steps and consider the entire neighborhood in this distance. As it turns out, even the one-step neighborhood is too large and hence we have to score candidate neighbors v. We use the following formula:

$$\text{score}(v) = \max_{u \in \text{ seed}} \text{BM25}(u) + \text{change}(u) + \text{change}(v). \quad (1)$$

2.3 Scores for change and relevance

We take a convex combination of the IR and change scores by a parameter α. Before combination, we transform both IR and change scores into $[0, 1]$. IR scores are normalized by the maximum while change scores are saturated by using parameter T in order to avoid extreme large values of change. The final formula becomes

$$p(u) = \alpha \cdot \frac{\text{IR(u)}}{\text{maxIR}} + (1 - \alpha) \cdot \frac{\text{change}(u)}{(\text{change}(u) + T)}. \quad (2)$$

The above score forms a class of baseline ranking schemes depending on the parameters. While the dependence on T turned out to be relative low, the values of α balance between two extremities. Case $\alpha = 1$ returns the text relevance score and case $\alpha = 0$ takes only the amount of change into account. Note that even in this case, text relevance scores are involved in the seed set and the expansion process.

2.4 Personalized PageRank, random walks and electric networks

Our first algorithm scores graph nodes by PageRank [12] personalized on the IR score. In [6] an electric network based method is presented to select a subgraph connecting a set of nodes; we show that this result is a special case of our personalized PageRank method. We briefly review some useful properties of personalized PageRank from e.g. [13] and their connection to the electric network formulation of [6]. Let $\text{PPR}_p(v)$ denote the personalized PageRank of vertex $v \in V$ where $p = p(u) \in \mathbb{R}^{|V|}$ is the personalization vector. Then personalized PageRank is the solution of the system of equations

$$\text{PPR}_p(v) = (1 - c) \sum_{uv \in E} \text{PPR}_p(u) \frac{w(uv)}{w(u)} + c \cdot p(v), \quad (3)$$

where $w(uv)$ denotes the edge weight normalized so that the total weight of out-edges from u is 1. PageRank is equal to the probability that a random walk of length drawn from a geometric distribution terminates at the given node:

$$\text{PPR}_p(v) = \sum_{k=0}^{\infty} c(1 - c)^k \sum_{v_0, v_1, \ldots, v_k = v} p(v_0) \cdot w(v_0 v_1) \cdots w(v_{k-1} v_k). \quad (4)$$

Next we consider special personalization vectors that apply for a single node only. With an abuse of notation, PPR_u will denote personalization to a vector p with $p(u) = 1$ and 0 otherwise. For such personalization vectors, the system of equations is equivalent to

$$PPR_u(x) = (1 - c) \sum_{uv \in E} PPR_v(x) \cdot w(uv) + c \cdot p(v). \quad (5)$$

The electric network formulation of [6] uses the equation

$$V(u) = \sum_v V(v) \cdot w(uv) \ \forall u \neq s, t \quad (6)$$

with boundary conditions $V(s) = 1$ and $V(t) = 0$, see [6] for details. Note that equations (5) and (6) have the same form with V_u corresponding to PPR_u except for the additive term $c \cdot p(v)$. These terms in equations (3) and (5) correspond to a universal sink S which can be added to the electric network with edge weight $w(v, S) = p(v)$. Universal sinks are also introduced in [6] with the difference that their method immediately taxes the large degree nodes while the uniform additive term in (5) taxes equally, regardless of the degree.

2.5 Personalized HITS

Our next class of graph ranking procedures are based on HITS [8]. HITS is known to be vulnerable to topic drift, the preference of nodes in a large but irrelevant clique or dense region in the neighborhood of the original topic. A few papers consider the question of personalizing HITS to reduce topic drift [2, 7] but these algorithms are rather complex.

We give a simple personalization to HITS by using a "supersource". We can think of the supersource as a new node of the graph which is connected with each node of the original graph, and the weight of an edge corresponds to the importance of the respective node in the personalization, with weight 0 also allowed. The supersource distributes a fixed amount of score in each iteration split proportional to the personalization distribution. At the end of each iteration, we normalize the authority and hub vectors, so the maximal element in the vector is 1. With the notation of a as the vector of authorities, h as the vector of hubs, c as the importance of the supersource and p as the personalization vector, we have

$$\hat{a}(v) = \sum_{uv \in E} w(uv) \cdot h(u) + c \cdot p(v), \quad a = \hat{a}/\|a\|_\infty; \quad (7)$$

$$\hat{h}(v) = \sum_{vu \in E} w(vu) \cdot a(u) + c \cdot p(v), \quad h = \hat{h}/\|h\|_\infty, \quad (8)$$

where $w(vu)$ denotes the weight of vu. We obtain personalized vectors a and h; the corresponding node scoring method will be denoted by *HitsAuth* and *HitsHub*.

3. EXPERIMENTS

Our experiments are based on three monthly Wikipedia snapshots of 2011-09-01, 2011-10-07 and 2011-11-15, with over 7M nodes and 180M edges. We selected 35 queries that are related to headline news events either from September or from October 2011. For each query we set a list of manually

	Month of change			The other month	
	NDCG	recall	MRR	recall	MRR
None	0.243	0.456	0.865	0.327	0.423
PageRank-0.9	0.258	**0.555**	0.964	0.368	0.731
HitsAuth-200	**0.315**	**0.543**	**1.101**	0.377	0.689
HitsHubs-200	0.266	**0.531**	0.953	0.378	0.527
Combined	0.268	**0.557**	0.994	0.365	0.746

Table 2: Recall@15 and MRR with the overall best parameter settings. Scores in bold show performances statistically significantly better than the baseline ($p < 0.01$).

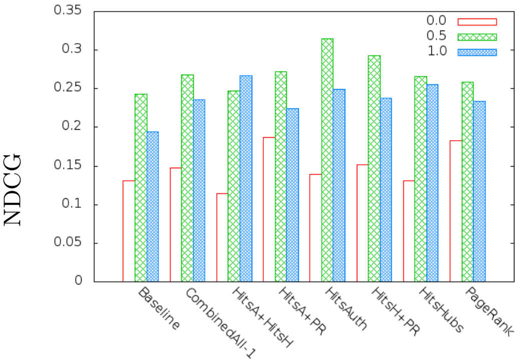

Figure 1: NDCG as a function of the values of α (0.0, 0.5 and 1.0), for different algorithms and the baseline. The top 100 BM25 score articles are expanded by another 100. Change is saturated with $T = 10$ and personalization is $c = 0.9$ for PageRank and 200 for HITS.

selected relevant articles. We made the list of queries and the assessment publicly available[2].

For all queries, we measured our methods focusing both on the change from September to October and from October to November. We expect that September events score higher in the first while October events in the second case. Results for the accuracy measures are found in Table 2.

3.1 Evaluation measures

We evaluate the performance by the Normalized Discounted Cumulative Gain (NDCG), Mean Reciprocal Rank (MRR), Early Recall, Graph Density Change and Tendency of short paths staying within the top hits. The last two measures consider the change in the connectivity of top ranked articles. We not only expect that the pages in the result set are relevant but we also show temporal evolution, growth or densification. Therefore we also measure the number of edges for the top ranked 15 articles.

We apply PageRank to measure the fraction of short paths staying inside the top 15 hits. By equation (4) if we define a personalization vector p that is identically distributed over the nodes of the selected subgraph, the sum of $\mathrm{PPR}_p(v)$ over nodes v of the subgraph gives a weighted sum of paths that terminate within the subgraph.

3.2 Retrieval performance

We overview the results of various combinations of change measures, graph ranking and the BM25 score in Table 2 with statistically significant improvements shown bold ($p < 0.01$). As best parameters we identified the following values or ranges: seed set size of 10-100; seed set expanded by 100 more articles; change saturation $T = 10$ (it has little effect); change and IR combination ratio $\alpha = 0.5$; PageRank damping $c = 0.9$ and HITS personalization $c = 200$.

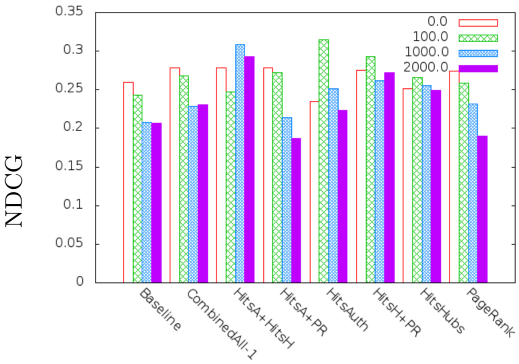

Figure 2: NDCG as the function of the size of the expansion (0, 100, 1000 and 2000), for different algorithms and the baseline. Here $\alpha = 0.5$, change is saturated with $T = 10$, and personalization is $c = 0.9$ for PageRank and 200 for HITS.

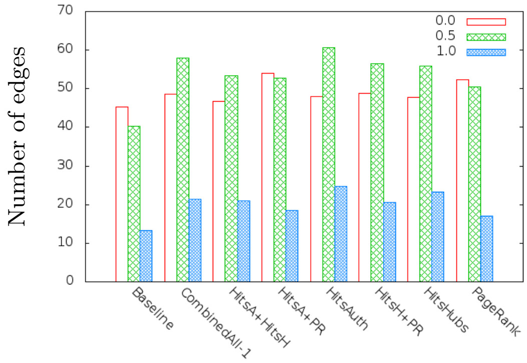

Figure 3: The number of edges among the top 15 hits in the snapshots before and after the event as well as their difference as the function of α (0.0, 0.5 and 1.0), for different algorithms and the baseline.

In Fig. 1 we show how NDCG is influenced by the value of α, and in Fig. 2 we show how NDCG is influenced by size of the expansion. We observe that the balanced mix of $\alpha = 0.5$ is the best choice by including both text based relevance and change measures. The importance of the temporal aspect of our queries is clear in the weaker performance of the BM25 score itself ($\alpha = 1$) while the change-only $\alpha = 0$ performs weakest.

We should be careful in expanding the seed set: best results are obtained if we extend the original top 100 articles with another 100 changing ones in the neighborhood. However, for much larger subgraphs, all algorithms show topic drift and strong personalization is needed with $c = 0.9$-0.95 for PageRank and 100-200 for HITS.

3.3 Graph density

We compare the quality and connectivity of the linkage within the top results both by counting the edges and computing the sum of personalized PageRank kept within the displayed result set in Figs. 3–4. Note that the denser sub-

[2] `http://dms.sztaki.hu/en/download/wimmut-searching-and-navigating-wikipedia`

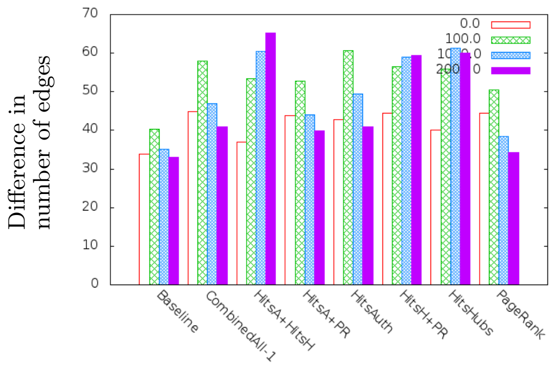

Figure 4: The difference of the number of edges among the top 15 hits between the snapshots before and after the event as the function of the size of the expansion (0, 100, 1000 and 2000), for different algorithms and the baseline.

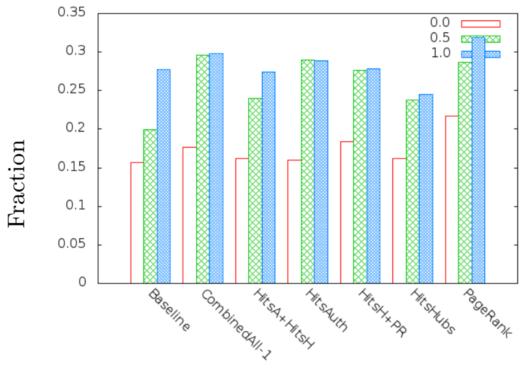

Figure 5: The fraction of short paths kept within the top 15 hits as the function of α (0.0, 0.5 and 1.0), for different algorithms and the baseline.

graphs are also of improved relevance, as seen in Section 3.2.

When comparing the effect of α (Fig. 3) and the expansion size (Fig. 4), we note that the graph algorithms tend to overfit for irrelevant lists, for example for very large expansion (1000-2000) and $\alpha = 1$ considering change only. The results show clear topic drift in these cases while good performance for low expansion. The findings are similar for the PageRank based measure of short paths staying inside the subgraph in Fig. 5.

4. RELATED RESULTS

Temporal information retrieval has mainly been considered as a task for either temporal query aspect detection or Web content dating. We are aware of no results for ranking with respect to a specified date as part of the input query. In a result with goals similar to ours for timestamped news [5], time sensitive queries are analyzed by relying on the publication date of documents. However, their goal is to identify relevant time ranges for queries, unlike in our result where we search for events in a given time related to the query.

Similar to our task is the identification of important events from the Blogosphere [10] as in the TREC Top Stories Identification task. Among others these results rely on timing and relevance as key factors but their results do not take connectivity into account. Topic detection by analyzing term bursts is first described in [9]; subsequent results rely on topic detection and tracking (TDT) achievements [3]. The general properties of the Wikigraph including degrees and their change in time is measured in [1].

As a different task, the extraction of a chronological order from free text turns out to be a difficult [11] and considered as part of the TAC Temporal Slot Filling task [14, 15]. In one application, the timeline of events related to G8 leaders is extracted [4] by starting with a query for a given politician and then identifying the date from free text. We believe that these tasks can be enhanced by our techniques.

Conclusions

We identified events in time by relying on edit dates aggregated in a neighborhood defined by hyperlinks. We proposed algorithms based on personalized HITS and PageRank that amplify changes and relevance in a given graph neighborhood. Part of our results is a query set with relevance assessment suited for the data set as well as the annotated document collection. We believe that our results find application in other related social networking and Web IR tasks.

5. REFERENCES

[1] L. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal analysis of the wikigraph. In *Web Intelligence*, pp. 45–51. IEEE, 2006.

[2] H. Chang, D. Cohn, and A. McCallum. Learning to Create Customized Authority Lists. In *Proc. 7th ICML*, pp. 127–134, 2000.

[3] K. Chen, L. Luesukprasert, S. Chou, et al. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE TKDE*, 19(8):1016–1025, 2007.

[4] H. Chieu and Y. Lee. Query based event extraction along a timeline. In *Proc. 27th ACM SIGIR*, pp. 425–432, 2004.

[5] W. Dakka, L. Gravano, and P. Ipeirotis. Answering general time-sensitive queries. *IEEE TKDE*, 24(2):220–235, 2012.

[6] C. Faloutsos, K. McCurley, and A. Tomkins. Fast discovery of connection subgraphs. In *Proc. 10th ACM SIGKDD*, pp. 118–127. ACM, 2004.

[7] K. Kim and S. Cho. Personalized mining of web documents using link structures and fuzzy concept networks. *Applied Soft Computing*, 7(1):398–410, 2007.

[8] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[9] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[10] Y. Lee, H. Jung, W. Song, and J. Lee. Mining the blogosphere for top news stories identification. In *Proc. 33rd ACM SIGIR*, pp. 395–402. ACM, 2010.

[11] I. Mani and G. Wilson. Robust temporal processing of news. In *Proc. 38th ACL*, pp. 69–76, 2000.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford University, 1998.

[13] T. Sarlós, A. A. Benczúr, K. Csalogány, D. Fogaras, and B. Rácz. To randomize or not to randomize: Space optimal summaries for hyperlink analysis. In *Proc. 15th WWW*, pp. 297–306, 2006.

[14] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proc. 4th International Workshop on Semantic Evaluations*, pp. 75–80. ACL, 2007.

[15] M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proc. 5th International Workshop on Semantic Evaluation*, pp. 57–62. ACL, 2010.

[16] S. White and P. Smith. Algorithms for estimating relative importance in networks. *Proc 9th ACM SIGKDD*, page 266, 2003.