Diversifying Search Results Using Time

Dhruv Gupta
Klaus Berberich

MPI–I–2016–5-001 February 2016

Authors' Addresses

Dhruv Gupta
Max-Planck-Institut für Informatik
Campus E1 4
66123 Saarbrücken
Germany

Klaus Berberich
Max-Planck-Institut für Informatik
Campus E1 4
66123 Saarbrücken
Germany

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 would benefit if returned results covered diverse dates, thus giving an overview of what has happened throughout history. Diversifying search results based on important dates can be a building block for applications, for instance, in *digital humanities*. Historians would thus be able to quickly explore longitudinal document collections by querying for entities or events without knowing associated important dates apriori.

In this work, we describe an approach to diversify search results using temporal expressions (e.g., in the 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 indeed is able to present search results diversified along time.

Keywords

Information Retrieval, Temporal Expressions, Novelty & Diversity

Contents

1 Introduction

"Those who do not remember the past are condemned to repeat it."
—George Santayana

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. In this article 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.

With growing amount of information on the Web, modern retrieval system focus more on recently published documents by using their creation time or publication dates. However, little attention is given to the temporal expressions in document contents which can help us uncover a trove of historical knowledge. This is particularly challenging as temporal expressions like natural language may occur implicitly or explicitly. For example, " during the last three years of his presidency " [23]; the temporal expression last three years is implicit and should be resolved and normalized. Temporal expressions are usually mentioned in a *relative* sense and are highly uncertain in nature. As an example consider, early 1990s; here the exact time interval conveyed by the temporal expression 1990s is not clearly demarcated. Finally, temporal expressions can be present at different granularities of time. Consider the example, " On July 25, 2000, Bush ... at the 2000 Republican National Convention " [23]; here we have 2000 present at year level of granularity while at day level of granularity we have July 25, 2000. Therefore, incorporating statistics about such temporal expressions in a search diversification method can be highly complex.

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 date may not necessarily be the time that the text refers to. It is quite common to have articles that contain a historical perspective of 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 [11], 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 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 from prior available research. As a ground truth we utilize the 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.

Outline. The article is structured as follows. Section 2 presents our adapted probabilistic framework for diversifying search results along time. Section 3 covers in details our evaluation framework. We discuss the results and some anecdotal results in Section 4. Section 5 covers related work in context of temporal information retrieval and text analytics. Finally we end with a brief conclusion and directions for future research in Section 6.

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

2 Method

We now describe the probabilistic framework to diversify search results using temporal expressions mentioned in their contents. The probabilistic frame consists of three key components. The first key component is the representation of time that incorporates temporal uncertainty. The second key component, described elsewhere, is that of generating time intervals of interest for the given keyword query. The final key component is the objective function for determining the maximal subset of documents that covers the time intervals of interest.

2.1 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 $|d_{time}|$.

2.2 Time Model

Several challenges are associated with the statistical analysis of temporal expressions. Identification of implicit and explicit temporal expressions can be done by temporal taggers such as SUTime [7] and HeidelTime [20]. The temporal taggers also allow for normalization and resolution of implicit temporal expressions. The hardest challenge is a representation of time that considers the uncertainty and different levels of granularity in temporal expressions. Berberich et al. [4] addressed this issue by presenting a time model that incorporates temporal uncertainty. For temporal expressions such as 1990s where the begin and end of the interval can not be identified; 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 $|T|$.

2.3 Time Intervals of Interest

In order to present historically diverse documents for a keyword query q_{text}; we need to first identify interesting time intervals. The identified interesting time intervals are then subsequently treated as aspects over which diversification will be performed. Time intervals of interest to the given keyword query q_{text} are identified using an approach outline by Gupta and Berberich [11]. A time interval $[b, e]$ is deemed *interesting* if its referred 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:

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

The first probability gives the likelihood of generating $[b, e]$ from temporal expressions in document d [4]:

$$P(\,[b, e] \mid d_{time}\,) = \frac{1}{|d_{time}|} \sum_{T \in d_{time}} \frac{\mathbb{1}([b, e] \in T)}{|T|}.$$

The second probability is obtained from the query likelihoods $P(\,q_{text} \,|\, d_{text}\,)$:

$$P(\,d_{text} \mid q_{text}\,) = \frac{P(\,q_{text} \mid d_{text}\,)}{\sum_{d'_{text} \in R} P(\,q_{text} \mid d'_{text}\,)}\,.$$

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

2.4 Temporal Diversification

Our aim is to present documents that cover a variety of historical aspects underlying a history-oriented query. To this end we use the identified interesting time intervals as explicit temporal aspects that need to be satisfied

by the documents. The diversified set of documents must thus try to cover these time intervals in proportion to the frequency of their occurrence.

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 $P(\, [b,e] | q_{text}\,)$ 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], which is known to give a $(1 - \frac{1}{e})$ approximation guarantee.

3 Evaluation

We next describe the setup of our experimental evaluation.

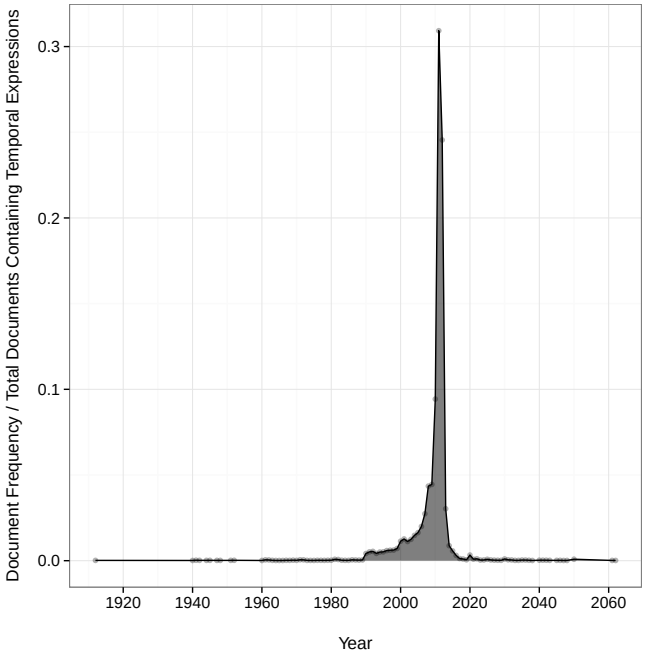
3.1 Document Collections

We used two document collections one from a news archive and one from a web archive. The Living Knowledge [9] corpus is a collection of news and blogs on the Web amounting to approximately 3.8 million documents [14]. The documents are provided with annotations for temporal expressions as well as named-entities.

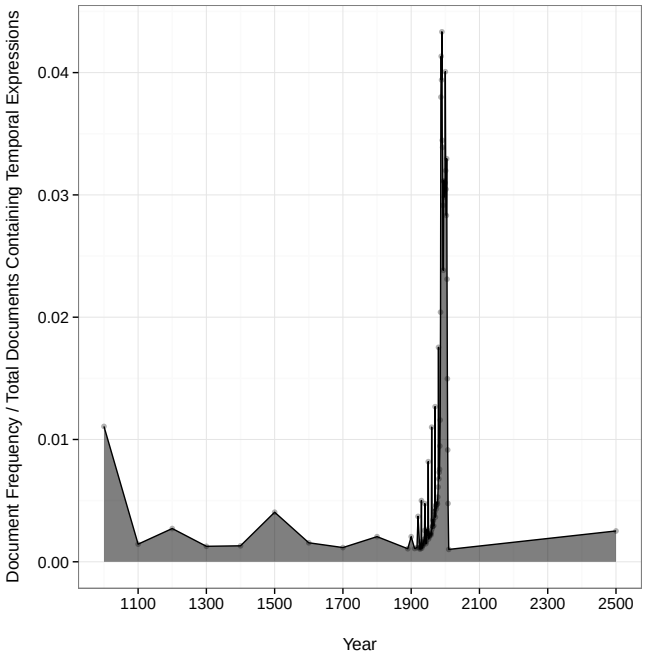
The New York Times (NYT) Annotated [10] corpus is a collection of news articles published in *The New York Times*. It reports articles from 1987 to 2007 and consists of around 2 million news articles. The temporal annotations for it were done via SUTime [7]. Both explicit and implicit temporal expressions were annotated, resolved, and normalized using SUTime.

Temporal Analysis. We did a simple analysis of temporal expressions in both document collections. This involved computing document frequency of temporal expressions at year granularity across the collections. The plots are shown in Figure 3.1.

For Living Knowledge corpus the kurtosis of the distribution of document frequency ordered by time is 421.4 and skewness of 19.9. For the NYT Annotated corpus the kurtosis is 6.8 and skewness is 2.8. This shows that the Living Knowledge corpus has a highly skewed nature of distribution of temporal expressions. Also take note of the time scale on the $x-$axis of both plots (3.1a, 3.1b). The NYT Annotated corpus has a wider temporal coverage in contrast to the Living Knowledge corpus. Given these observations we can conclude that the Living Knowledge corpus is *not* a true longitudinal corpus as the time-period the corpus covers is limited and distorted. These aspects affect any probabilistic analysis performed on the Living Knowledge corpus. The resulting effects show up in our results; which we later discuss.



(a) Living Knowledge corpus.



(b) New York Times Annotated corpus.

Figure 3.1: Distribution of temporal expressions at year granularity by document frequency.

Indexing. The document collections were preprocessed and subsequently indexed using the ElasticSearch software [22]. 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.

3.2 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 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 [11, 16]. We describe the details next.

Queries for long-lasting wars were constructed from the *WikiWars* corpus [16]. The corpus was created for the purpose of temporal information extraction. The keyword for the wars are given in Table 3.1a. For ambiguous important events we utilized the set of ambiguous queries used in our earlier work [11]. The queries used are listed in Table 3.1b. For famous personalities we utilized a list of most influential people available on the USA Today [1] website. The names of these famous personalities were used based on the intuition that there would be important events associated with them at different points of time. The list of all the entities is given in in Table 3.1c.

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* [2] pages of the queries as ground truth summaries.

The entire testbed of history-oriented queries along with their corresponding *Wikipedia* articles is made publicly available at the following url:

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

[1] `http://usatoday30.usatoday.com/news/top25-influential.htm`
[2] `https://en.wikipedia.org/`

Americas | american civil war | american revolution | mexican revolution
Europe | world war II | world war I | french revolution | punic wars | spanish civil war | russo-polish war | second italo abyssinian war
Africa | french algreian war | biafran nigerian civil war
Asia | vietnam war | korean war | iraq war | persian wars | chinese civil war | iran iraq war | russian civil war | french indochina war | russo-japanese car

(a) Wars.

Sports | commonwealth games | asian games | summer olympics | winter olympics | super bowl winners
Music | u2 album | nirvana album | beatles album | red hot chilli peppers album | michael jackson album
Movies | harry potter movie | oscar academy awards | lord of the ring movie
Politics | german federal elections | us presidential elections | australia federal elections

(b) Events.

Business | bill gates | sergey brin | larry page | howard schultz | sam walton
Science | stephen hawking | francis collins | craig venter
Politics | ronald reagan | mikhail gorbachev | george w. bush | deng xiaoping | nelson mandela | bill clinton | hillary clinton
Arts | j. k. rowling | oprah winfrey | russell simmons | bono
Religion | pope john paul II
Sports | lance armstrong | michael jordan
Other | ryan white | homer simpson | osama bin laden

(c) Entities.

Table 3.1: History-oriented queries.

3.3 Systems and Metrics

Baselines. We considered three baselines, in order of increasing sophistication. As a naïve baseline, we first consider the pseudo-relevant documents retrieved for the given keyword query. The next two baselines utilize a well known implicit diversification algorithm *maximum marginal relevance* (MMR) [6]. Formally it is defined as:

$$\underset{d \notin S}{\operatorname{argmax}} \left(\lambda \cdot \operatorname{sim}_1 (q, d) - (1 - \lambda) \cdot \max_{d' \in S} \operatorname{sim}_2 (d', d) \right).$$

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. First one is the number of documents considered for generating time intervals of interest $|R|$. The second parameter is 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 [26] (implemented in [15]) 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 in the candidate summary S and the reference summary G with respect to the reference summary G:

$$recall = \frac{\sum_{g \in ngram(G)} \sum_{s \in ngram(S)} \mathbb{1}(s = g)}{\mid ngram(G) \mid},$$

where, $ngram(\cdot)$ returns n-grams for piece of text, $\mid ngrams(\cdot) \mid$ gives the total number n-grams, and $\mathbb{1}(\cdot)$ is an indicator function that tests the equivalence of n-grams. Note that n in $ngram$ is the length of the gram to be considered; we limit ourselves to $n \in \{1, 3\}$. The precision is calculated in a similar manner although with respect to the candidate summary S. To combine both measures, F_β is used:

$$F_\beta = \frac{(1 + \beta^2) \cdot recall \cdot precision}{recall + \beta^2 \cdot precision}$$

4 Results

Results are shown for three different categories of history-oriented queries per document collection. For each category of history-oriented query we show *recall*, *precision*, and $F_{\beta=1.0}$ scores for ROUGE-1 and ROUGE-3 metrics. $F_{\beta=1.0}$ is a balanced metric that gives equal weight to both precision and recall. All values are 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 Tables 4.1 and for the Living Knowledge corpus are shown in Table 4.2.

Given a history-oriented query, an ideal list of documents should either give a timeline overview of the event or portray the biography of the entity. Therefore all the documents that the system presents must *recall* as many facts as possible when compared to a ground truth summary. The *precision* as it is computed with respected to the system generated summary may vary as we increase the number of pseudo-relevant documents. Regardless of this we present the $F_{\beta=1.0}$ scores that give equal weight to both precision and recall.

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.

For the Living Knowledge corpus we see that our method performs better than the baselines when considering the $F_{\beta=1.0}$ scores. In recall it performs at

	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

Table 4.1: 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	14.5	1.7	62.0	16.8	18.4	2.3	21.5	6.4	53.2	16.4	22.4	6.5	4.6	1.2	80.2	37.0	7.6	2.0
	MMR (λ=0.5)	14.5	1.7	62.0	16.6	18.4	2.4	21.3	6.4	52.6	15.9	22.3	6.6	4.6	1.2	79.0	34.6	7.7	2.0
	MMR (λ=0.0)	28.3	1.8	49.7	6.8	23.4	1.7	37.5	6.7	38.5	5.2	24.9	3.6	14.1	1.5	66.5	9.3	20.2	2.1
	TIME-DIVERSE	26.1	1.9	52.1	6.1	24.9	1.9	36.7	6.9	40.7	5.8	26.9	4.1	15.0	1.5	68.8	12.7	19.9	2.0
$\|R\|$=100 $\|S\|$=10	NAÏVE	24.8	2.9	55.0	11.3	25.7	3.3	38.2	11.2	42.9	11.5	27.7	7.0	9.2	2.0	77.0	32.1	13.1	3.1
	MMR (λ=0.5)	25.4	2.9	54.7	10.8	26.4	3.3	38.4	11.2	42.6	10.2	28.0	7.0	9.5	2.0	75.4	28.6	13.5	3.1
	MMR (λ=0.0)	40.7	3.3	40.8	4.2	29.4	2.5	50.3	11.3	32.7	4.4	26.5	4.1	22.4	2.4	62.5	8.5	27.1	2.9
	TIME-DIVERSE	39.2	3.4	43.0	4.8	30.6	2.7	50.6	11.7	34.1	4.8	29.4	4.7	22.5	2.5	64.6	10.5	25.8	3.0
$\|R\|$=150 $\|S\|$=5	NAÏVE	14.8	1.7	60.7	14.3	18.8	2.4	21.5	6.4	53.2	16.4	22.4	6.5	4.3	1.1	81.4	39.2	7.2	2.0
	MMR (λ=0.5)	14.8	1.7	60.6	14.2	18.8	2.4	21.3	6.4	52.6	15.9	22.3	6.6	4.4	1.1	80.4	36.9	7.4	2.0
	MMR (λ=0.0)	31.5	2.3	46.5	4.5	26.2	1.9	40.3	7.1	37.3	4.7	24.0	3.3	15.1	1.5	63.2	9.1	19.7	2.1
	TIME-DIVERSE	28.6	2.1	49.9	4.5	27.5	2.1	36.7	6.8	39.2	5.8	25.6	4.1	16.4	1.5	66.5	9.0	20.7	2.0
$\|R\|$=150 $\|S\|$=10	NAÏVE	24.8	2.9	55.0	11.3	25.7	3.3	38.2	11.2	42.9	11.5	27.7	7.0	9.2	2.0	77.0	32.1	13.1	3.1
	MMR (λ=0.5)	25.4	3.0	54.8	10.8	26.5	3.3	38.8	11.2	42.4	10.2	28.1	7.0	9.5	2.0	75.3	28.6	13.5	3.1
	MMR (λ=0.0)	42.9	3.8	39.1	3.8	30.5	2.6	53.5	11.8	31.4	3.9	26.6	3.8	25.4	2.6	58.9	7.3	28.1	2.9
	TIME-DIVERSE	41.5	3.6	41.2	3.7	32.3	2.7	50.0	11.7	33.2	4.8	28.1	4.7	25.1	2.6	61.3	7.8	28.0	2.9
$\|R\|$=200 $\|S\|$=5	NAÏVE	14.8	1.7	60.7	14.3	18.8	2.4	21.5	6.4	53.2	16.4	22.4	6.5	4.3	1.1	81.4	39.2	7.2	2.0
	MMR (λ=0.5)	14.8	1.7	60.4	13.9	18.9	2.4	21.3	6.4	52.6	15.9	22.3	6.6	4.4	1.1	80.4	36.9	7.4	2.0
	MMR (λ=0.0)	31.5	2.3	45.3	4.1	26.1	1.9	38.5	6.8	36.9	4.9	23.3	3.2	18.2	1.5	61.3	7.3	21.7	1.8
	TIME-DIVERSE	30.2	2.1	48.8	4.3	27.8	2.0	38.3	7.0	41.0	5.5	27.4	4.1	15.0	1.4	68.4	8.9	20.7	2.0
$\|R\|$=200 $\|S\|$=10	NAÏVE	24.8	2.9	55.0	11.3	25.7	3.3	38.2	11.2	42.9	11.5	27.7	7.0	9.2	2.0	77.0	32.1	13.1	3.1
	MMR (λ=0.5)	25.4	3.0	54.7	10.7	26.5	3.3	38.8	11.2	42.4	10.2	28.1	7.0	9.5	2.0	75.3	28.6	13.5	3.1
	MMR (λ=0.0)	43.0	3.7	37.9	3.3	30.5	2.4	54.0	11.7	30.7	3.8	27.0	3.7	27.9	2.5	56.8	6.2	28.7	2.6
	TIME-DIVERSE	42.3	3.6	40.1	3.6	31.0	2.6	53.2	12.1	33.3	4.5	29.4	4.4	25.4	2.6	61.6	7.1	29.5	3.0

Table 4.2: Results for the Living Knowledge corpus.

par with MMR ($\lambda = 0.0$). While looking at the precision of the summaries our method TIME-DIVERSE constantly outperforms the baselines; considering that the length of the summaries in this case tend to be uniform for all methods. As discussed in Section 3.1, Living Knowledge presents us with the challenge of having a skewed and biased distribution of temporal expressions. Even taking into account this factor; our method shows an overall improvement over the baselines. This again empirically shows that our method is highly sensitive to temporal expressions and aids in *temporal diversification*.

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 number of documents that are presented to the user $|S|$ increases the performance; indicating that $|S| = 10$ for an optimal value.

Overall, the results conclusively show that using our diversification algorithm taking into account temporal expressions gives us a better retrospective overview of a history-oriented query.

4.1 History by Algorithms

Here we present anecdotal results for two example history-oriented queries. Queries considered were george w. bush and economic depression. The results shown in Figures 4.1 and 4.2 are from the New York Times Annotated corpus. Individual results are shown with their article headline and their contained temporal expressions. In addition we show the time intervals identified and used for diversification.

For the query george w. bush the identified time intervals include the time intervals [2000, 2000] and [2000, 2004] marking the year of his first election and his first presidential term. This is covered by documents $D_{1117027}$, $D_{1461580}$, & $D_{1255229}$ returned by our diversification method. The second term of his presidency is marked by the category [2004, 2004] and [2004, 2007] (corpus covers the time period 1987-2007) described by the document $D_{1610342}$. The last temporal category is [1992, 1992] in which George W. Bush worked as campaign advisor for his father's presidential campaign. The 1992 presidential campaign of his father is given in D_{537116}. This is a more temporally diverse set of documents as compared against the baseline set of documents centered around 2000.

For the query economic depression identified time intervals explicitly cover the various periods when a downturn occurred. The periods [1990, 1991]

Query: george w. bush	
Identified Time Intervals: [2000, 2000] ; [2000, 2004] ; [2004, 2004] ; [2004, 2007] ; [1992, 1992]	
NAÏVE	TIME-DIVERSE
$D_{1181696}$ – Heir Apparent? – 2000	$D_{1117027}$ – Ideas & Trends; Republicans Stalk a Slogan, Hunting for Themselves – 1999; 1992; 1996
$D_{1142543}$ – Rival Biographies of Bush Are Rushing to Print – 1999; 1992	$D_{1461580}$ – The World: A Calling to Heal; Getting Religion on AIDS – 2003; 1999; 1986; 1991; 1995
$D_{1242996}$ – THE 2000 CAMPAIGN: THE CANDIDATES; In Final Days, Rallying Supporters and Attempting to Sidestep a Volatile Issue – 2000; 1996	$D_{1610342}$ – THE 2004 CAMPAIGN: THE DEMOCRATIC NOMINEE; Kerry Invokes the Bible In Appeal for Black Votes – 2004; 2000
$D_{1609737}$ – A Ketchup Too Spicy for the G.O.P. – 2004	$D_{1255229}$ – THE 43rd PRESIDENT: THE MOOD IN TEXAS; Victory Celebration Is Tempered by Bush's Need to Focus on Reconciliation – 2000
$D_{1255563}$ – The George Bush I Knew – 2000; mid-1960's	D_{537116} – THE MEDIA BUSINESS: ADVERTISING; Bringing Madison Avenue Polish to Bush's Campaign Ads – 1992; 1988; 1974

Figure 4.1: Top-5 results for george w. bush

Query:economic depression	
Identified Time Intervals: [1990, 1990] ; [1990, 1995] ; [1987, 1987] ; [1930, 1930] ; [1998, 1998]	
NAÏVE	TIME-DIVERSE
D_{175692} – Economic Scene; Forecasters' Art In 1929 and Now – 1988; 1920's ;1929 ; 1930-31 ; 1929 ; mid-1931 ; 1929-30	D_{491246} – U.N. Report Warns of Crisis in Eastern Europe – 1988; 1929 to 1933 ; 1992 ; 1992
D_{653581} – Costs of Depression Are on a Par With Heart Disease, a Study Says – 1993; 1990	D_{113808} – WASHINGTON TALK: BRIEFING; 'The Great Correction' – 1988
D_{317417} – The Reagan Boom - Greatest Ever – 1990; 1930's;1980's;1989; 1982; 1990's;1960's;1990; 1970 to 1982; 1982; 1983; 1980's	D_{741011} – Let's Look at Biases in the History Standards; Skewed Economics – 1995; 1920's ;1995
D_{741011} – Let's Look at Biases in the History Standards; Skewed Economics – 1995; 1920's ;1995	D_{88390} – THE ECONOMIC HISTORIANS' VIEW: Comparing the Collapses; C.P. Kindleberger: Watch Prayerfully – 1987; 1929; 18th and 19th century; 1836; 1857; 1873; 1907; 1929; 1987; 1920
D_{327066} – Economic Scene; High Hopes And Deep Fears – 1990;1991	$D_{1349556}$ – Don't Blame Wall Street – 2001; 1929; 1920's;early 20's;90's;1921; 1991; 1920-21; 1921; 1913; 1927; 1912; 1920 to 1929; 1931; 1929; 1955; 1969; 1998; 1989; 1926; 1990's; 1939

Figure 4.2: Top-5 results for economic depression

& [1990, 1995] represent the *early 90's depression* covered by documents D_{491246}, D_{113808}, & D_{741011} ; [1987, 1987] covers the *Black Monday* when stock markets crashed, the story is reported in D_{88390} and finally [1930,1930] marks the year that begins the *Great Depression* covered in document $D_{1349556}$. Clearly, this is a more temporally diverse and interesting set of documents output by our approach as compared to the baseline where documents focus more on the slump in markets in the 1990s.

5 Related Work

Our research provides a bridging gap between two very important research themes: temporal information retrieval and temporal text analytics. Temporal IR centric methods have largely avoided leveraging complex temporal expressions in favor of document publication dates. Temporal text analytical methods on the other hand have largely relied on these temporal expressions for mining time-sensitive facts. We discuss some of these works next.

Temporal Information Retrieval. Diversifying search results using time was explored in [3]. In their preliminary study the authors limited themselves to using document publications dates. However they 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 [18] 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. Also a recent survey of temporal information retrieval by Campos et al. [5] also highlights the lack of any research that address the challenges of utilizing temporal expressions in document contents for search result diversification along time.

Temporal Text Analytics. Using text analytics and temporal expressions, Yeung and Jatowt [25] study how the past is remembered. They study varying trends of topics identified via *latent Dirichlet allocation* and time in a text collection. Special emphasis has been laid on predictions of future events. In the seminal work by Baeza-Yates [2]; the author explores how to model a future retrieval (FR) system that would take in to account temporal expressions in a document body. More recently, Jatowt and Yeung [13] explore this more research direction by proposing a model-based clustering algorithm for extracting representative summaries for future events from the Google news archive.

6 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.

Future Work. In this article we focused on how to incorporate temporal expression to diversify search results. Following problems still remain open:

- How can we utilize the context around the temporal expressions for presenting historical information at a finer level of textual granularity?
- How about considering other semantic annotations in form of named entities and geographical locations ? How do we model these additional annotations in an information retrieval method efficiently and effectively?

Bibliography

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009*, pages 5–14, 2009.

[2] R. Baeza-Yates. Searching the future. In *SIGIR Workshop MF/IR*, 2005.

[3] K. Berberich and S. Bedathur. Temporal Diversification of Search Results. In F. Diaz, S. Dumais, K. Radinsky, M. de Rijke, and M. Shokouhi, editors, *SIGIR 2013 Workshop on Time-aware Information Access*. Microsoft Research, 2013.

[4] K. Berberich, S. J. Bedathur, O. Alonso, and G. Weikum. A language modeling approach for temporal information needs. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, pages 13–25, 2010.

[5] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt. Survey of temporal information retrieval and related applications. *ACM Comput. Surv.*, 47(2):15:1–15:41, 2014.

[6] J. G. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 335–336, 1998.

[7] A. X. Chang and C. D. Manning. Sutime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 3735–3740, 2012.

[8] W. Dakka, L. Gravano, and P. G. Ipeirotis. Answering general time-sensitive queries. *IEEE Trans. Knowl. Data Eng.*, 24(2):220–235, 2012.

[9] Living knowledge corpus, 2015. [Online; accessed 23-September-2015]`http://livingknowledge.europarchive.org/`.

[10] The new york times annotated corpus, 2015. [Online; accessed 23-September-2015]`https://catalog.ldc.upenn.edu/LDC2008T19`.

[11] D. Gupta and K. Berberich. Identifying time intervals of interest to queries. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 1835–1838, 2014.

[12] D. Gupta and K. Berberich. Temporal query classification at different granularities. In *String Processing and Information Retrieval - 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings*, pages 156–164, 2015.

[13] A. Jatowt and C. A. Yeung. Extracting collective expectations about the future from large text collections. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1259–1264, 2011.

[14] H. Joho, A. Jatowt, and R. Blanco. Ntcir temporalia: A test collection for temporal information access research. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 845–850, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[15] C. Y. Lin. Rouge: Recall-oriented understudy of gisting evaluation. a software package for automated evaluation of summaries, 2015. [Online; accessed 23-September-2015]`http://www.berouge.com/Pages/default.aspx`.

[16] P. P. Mazur and R. Dale. Wikiwars: A new corpus for research on temporal expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 913–922, 2010.

[17] D. Metzler, R. Jones, F. Peng, and R. Zhang. Improving search relevance for implicitly temporal queries. In *Proceedings of the 32Nd International*

ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, pages 700–701, New York, NY, USA, 2009. ACM.

[18] T. N. Nguyen and N. Kanhabua. Leveraging dynamic query subtopics for time-aware search result diversification. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 222–234, 2014.

[19] M.-H. Peetz, E. Meij, and M. Rijke. Using temporal bursts for query modeling. *Inf. Retr.*, 17(1):74–108, Feb. 2014.

[20] J. Strötgen and M. Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013.

[21] Top-25 influential people, 2015. [Online; accessed 23-September-2015]`http://usatoday30.usatoday.com/news/top25-influential.htm`.

[22] Elastic — revealing insights from data (formerly elasticsearch), 2015. [Online; accessed 23-September-2015]`https://www.elastic.co/`.

[23] George w. bush — Wikipedia, the free encyclopedia, 2015. [Online; accessed 23-September-2015]`https://en.wikipedia.org/wiki/Ronald_Reagan`.

[24] Wikipedia, the free encyclopedia, 2015. [Online; accessed 23-September-2015]`http://en.wikipedia.org/`.

[25] C. A. Yeung and A. Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1231–1240, 2011.

[26] C. Y. Lin. Rouge: a package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop* pages 25–26, 2004.

Below you find a list of the most recent research reports of the Max-Planck-Institut für Informatik. Most of them are accessible via WWW using the URL `http://www.mpi-inf.mpg.de/reports`. Paper copies (which are not necessarily free of charge) can be ordered either by regular mail or by e-mail at the address below.

Max-Planck-Institut für Informatik
– Library and Publications –
Campus E 1 4

D-66123 Saarbrücken

E-mail: `library@mpi-inf.mpg.de`

MPI-I-2014-5-002	A. Anand, I. Mele, S. Bedathur, K. Berberich	Phrase Query Optimization on Inverted Indexes
MPI-I-2014-5-001	M. Dylla, M. Theobald	Learning Tuple Probabilities in Probabilistic Databases
MPI-I-2014-4-002	S. Sridhar, A. Oulasvirta, C. Theobalt	Fast Tracking of Hand and Finger Articulations Using a Single Depth Camera
MPI-I-2014-4-001	K.I. Kim, J. Tompkin, C. Theobalt	Local high-order regularization on data manifolds
MPI-I-2013-RG1-002	P. Baumgartner, U. Waldmann	Hierarchic superposition with weak abstraction
MPI-I-2013-5-002	F. Makari, R. Gemulla, R. Khandekar, J. Mestre, M. Sozio	A distributed algorithm for large-scale generalized matching
MPI-I-2013-1-001	S. Ott	New results for non-preemptive speed scaling
MPI-I-2012-RG1-002	A. Fietzke, E. Kruglov, C. Weidenbach	Automatic generation of inductive invariants by SUP(LA)
MPI-I-2012-RG1-001	M. Suda, C. Weidenbach	Labelled superposition for PLTL
MPI-I-2012-5-004	F. Alvanaki, S. Michel, A. Stupar	Building and maintaining halls of fame over a database
MPI-I-2012-5-003	K. Berberich, S. Bedathur	Computing n-gram statistics in MapReduce
MPI-I-2012-5-002	M. Dylla, I. Miliaraki, M. Theobald	Top-k query processing in probabilistic databases with non-materialized views
MPI-I-2012-5-001	P. Miettinen, J. Vreeken	MDL4BMF: Minimum Description Length for Boolean Matrix Factorization
MPI-I-2012-4-001	J. Kerber, M. Bokeloh, M. Wand, H. Seidel	Symmetry detection in large scale city scans
MPI-I-2011-RG1-002	T. Lu, S. Merz, C. Weidenbach	Towards verification of the pastry protocol using TLA+
MPI-I-2011-5-002	B. Taneva, M. Kacimi, G. Weikum	Finding images of rare and ambiguous entities
MPI-I-2011-5-001	A. Anand, S. Bedathur, K. Berberich, R. Schenkel	Temporal index sharding for space-time efficiency in archive search
MPI-I-2011-4-005	A. Berner, O. Burghard, M. Wand, N.J. Mitra, R. Klein, H. Seidel	A morphable part model for shape manipulation
MPI-I-2011-4-003	J. Tompkin, K.I. Kim, J. Kautz, C. Theobalt	Videoscapes: exploring unstructured video collections
MPI-I-2011-4-002	K.I. Kim, Y. Kwon, J.H. Kim, C. Theobalt	Efficient learning-based image enhancement : application to compression artifact removal and super-resolution
MPI-I-2011-4-001	M. Granados, J. Tompkin, K. In Kim, O. Grau, J. Kautz, C. Theobalt	How not to be seen inpainting dynamic objects in crowded scenes
MPI-I-2010-RG1-001	M. Suda, C. Weidenbach, P. Wischnewski	On the saturation of YAGO
MPI-I-2010-5-008	S. Elbassuoni, M. Ramanath, G. Weikum	Query relaxation for entity-relationship search
MPI-I-2010-5-007	J. Hoffart, F.M. Suchanek, K. Berberich, G. Weikum	YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia
MPI-I-2010-5-006	A. Broschart, R. Schenkel	Real-time text queries with tunable term pair indexes
MPI-I-2010-5-005	S. Seufert, S. Bedathur, J. Mestre, G. Weikum	Bonsai: Growing Interesting Small Trees

MPI-I-2010-5-004	N. Preda, F. Suchanek, W. Yuan, G. Weikum	Query evaluation with asymmetric web services
MPI-I-2010-5-003	A. Anand, S. Bedathur, K. Berberich, R. Schenkel	Efficient temporal keyword queries over versioned text
MPI-I-2010-5-002	M. Theobald, M. Sozio, F. Suchanek, N. Nakashole	URDF: Efficient Reasoning in Uncertain RDF Knowledge Bases with Soft and Hard Rules
MPI-I-2010-5-001	K. Berberich, S. Bedathur, O. Alonso, G. Weikum	A language modeling approach for temporal information needs
MPI-I-2010-1-001	C. Huang, T. Kavitha	Maximum cardinality popular matchings in strict two-sided preference lists
MPI-I-2009-RG1-005	M. Horbach, C. Weidenbach	Superposition for fixed domains
MPI-I-2009-RG1-004	M. Horbach, C. Weidenbach	Decidability results for saturation-based model building
MPI-I-2009-RG1-002	P. Wischnewski, C. Weidenbach	Contextual rewriting
MPI-I-2009-RG1-001	M. Horbach, C. Weidenbach	Deciding the inductive validity of $\forall\exists^*$ queries
MPI-I-2009-5-007	G. Kasneci, G. Weikum, S. Elbassuoni	MING: Mining Informative Entity-Relationship Subgraphs
MPI-I-2009-5-006	S. Bedathur, K. Berberich, J. Dittrich, N. Mamoulis, G. Weikum	Scalable phrase mining for ad-hoc text analytics
MPI-I-2009-5-005	G. de Melo, G. Weikum	Towards a Universal Wordnet by learning from combined evidenc
MPI-I-2009-5-004	N. Preda, F.M. Suchanek, G. Kasneci, T. Neumann, G. Weikum	Coupling knowledge bases and web services for active knowledge
MPI-I-2009-5-003	T. Neumann, G. Weikum	The RDF-3X engine for scalable management of RDF data
MPI-I-2009-5-003	T. Neumann, G. Weikum	The RDF-3X engine for scalable management of RDF data
MPI-I-2009-5-002	M. Ramanath, K.S. Kumar, G. Ifrim	Generating concise and readable summaries of XML documents
MPI-I-2009-4-006	C. Stoll	Optical reconstruction of detailed animatable human body models
MPI-I-2009-4-005	A. Berner, M. Bokeloh, M. Wand, A. Schilling, H. Seidel	Generalized intrinsic symmetry detection
MPI-I-2009-4-004	V. Havran, J. Zajac, J. Drahokoupil, H. Seidel	MPI Informatics building model as data for your research
MPI-I-2009-4-003	M. Fuchs, T. Chen, O. Wang, R. Raskar, H.P.A. Lensch, H. Seidel	A shaped temporal filter camera
MPI-I-2009-4-002	A. Tevs, M. Wand, I. Ihrke, H. Seidel	A Bayesian approach to manifold topology reconstruction
MPI-I-2009-4-001	M.B. Hullin, B. Ajdin, J. Hanika, H. Seidel, J. Kautz, H.P.A. Lensch	Acquisition and analysis of bispectral bidirectional reflectance distribution functions
MPI-I-2008-RG1-001	A. Fietzke, C. Weidenbach	Labelled splitting
MPI-I-2008-5-004	F. Suchanek, M. Sozio, G. Weikum	SOFIE: a self-organizing framework for information extraction
MPI-I-2008-5-003	G. de Melo, F.M. Suchanek, A. Pease	Integrating Yago into the suggested upper merged ontology
MPI-I-2008-5-002	T. Neumann, G. Moerkotte	Single phase construction of optimal DAG-structured QEPs
MPI-I-2008-5-001	G. Kasneci, M. Ramanath, M. Sozio, F.M. Suchanek, G. Weikum	STAR: Steiner tree approximation in relationship-graphs
MPI-I-2008-4-003	T. Schultz, H. Theisel, H. Seidel	Crease surfaces: from theory to extraction and application to diffusion tensor MRI
MPI-I-2008-4-002	D. Wang, A. Belyaev, W. Saleem, H. Seidel	Shape complexity from image similarity