Diversifying Search Results Using Time
An Information Retrieval Method for Historians

Dhruv Gupta[1,2] and Klaus Berberich[1(✉)]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany
kberberi@mpi-inf.mpg.de
[2] Saarbrücken Graduate School of Computer Science, Saarbrücken, Germany
dhgupta@mpi-inf.mpg.de

Abstract. Getting an overview of a historic entity or event can be difficult in search results, especially if important dates concerning the entity or event are not known beforehand. For such information needs, users benefit if returned results covered diverse dates, thus giving an overview of what has happened throughout history. Such a method can be a building block for applications, for instance, in *digital humanities*. We describe an approach to diversify search results using temporal expressions (e.g., 1990s) from their contents. Our approach first identifies time intervals of interest to the given keyword query based on pseudo-relevant documents. It then re-ranks query results so as to maximize the coverage of identified time intervals. We present a novel and objective evaluation for our proposed approach. We test the effectiveness of our methods on The New York Times Annotated corpus and the Living Knowledge corpus, collectively consisting of around 6 million documents. Using history-oriented queries and encyclopedic resources we show that our method is able to present search results diversified along time.

1 Introduction

Large born-digital document collections are a treasure trove of historical knowledge. Searching these large longitudinal document collections is only possible if we take into account the temporal dimension to organize them. We present a method for diversifying search results using temporal expressions in document contents. Our objective is to specifically address the information need underlying *history-oriented* queries; we define them to be keyword queries describing a historical event or entity. An ideal list of search results for such queries should constitute a *timeline* of the event or portray the *biography* of the entity. This work shall yield a useful tool for scholars in history and humanities who would like to search large text collections for *history-oriented* queries without knowing relevant dates for them apriori.

No work, to the best of our knowledge, has addressed the problem of diversifying search results using temporal expressions in document contents. Prior approaches in the direction of diversifying documents along time have relied largely on publication dates of documents. However a document's publication

N. Ferro et al. (Eds.): ECIR 2016, LNCS 9626, pp. 789–795, 2016.
DOI: 10.1007/978-3-319-30671-1_69

date may not necessarily be the time that the text refers to. It is quite common to have articles that contain a historical perspective on a past event from the current time. Hence, the use of publication dates is clearly insufficient for history-oriented queries.

In this work, we propose a probabilistic framework to diversify search results using temporal expressions (e.g., 1990s) from their contents. First, we identify time intervals of interest to a given keyword query, using our earlier work [7], which extracts them from pseudo-relevant documents. Having identified time intervals of interest (e.g., [2000,2004] for the keyword query george w. bush), we use them as aspects for diversification. More precisely, we adapt a well-known diversification method [1] to determine a search result that consists of relevant documents which cover all of the identified time intervals of interest.

Evaluation of historical text can be highly subjective and biased in nature. To overcome this challenge; we view the evaluation of our approach from a statistical perspective and take into account an objective evaluation for automatic summarization to measure the effectiveness of our methods. We create a large history-oriented query collection consisting of long-lasting wars, important events, and eminent personalities from reliable encyclopedic resources and prior research. As a ground truth we use articles from *Wikipedia*[1] concerning the queries. We evaluate our methods on two large document collections, the New York Times Annotated corpus and the Living Knowledge corpus. Our approach is thus tested on two different types of textual data. One being highly authoritative in nature; in form of news articles. Another being authored by real-world users; in form of web documents. Our results show that using our method of diversifying search results using time; we can present documents that serve the information need in a history-oriented query very well.

2 Method

Notation. We consider a document collection $\mathcal{D}$. Each document $d \in \mathcal{D}$ consists of a multiset of keywords d_{text} drawn from vocabulary $\mathcal{V}$ and a multiset of temporal expressions d_{time}. Cardinalities of the multisets are denoted by $|d_{text}|$ and $|d_{time}|$. To model temporal expressions such as 1990s where the begin and end of the interval can not be identified, we utilize the work by Berberich et al. [3]. They allow for this uncertainty in the time interval by associating lower and upper bounds on begin and end. Thus, a temporal expression T is represented by a four-tuple $\langle b_l, b_u, e_l, e_u \rangle$ where time interval $[b, e]$ has its begin bounded as $b_l \leq b \leq b_u$ and its end bounded as $e_l \leq l \leq e_u$. The temporal expression 1990s is thus represented as $\langle 1990, 1999, 1990, 1999 \rangle$. More concretely, elements of temporal expression T are from time domain $\mathcal{T}$ and intervals from $\mathcal{T} \times \mathcal{T}$. The number of such time intervals that can be generated is given by $|T|$.

Time Intervals of Interest to the given keyword query q_{text} are identified using our earlier work [7]. A time interval $[b, e]$ is deemed *interesting* if its referred

[1] https://en.wikipedia.org/.

frequently by highly relevant documents of the given keyword query. This intuition is modeled as a two-step generative model. Given, a set of pseudo-relevant documents R, a time interval $[b, e]$ is deemed interesting with probability:

$$P([b,e] \mid q_{text}) = \sum_{d \in R} P([b,e] \mid d_{time}) P(d_{text} \mid q_{text}).$$

To diversify search results, we keep all the time intervals generated with their probabilities in a set q_{time}.

Temporal Diversification. To diversify search results we adapt the approach proposed by Agrawal et al. [1]. Formally, the objective is to maximize the probability that the user sees at least one result relevant to her time interval of interest. We thus aim to determine a query result $S \subseteq R$ that maximizes

$$\sum_{[b,e] \in q_{time}} \left(P([b,e] \mid q_{text}) \left(1 - \prod_{d \in S} (1 - P(q_{text} \mid d_{text}) P([b,e] \mid d_{time})) \right) \right).$$

The probability $P([b, e] \mid q_{text})$ is estimated as described above and reflects the salience of time interval $[b, e]$ for the given query. We make an independence assumption and estimate the probability that document d is relevant and covers the time interval $[b, e]$ as $P(q_{text} \mid d_{text})\, P([b, e] \mid d_{time})$. To determine the diversified result set S, we use the greedy algorithm described in [1].

3 Evaluation

Document Collections. We used two document collections one from a news archive and one from a web archive. The Living Knowledge[2] corpus is a collection of news and blogs on the Web amounting to approximately 3.8 million documents [8]. The documents are provided with annotations for temporal expressions as well as named-entities. The New York Times (NYT) Annotated[3] corpus is a collection of news articles published in *The New York Times*. It reports articles from 1987 to 2007 and consists of around 2 million news articles. The temporal annotations for it were done via SUTime [6]. Both explicit and implicit temporal expressions were annotated, resolved, and normalized.

Indexing. The document collections were preprocessed and subsequently indexed using the ElasticSearch software[4]. As an ad-hoc retrieval baseline and for retrieval of pseudo–relevant set of documents we utilized the state-of-the-art *Okapi-BM25* retrieval model implemented in ElasticSearch.

Collecting History-Oriented Queries. In order to evaluate the usefulness of our method for scholars in history, we need to find keyword queries that are highly

[2] http://livingknowledge.europarchive.org/.
[3] https://catalog.ldc.upenn.edu/LDC2008T19.
[4] https://www.elastic.co/.

ambiguous in the temporal domain. That is multiple interesting time intervals are associated with the queries. For this purpose we considered three categories of history-oriented queries: long-lasting wars, recurring events, and famous personalities. For constructing the queries we utilized reliable sources on the Web and data presented in prior research articles [7,9]. Queries for long-lasting wars were constructed from the *WikiWars* corpus [9]. The corpus was created for the purpose of temporal information extraction. For ambiguous important events we utilized the set of ambiguous queries used in our earlier work [7]. For famous personalities we use a list of most influential people available on the USA Today[5] website. The names of these famous personalities were used based on the intuition that there would important events associated with them at different points of time. The keyword queries are listed in our accompanying technical report [11]. The entire testbed is publicly available at the following url:

http://resources.mpi-inf.mpg.de/dhgupta/data/ecir2016/.

The objective of our method is to present documents that depict the historical timeline or biography associated with keyword query describing event or entity. We thus treat the set of diversified set of documents as a *historical summary* of the query. In order to evaluate this diversified summary we obtain the corresponding *Wikipedia* (see footnote 1) pages of the queries as ground truth summaries.

Baselines. We considered three baselines, with increasing sophistication. As a naïve baseline, we first consider the pseudo-relevant documents retrieved for the given keyword query. The next two baselines use a well-known implicit diversification algorithm *maximum marginal relevance* (MMR) [5]. Formally it is defined as: $\text{argmax}_{d \notin S} \left[\lambda \cdot \text{sim}_1 (q, d) - (1 - \lambda) \cdot \max_{d' \in S} \text{sim}_2 (d', d)\right]$. MMR was simulated with sim_1 using query likelihoods and sim_2 using cosine similarity between the term-frequency vectors for the documents. The second baseline considered MMR with $\lambda = 0.5$ giving equal importance to query likelihood and diversity. While the final baseline considered MMR with $\lambda = 0.0$ indicating complete diversity. For all methods the summary is constructed by concatenating all the top-k documents into one large document.

Parameters. There are two parameters to our system: (i) The number of documents considered for generating time intervals of interest $|R|$ and (ii) The number of documents considered for *historical summary* $|S|$. We consider the following settings of these parameters: $|R| \in \{100, 150, 200\}$ and $|S| \in \{5, 10\}$.

Metrics. We use the ROUGE-N measure [12] (implementation[6]) to evaluate the *historical summary* constituted by diversified set of documents with respect to the ground truth. ROUGE-N is a recall-oriented metric which reports the number of n-grams matches between a candidate summary and a reference summary. The n in *ngram* is the length of the gram to be considered; we limit ourselves to $n \in \{1, 3\}$. We report the *recall*, *precision*, and $F_{\beta=1}$ for each ROUGE-N measure.

[5] http://usatoday30.usatoday.com/news/top25-influential.htm.
[6] http://www.berouge.com/Pages/default.aspx.

Table 1. Results for the New York Times Annotated corpus.

	Category	Historical Wars						Historical Events						Historical Entity					
	Metric	R		P		$F_{\beta=1.0}$		R		P		$F_{\beta=1.0}$		R		P		$F_{\beta=1.0}$	
	Rouge-N	1	3	1	3	1	3	1	3	1	3	1	3	1	3	1	3	1	3
$\|R\|$=100 $\|S\|$=5	Naïve	30.5	12.0	62.7	23.5	33.9	13.2	43.3	18.0	42.4	15.7	21.0	8.4	19.9	7.9	74.6	29.8	24.4	9.8
	MMR (λ=0.5)	30.5	12.0	62.8	23.6	33.9	13.2	43.3	18.0	42.6	15.6	21.1	8.4	20.0	7.9	74.3	29.6	24.6	9.8
	MMR (λ=0.0)	30.5	12.0	62.8	23.6	33.9	13.2	43.3	18.0	42.6	15.6	21.1	8.4	20.0	7.9	74.3	29.6	24.6	9.8
	Time-Diverse	46.4	17.5	55.7	21.1	41.0	15.5	56.7	22.0	35.9	13.0	26.3	9.9	35.3	13.4	67.0	25.3	34.5	13.1
$\|R\|$=100 $\|S\|$=10	Naïve	48.0	18.4	51.0	18.9	39.2	15.0	57.6	22.9	33.4	12.0	23.1	8.7	35.4	13.6	67.4	26.7	34.4	13.5
	MMR (λ=0.5)	48.4	18.5	50.6	18.8	39.2	15.0	57.5	22.9	33.4	11.9	23.1	8.7	35.8	13.7	67.2	26.8	34.7	13.6
	MMR (λ=0.0)	48.4	18.5	50.6	18.8	39.2	15.0	57.5	22.9	33.4	11.9	23.1	8.7	35.8	13.7	67.2	26.8	34.7	13.6
	Time-Diverse	64.8	24.4	43.2	16.5	42.6	16.3	66.1	24.3	27.1	8.9	23.1	8.0	48.2	17.8	56.9	21.1	36.8	13.7
$\|R\|$=150 $\|S\|$=5	Naïve	30.5	12.0	62.7	23.5	33.9	13.2	43.3	18.0	42.4	15.7	21.0	8.4	19.9	7.9	74.6	29.8	24.4	9.8
	MMR (λ=0.5)	30.5	12.0	62.8	23.6	33.9	13.2	43.3	18.0	42.6	15.6	21.1	8.4	20.0	7.9	74.3	29.6	24.6	9.8
	MMR (λ=0.0)	30.5	12.0	62.8	23.6	33.9	13.2	43.3	18.0	42.6	15.6	21.1	8.4	20.0	7.9	74.3	29.6	24.6	9.8
	Time-Diverse	48.2	18.6	55.1	21.1	42.0	16.2	58.1	22.6	33.4	12.2	25.7	9.6	38.0	14.1	65.3	23.9	36.7	13.7
$\|R\|$=150 $\|S\|$=10	Naïve	48.0	18.4	51.0	18.9	39.2	15.0	57.6	22.9	33.4	12.0	23.1	8.7	35.4	13.6	67.4	26.7	34.4	13.5
	MMR (λ=0.5)	48.5	18.6	50.7	18.8	39.3	15.1	57.5	22.9	33.4	11.9	23.1	8.7	35.7	13.7	67.3	26.8	34.7	13.7
	MMR (λ=0.0)	48.5	18.6	50.7	18.8	39.3	15.1	57.5	22.9	33.4	11.9	23.1	8.7	35.7	13.7	67.3	26.8	34.7	13.7
	Time-Diverse	65.4	24.9	42.1	16.4	42.2	16.3	67.0	24.9	26.4	9.2	23.1	8.1	54.2	20.1	55.7	20.9	40.8	15.5
$\|R\|$=200 $\|S\|$=5	Naïve	30.5	12.0	62.7	23.5	33.9	13.2	43.3	18.0	42.4	15.7	21.0	8.4	19.9	7.9	74.6	29.8	24.4	9.8
	MMR (λ=0.5)	30.5	12.0	62.8	23.6	33.9	13.2	43.3	18.0	42.6	15.6	21.1	8.4	20.0	7.9	74.3	29.6	24.6	9.8
	MMR (λ=0.0)	30.5	12.0	62.8	23.6	33.9	13.2	43.3	18.0	42.6	15.6	21.1	8.4	20.0	7.9	74.3	29.6	24.6	9.8
	Time-Diverse	51.7	20.0	53.2	20.3	43.7	16.8	59.4	23.0	34.8	12.7	27.7	10.4	39.6	15.2	64.6	23.8	37.6	14.5
$\|R\|$=200 $\|S\|$=10	Naïve	48.0	18.4	51.0	18.9	39.2	15.0	57.6	22.9	33.4	12.0	23.1	8.7	35.4	13.6	67.4	26.7	34.4	13.5
	MMR (λ=0.5)	48.5	18.6	50.7	18.8	39.3	15.1	57.5	22.9	33.4	11.9	23.1	8.7	35.7	13.7	67.3	26.8	34.7	13.7
	MMR (λ=0.0)	48.5	18.6	50.7	18.8	39.3	15.1	57.5	22.9	33.4	11.9	23.1	8.7	35.7	13.7	67.3	26.8	34.7	13.7
	Time-Diverse	66.4	24.8	38.2	14.3	39.4	14.8	69.5	25.9	25.2	8.8	24.1	8.7	54.7	20.0	54.2	19.5	41.5	15.3

Results. Are shown for three different categories of history-oriented queries per document collection. All values reported are percentages of the metrics and averaged over all the queries in a group. The results for the New York Times Annotated corpus are presented in Table 1 and for the Living Knowledge corpus can be found in our accompanying technical report [11].

For The New York Times Annotated corpus we can clearly see that our method Time-Diverse outperforms all three baselines by a large margin in recalling most important facts concerning the history-oriented queries. This shows that using retrieval method informed by temporal expressions presents documents that are *retrospectively relevant* for history-oriented queries. The slightly higher precision values for baseline system in all the findings above can be attributed to the fact that most of the baseline summaries tended to be of shorter length than the summaries produced by Time-Diverse method. When increasing the size of $|R|$ we notice that recall also increases for Time-Diverse as compared to the baselines. Since the increase in $|R|$ also implies a increase in the length of the summary; the precision also drops.

There is no clear correlation between a *good summary* and the number of top-k documents $|R|$ considered for generating time intervals of interest; in most cases though it seems increasing the size of pseudo-relevant set generation of time intervals hurts the performance of the diversification algorithm. Considering more documents that are presented to the user $|S|$ increases the performance; indicating that $|S| = 10$ for an optimal value.

Overall, the results show that using our diversification algorithm taking into account temporal expressions gives a better overview for history-oriented queries.

4 Related Work

Diversifying search results using time was first explored in [2]. In their preliminary study the authors limited themselves to using document publications dates, but posed the open problem of diversifying search results using temporal expressions in document contents and the challenging problem of evaluation. Both these aspects have been adequately addressed in our article. More recently, Nguyen and Kanhabua [10] diversify search results based on dynamic latent topics. The authors study how the subtopics for a multi-faceted query change with time. For this they utilize a time-stamped document collection and an external query log. However for the temporal analysis they limit themselves to document publication dates. The recent survey of temporal information retrieval by Campos et al. [4] also highlights the lack of any research that addresses the challenges of utilizing temporal expressions in document contents for search result diversification.

5 Conclusion

In this work, we considered the task of diversifying search results by using temporal expressions in document contents. Our proposed probabilistic framework utilized time intervals of interest derived from the temporal expressions present in pseudo-relevant documents and then subsequently using them as aspects for diversification along time. To evaluate our method we constructed a novel testbed of history-oriented queries derived from authoritative resources and their corresponding *Wikipedia* entries. We showed that our diversification method presents a more complete retrospective set of documents for the given history-oriented query set. This work is largely intended to help scholars in history and humanities to explore large born-digital document collections quickly and find relevant information without knowing time intervals of interest to their queries.

References

1. Agrawal, R., et al.: Diversifying search results. In: WSDM (2009)
2. Berberich, K., Bedathur, S.: Temporal diversification of search results. In: TAIA (2013)
3. Berberich, K., Bedathur, S., Alonso, O., Weikum, G.: A language modeling approach for temporal information needs. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 13–25. Springer, Heidelberg (2010)
4. Campos, R.: Survey of temporal information retrieval, related applications. ACM Comput. Surv. **47**(2), 15:1–15:41 (2014)
5. Carbonell, J.G., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR (1998)
6. Chang, A.X., Manning, C.D: A library for recognizing and normalizing time expressions. In: LREC, SUTIME (2012)
7. Gupta, D., Berberich, K.: Identifying time intervals of interest to queries. In: CIKM (2014)

8. Joho, H., et al.: NTCIR temporalia: A test collection for temporal information access research. In: WWW (2014)
9. Mazur, P.P., Dale, R.: A new corpus for research on temporal expressions. In: EMNLP, Wikiwars (2010)
10. Nguyen, T.N., Kanhabua, N.: Leveraging dynamic query subtopics for time-aware search result diversification. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C.X., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 222–234. Springer, Heidelberg (2014)
11. Gupta, D., Berberich, K.: Diversifying search results using time. Research Report MPI-I–5-001 (2016)
12. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: ACL (2004)