History by Diversity: Helping Historians search News Archives

Jaspreet Singh, Wolfgang Nejdl, Avishek Anand
L3S Research Center
Appelstr. 9a
30167 Hanover, Germany
{singh,nejdl,anand}@L3S.de

ABSTRACT

Longitudinal corpora like newspaper archives are of immense value to historical research, and time as an important factor for historians strongly influences their search behaviour in these archives. While searching for articles published over time, a key preference is to retrieve documents which cover the important aspects from important points in time which is different from standard search behavior. To support this search strategy, we introduce the notion of a *Historical Query Intent* to explicitly model a historian's search task and define an aspect-time diversification problem over news archives.

We present a novel algorithm, HistDiv, that explicitly models the aspects and important time windows based on a historian's information seeking behavior. By incorporating temporal priors based on publication times and temporal expressions, we diversify both on the aspect and temporal dimensions. We test our methods by constructing a test collection based on *The New York Times Collection* with a workload of 30 queries of historical intent assessed manually. We find that HistDiv outperforms all competitors in subtopic recall with a slight loss in precision. We also present results of a qualitative user study to determine wether this drop in precision is detrimental to user experience. Our results show that users still preferred HistDiv's ranking.

1. INTRODUCTION

Newspaper articles encode history as it happens by capturing events and their immediate impact on society, politics, business and other important spheres. These are of immense value to historians, sociologists, and journalists who rely on fairly reliable, accurate and time-aligned information sources. Specifically for historians, whose desired corpus of study is an archive, browsing and searching such archives has emerged as an important aspect in their research [42]. Consequently, designing access methods and retrieval models tailored to their search patterns and information need is an important problem.

The information seeking behavior of a historian is slightly different from the traditional user search behavior, for which classical retrieval tasks are designed, in two respects. First, historians are interested in obtaining an overview of the topic they wish to research in order to contextualize results. They desire to look at relevant results from *important subtopics* from the most *relevant time points of interest*. Currently, this is realized by issuing an underspecified broad query on the topic and then trying to identify relevant articles from important subtopics by applying various filters which are time, source, region or domain-based. Secondly, the major preoccupation of historians is in finding *primary sources* of information (accounts/reports/documents made by an observer of an event). Secondary information sources (accounts made in retrospect) are also important, however they are intrinsically used to identify primary information sources.

Consider a historian interested in Rudolph Giulaini, a U.S. Republican politician, in the period between 1987 and 2007. Giuliani started out as an attorney and rose to prominence to challenge for the mayorality of New York City in 1989. Though he lost that year to David Dinkins he went on to win in 1993, again in 1997 and stayed mayor until 2001. He was known for his tough stance on crime, his efforts after 9/11 and is responsible for many forward reforms in the city. In 2000, he ran for senate against Hilary Clinton before being diagnosed with cancer. At the same time he was involved in an extramarital affair with Judith Nathan. He then decided to run for president in 2007.

Identifying the New York Times newspaper archive[2] as the best source of primary material, the historian formulates her intent of finding information related to Giuliani with the keywords `rudolph giuliani` and sets the publication date filter to 1987 – 2007 only to get the following results:

Rank	Year	Headline
1	2007	In His Own Words
2	2007	Giuliani Is Expected to Sell One of His Three Businesses
3	2007	Giuliani Is Selling Investment Firm
4	2007	'08 Candidacy Could Shake Up Giuliani's Firm
5	2006	Giuliani Building Network of Donors, a Backer Says

None of the result documents mention his stance on crime, cancer, Hilary Clinton, World Trade Center, David Dinkins or his reforms. Arguably a better set of results for her to scope out her topic is a diversified set of top documents covering important aspects which could be entities like Hilary Clinton and David Dinkins but also from important time points so that she can contextualize or reformulate her query.

Learning from the experiences of our colleagues at the British Library and the Institute of Historical Research in London (cf. Section 2), in this work, we propose a novel document retrieval task which intends to present the most relevant results from a topic-temporal space. This problem can be seen as a generalization of the classical diversity problem by adding a temporal dimension. The topical diversity focuses on presenting results from different subtopics while the temporal diversity ensures that the documents returned are primary in nature. However, the challenge in adapting existing diversity-based approaches are the following. Firstly, tra-

ditional diversity approaches [14, 39, 3, 15, 23] tend to diversify typically on the topical aspect and do not take time into account. As a result, documents retrieved might still cover a good number of aspects but (a) might be from the same time period disregarding the temporal salience of the aspect and (b) might return secondary sources even when more relevant primary sources are present. For the query `rudolph giuliani`, such methods would not guarantee that documents are from the important time periods like '89, '93, '00, '01 and '07.

Time-aware approaches which take into account latent topics or aspects like [36] are optimized to present results which are valid at querying time or in other words reward recency. On the other hand [7] diversifies based on time without explicitly considering topical aspects. Although this ensures that results are temporally distant from each other, as a consequence of the inherent topic-agnostic nature they still might belong to similar aspects. In our example, if the historian uses a temporal diversification retrieval model there is no guarantee that results returned from 2000 and 2001 will definitely cover the WTC, cancer, Hilary Clinton and Judith Nathan.

Finally, multi-dimensional approaches to diversity treat both time and aspects similarly which is not always desirable since both these dimensions have different semantics. We on the contrary, explicitly model both the topical and temporal aspects of a document by treating time as a first-class citizen in our model. In our approach, called HistDiv, the temporal space is based on primary (publication times) and secondary sources (temporal references in text) and the topic space is based on the entities present in the news article. We then jointly diversify both in the aspect and time dimensions discounting each of these dimensions based on semantics unique to each. In sum we make the following contributions:

- We introduce the notion of *Historical Query Intents* and model this as a search result diversification task on both the aspect and time dimensions for historical search.
- We develop a novel retrieval algorithm called HistDiv which jointly diversifies both dimensions by appropriately discounting the contribution of aspects and time.
- We establish the effectiveness of our method by building a test collection based on the 20 years of the *New York Times Collection* as a dataset and a workload of 30 manually judged queries which will be made available to the community. The quantitative results show we outperform our competitors in subtopic recall at the cost of precision.
- Finally, we conducted a qualitative study with the target users to confirm if this loss in precision truly harms the quality of the overview derived from the ranking.

2. HOW DO HISTORIANS SEARCH?

A historian's corpus of study is often an archive. Archives consist of time annotated records, from the distant and recent past, categorized as primary and secondary sources.

Need for Overview and Context: A vital step in a historians search process is to browse the archive in order to get an *overview* of material available on a topic. This allows them to identify potential areas of interest which subsequently lead to more focused queries in the next search phase. Undoubtedly, the onset of digital archives has greatly improved information access but a key requirement of historians still is to obtain an overview of multiple aspects of the topic [41]. This overview allows them to not only find more specific topics of study but also to contextualize their results by studying the temporal and spatial vicinity of the topic.

Focus on primary sources: The main preoccupation of historians is the reading of primary sources while secondary sources are useful when no primary evidence exists and for tracing references to cited primary sources [16]. Even with the onset of digital archives, [41] re-enforces the historian's preference and focus on primary sources.

Historian Search Behavior: Based on existing literature and a short survey in collaboration with the British Library (BL) and the Institute of Historical Research in London, we found that the current way of searching digital archives is a two-stage iterative querying process. In the first stage, keyword queries on broader topics are issued and further reformulated by the use of a combination of filters (time, source, entity, etc.) and facets to gain an overview of the results. Subsequently, more specific queries on each aspect are prepared to serve their information need. [10] finds keyword search (labeled a *blunt instrument*) in archives to be most effective when the user is very precise and focused in his search. Hence, the initial stage is indeed the most cumbersome, firstly because it leads to a large number of results which then have to be read in its entirety, and secondly because of the increased usage of filters. Examples of such general queries are queries about entities, recurring events, conflicts etc.

In general, historians working with physical archives are privy to themed sub-collections created by archivists. An entity search in these smaller collections may produce plenty of results but nearly all tend to be relevant. On the other hand, a news archive is a single large collection only subdivided based on time, hence searching for entities, especially popular ones, can return many records.

As we will show in the remainder of this paper, the joint diversification of aspects and time can lead to better ranking of documents for the initial search phase, involving these broad keyword queries, in newspaper archives.

3. HISTORICAL SEARCH TASK

A *Historical Query Intent* is the moniker we choose to describe a user's intent to cover as many historically relevant subtopics and time windows for a given topic. According to [32], a temporal query intent is used specify queries which are either atemporal, temporally ambiguous or unambiguous. We however deal with a special case of temporally ambiguous queries which have an explicit information need for the past. Additionally, historical query intents also deal with ambiguity with respect to the aspect and time dimensions. In this section we formally present our input model and define our historical search task problem based on historical query intents, and discuss how to measure the effectiveness of proposed approaches.

3.1 Model

Document Model: We operate on a document collection $\mathcal{D}$ where each document $d_p \in \mathcal{D}$ has a publication time point of p. The content of d_p, for instance "Dinkins pulls negative ad about Giuliani as the race for Mayor draws closer", can be represented by a set of aspects such as {David Dinkins, Rudolph Giuliani}. The set of aspects describing the content of a document d_p is denoted by the set $A(d_p) = \{a_1, \ldots a_n\}$ where a_i is a single aspect like David Dinkins. The content of the document can also contain temporal references such as "last year" or "2001" which can be useful indicators of important time intervals. The set of temporal expressions contained in d_p is given by the set $E(d_p) = \{I_1, I_2, \ldots I_n\}$ where I is an arbitrary time interval with a definite begin and end timestamp denoted by $begin(I)$ and $end(I)$ respectively.

Temporal Model: We adopt a discrete notion of time for the collection and assume that a time-stamp t_i is a positive integer and is computed periodically, with a fixed granularity Δ, from a refer-

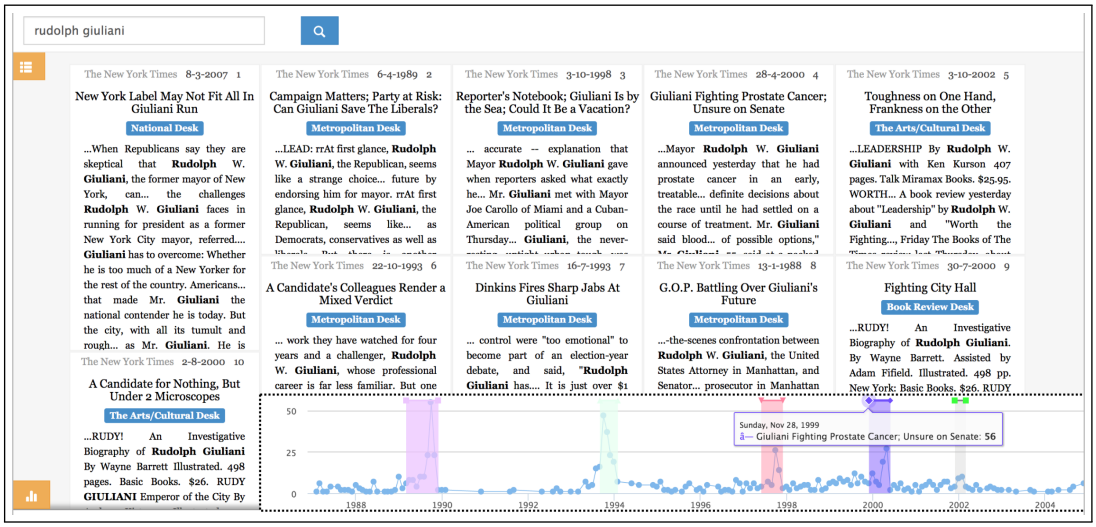

Figure 1: The HistDiv search system. The search results are layed out to mimic a newspaper. The coloured areas in the timeline represent bursts detected from the temporal distribution. Results shown are for the top 10 out of 26,000 results returned for the query Rudolph Giuliani.

ence point in the past t_0. The discretized time span of the collection is denoted by an ordered set $W = \langle t_0, t_1, \ldots t_n \rangle$ such that $\Delta = t_{i+1} - t_i$. We define $\delta_i = [t_i, t_{i+1})$ as the elementary time intervals between two consecutive time points. A document d_p is published in the interval δ_i if $t_i \leq p < t_{i+1}$ and is given by the function $\Lambda(p) = \delta_i$. The set of all elementary time intervals of size Δ in W is called the temporal space $\mathcal{T} = \{\delta_i \mid \delta_i \in \bigcup_{d_p \in \mathcal{D}} \Lambda(p)\}$.

Aspect-Time Space: For a given query q, $\mathcal{R}_q$ is set of top-K documents retrieved such that $\mathcal{R}_q \subseteq \mathcal{D}$. The time space relevant to q is the set $\mathcal{T}_q = \{\delta_i \mid \delta_i \in \bigcup_{d_p \in \mathcal{R}_q} \Lambda(p)\}$ such that $\mathcal{T}_q \subseteq \mathcal{T}$. Similarly we define the aspect-space of q as the set of all aspects found in documents from $\mathcal{R}_q$ denoted by $\mathcal{A}_q = \{a_i \mid a_i \in \bigcup_{d_p \in \mathcal{R}_q} A(d_p)\}$ where $\mathcal{A}_q \subseteq \mathcal{A}$. However, not all aspects are relevant in all time intervals. For example, the aspect World Trade Center for the query `rudolph giuliani` is historically irrelevant for time intervals before 2001. To this end, we define a combined aspect-time space $\mathcal{AT}_q \subseteq \mathcal{A}_q \times \mathcal{T}_q$ which contains aspect-time pairs encoding subspaces which are both temporally and aspect-wise relevant.

$$\mathcal{AT}_q = \{(a_i, \delta_j) \mid a_i \in \mathcal{A}_q \wedge \delta_j \in \mathcal{T}_q\}.$$

A result document d_p published at p for query q with aspects $A(d_p)$ and temporal expressions $E(d_p)$ is said to be relevant to a query aspect-time pair $(a_i, \delta_j) \in \mathcal{AT}_q$ if $a_i \in A(d_p)$ and $\Lambda(p) = \delta_j$.

3.2 Problem Definition

The historical search result diversification problem or simply the *historical search task* intends to find a re-ranking S of an initial result set $\mathcal{R}_q$ for a query q that has maximum coverage and minimum redundancy with respect to different aspect-times underlying q. In other words, it is the standard search result diversification task but over the $\mathcal{AT}_q$-space which encodes the historical query intent of obtaining relevant documents from the most important aspects from the time-period/s it is important in. As shown by [3], search result diversification is a bi-criterion optimization problem which can be reduced from the maximum k-coverage problem and hence is $\mathcal{NP}-$hard.

3.3 Evaluation Measures

Given that we define a new two-dimensional solution space, we could re-use the standard diversity-based retrieval measures for evaluating approaches to the historical search task, considered on the $\mathcal{AT}_q$-space. Subtopic-recall $\mathsf{SBR}_{q,k}$, for instance, for the result set $\mathcal{R}_q$ at depth k, can be computed as:

$$\frac{\left| \bigcup_{d_p \in \mathcal{R}_q^k} \{(a_i, \delta_j) \mid a_i \in A(d_p) \wedge \Lambda(p) = \delta_j\} \right|}{|\mathcal{AT}_q|}.$$

Similarly for other diversity metrics suggested in [24] like intent aware ERR (IA-ERR), intent aware precision (IA-P), mean average precision MAP and α-NDCG (NDCG) we can substitute the subtopic space with the $\mathcal{AT}_q$-space.

4. RELATED WORK

Before we describe the HistDiv approach for historical search tasks, in this section we outline the relevant existing literature. Our problem has overlap mostly with areas relating to temporal representation and temporal retrieval models under temporal information retrieval, and with works on search result diversification.

4.1 Temporal Information Retrieval

Temporal information retrieval has emerged as an important subfield in IR with the goal to improve search effectiveness by exploiting temporal information in documents and queries [12]. The value of the temporal dimension was clearly identified in [4] and has led to a plethora of work which utilizes temporal features in query understanding [32, 35], retrieval models [8, 25, 26, 11], temporal indexing [9, 6], clustering [5] and query modelling [38, 37, 17]. A survey by Campos et. al. [12] gives an elaborate overview of the field.

Improving Ranking using temporal features One of the first algorithms to incorporate time in search result ranking was suggested in [33]. They used a temporal language model approach where time and term importance are handled implicitly. Various approaches have been suggested that consider time more explicitly. [8] proposes a language modelling approach taking into account the temporal expressions in the query and document text. [11] on the other hand, taking a non-probabilistic interpretation of relevance, defines *temporal scope similarity* between queries and documents in *metric spaces*. In both these works, similarity between the tem-

poral references in the query and documents are used to rank documents. In our query model, we never make any assumptions on the presence of temporal references nor do we model the similarity of query and document based on temporal references. Another line of work in this domain considers the freshness or recency of a document when ranking [25].

Finding important time periods An important ingredient in our retrieval model is finding temporal priors for different time points. [32] estimates a probability distribution over different time points for each query called the temporal query profiles using publication times of the documents. In [40], the authors exploit the publication dates to identify important time points for a given query by contrasting rankings for adjacent time points. However, neither of them utilize secondary sources or temporal references in text. Recently, [29] suggests using temporal references for ranking time intervals for a given temporal query. In our work, we use both publication dates (primary sources) and temporal references (secondary sources) to assign temporal priors to each granular time window akin to a temporal profile [32].

4.2 Search Result Diversification

Diversity in search (both explicit and implicit) has seen a rich body of literature lately in [22, 14, 3, 39, 15, 45, 34, 23]. Search result diversification aims to maximise the overall relevance of a document ranking to multiple query aspects, while minimising its redundancy with respect to these aspects. Existing approaches differ in the way they model different query aspects. Implicit approaches, like [14], assume similar documents cover similar aspects and do not model aspects. Explicit approaches model aspects in a variety of ways using query logs, taxonomies etc. We also model aspects explicitly by using entities found in text documents as their aspects. Also, none of the previous approaches take time into account or model the historical information intent.

Temporal diversification [7] proposes a diversification model which considers time windows as a set of intents for a query while modeling the importance of each intent as the weight of its burst. In traditional aspect-based diversification tasks like [20, 18], intent importance is considered static over time. However, intent importance was shown to vary across time; thus affecting the diversity evaluation of queries issued at different time points [44]. Keeping this in mind, [36] considers the time at which the query is issued to diversify intents based on their temporal significance at that time. Their approach also explicitly models time and aspects, although latent, but rewards recency. HistDiv, on the other hand, is query-time agnostic, since it is intended for historical search, and seeks to diversify documents based on both time and aspects.

5. THE HISTDIV APPROACH

5.1 Approach Overview

The challenge in designing a retrieval model for the historical search task is in identification of important time intervals and aspects. More importantly, the documents which optimize both dimensions and yet are relevant. Towards this, we first model the temporal space for the initial result set $\mathcal{R}_q$. We build a probability distribution $P(\delta_i|q)$ for $\delta_i \in \mathcal{T}_q$ over the entire time span W taking into account the publication times and temporal references (mined from document text). Such a temporal profile, as shown in Figure 1 helps us isolate the important time-periods in the result timeline (described in Section 5.2). Next in Section 5.3, we detail how we build priors for the aspects of the documents. Finally in Section 5.4, we present our diversification algorithm HistDiv which takes into account textual relevance, temporal sensitivity and aspect importance along with the *typical semantics* of the temporal and aspect domains to maximize coverage in the $\mathcal{AT}_q$ space.

Algorithm 1: The HistDiv Algorithm

Input: $k, q, \mathcal{A}_q, \mathcal{R}_q, \mathcal{T}_q, V(d|q), S = \emptyset$
Output: Set S of diversified documents

$\forall a \in \mathcal{A}_q$, $\forall \delta_i \in \mathcal{T}_q$, $U_{aspect}(a|q, S, \delta_i) =$ refer Eqn. 1
$\forall \delta_i \in \mathcal{T}_q$, $U_{time}(\delta_i|q, S) =$ refer Eqn. 2
while $|S| \leq k$ **do**
  while $d \in R$ **do**
    $g(d|q, S) \leftarrow \alpha.V(d|q) + (1-\alpha).(\beta.\sum_a^{A(d)} U_{aspect} + (1-\beta).U_{time})$
  $d^* \leftarrow argmax_d\ g(d|q, S)$
  $S \leftarrow S \cup \{d^*\}$
return S

5.2 Building Temporal Priors

To find important time intervals, we build a probability distribution $P(\delta_i|q)$ over the entire time span by projecting both the publication times and the reference times into $\mathcal{T}_q$. The temporal references are treated as secondary sources and can be used as indicators of relevant primary sources.

For the computing distributions we use the document counts published in a time interval δ_i (contrary to using top-k relevance scores for profile generation [32]). The probability $P_{pub}(\delta_i|q)$ is the fraction of all documents in $\mathcal{R}_q$ published in δ_i. To compute $P_{ref}(\delta_i|q)$ we first estimate the contribution of an interval $I \in E(d_p)$ as $\frac{1}{|I|}$ for all constituent time intervals δ_i. Finally, we employ a language modeling strategy to smooth the probability distribution of the publication times P_{pub} with the background distribution of the temporal references P_{ref} with a mixing parameter θ to arrive at a distribution $P(\delta_i|q)$:

$$P(\delta_i|q) = \theta.P_{pub}(\delta_i|q) + (1-\theta).P_{ref}(\delta_i|q).$$

In our experiments we demonstrate the value of estimating the time prior by comparing it to a temporal diversification baseline that assumes equal distribution of δ_i called EqT.

5.3 Aspect Modeling

Historians are particularly interested in events which can be described using groups of entities associated with specific time intervals. Keeping this in mind, we use entities mentioned in the document text as our aspects.Traditional aspect-space diversification methods, like IA-Select [3] and PM2 [23], estimate the probability $P(a_i|q)$ assuming the collection is static. For historical search, time is an essential factor and needs to be considered when estimating $P(a_i|q)$. Consider a document d_p published in time interval $\Lambda(p) = \delta_j$; an aspect $a_i \in A(d_p)$ can be temporally diverse if it occurs in documents from different time intervals. Hence the aspect a_i has a probability distribution across time intervals $\delta_j \in \mathcal{T}_q$. Consequently, the prior probability of an aspect $a_i \in \mathcal{A}_q$ in a time interval $\delta_j \in \mathcal{T}_q$ is given by:

$$P(a_i|q, \delta_j) = \frac{\left|\{d_{a_i,p} \mid a_i \in A(d_p) \wedge \Lambda(p) = \delta_j\}\right|}{|\{d_p \mid \Lambda(p) = \delta_j\}|}$$

where $d_{a_i,p}$ is a document tagged with aspect a_i published in the interval δ_j. Notice in the Figure 1 that the event mayoral campaigns is recurring every 4 years. Hence the aspects representing mayoral campaigns will have a higher $P(a_i|q, \delta_j)$ in certain time intervals $\{\delta_{1989}, \delta_{1993}, \delta_{1997}\}$ when compared to the others.

5.4 The HistDiv Algorithm

In classical diversification approaches like [3, 23, 27], each document is assigned a utility score computed using textual relevance

(denoted as $V(d|q)$ in Algorithm 1) and aspect importance. Most approaches employ a greedy algorithm which selects the candidate documents that maximize utility with respect to the uncovered aspects in each iteration. HistDiv considers both aspect and time dimensions (with their special semantics) and operates in a similar manner treating both the topical and temporal aspects as sets thereby retaining the $(1 - 1/e)$ approximation guarantee. However, we differ significantly from previous approaches in the way we interpret and compute the utility of each dimension.

Traditional diversification algorithms model only aspects and maximize coverage in the space $\mathcal{A}_q$. Since the objective is to maximize coverage in the $\mathcal{AT}_q$ space we first consider how they can be adapted to model the required space $\mathcal{AT}_q$.

Temporally augmented aspect space: A naïve approach to introduce time could be to enrich the aspect space by adding time intervals as new aspects. For instance for d_p published in δ_i we can add δ_i to $A(d_p)$. In our experiments we use this method to create two variations of IA-Select and PM2 called E-IA-Select and E-PM2.

Linearizing aspect space with time: Since we deal with two dimensions we can project or *linearize* the temporal dimension onto the aspect dimension. More formally, the result of linearization is the set of m aspects $\{a_1\delta_i, \ldots, a_m\delta_i\}$ for document $d_p \in \delta_i$ which is used to alter IA-Select and PM2. We name these two variations T-IA-Select and T-PM2 and also use them as baselines in our experiments.

Another alternative would be to keep the dimensions separate like the multi-dimensional approach proposed in [27] (referred to as MDiv henceforth). In this general framework, for the diversification of n arbitrary dimensions, the utility score $g(d|q, S)$ computation reflects how the dimensions are combined. The marginal utility of aspects given a document d is computed based on rank of d for the given aspect a_i. We can naturally add time as a second dimension and use it for diversification. We also use MDiv as a baseline in our experiments.

A key drawback of both these approaches is that they do not consider the fact that: (a) temporal aspects are ordered and thus have special semantics (b) temporal and topical aspects are interrelated. Hence MDiv's assumption of dimensional independence and identical discounting function for both dimensions might not yield optimal results. A retrieval model designed for a historical search task however should take this into account while computing the utility of aspects and time intervals.

In HistDiv, the utility of a document in the aspect space is measured by a function $U_{aspect}(a_i|q, S, \delta_j)$ with the exception that we treat an aspect in various time windows differently. We discount aspects in a neighborhood defined by the w so that $P(a_i|q, \delta_j)$ is strongly discounted if δ_j is temporally closer to a document $d_p \in S$ and $a_i \in A(d_p)$. We use the decay function suggested in OnlyTime [7] to discount aspects across time. In this way, we avoid the time agnostic property of standard topical diversification retrieval models which may select the right aspects but will discount the aspect for the entire span of the collection thereby reducing the probability of selecting documents for the same aspect from other important time intervals. The utility of an aspect, $U_{aspect}(a_i|q, S, \delta_j)$, is

$$P(a_i|q, \delta_j) \prod_{d_p \in S} \left(1 - \frac{1}{1 + e^{-w+|t_j - p|}}\right) \quad (1)$$

t_j denotes the boundary time point of δ_j. Like OnlyTime we can set w to the size of Δ. The limitation of modeling utility this way is that we are restricted by the fixed w. Consider bursts, where a high number of documents from multiple consecutive δ typically discuss a single event (especially if Δ is small). If we select a document about this event from the edge of the burst then we face two potential issues: (i) we may assign high utility to documents which are about similar aspects from two temporally distant intervals but still refer to the same event (ii) we heavily discount documents just outside the burst (temporally close but sufficiently different) unfavorably. Both of these issues lead to a potential drop in subtopic recall whcih we address by using the burst detection technique suggested in [46]. To detect the set of bursts B_q, this technique utilizes the mean and standard deviation of a fixed width sliding window across W.

We can now vary w depending on the position of d_p within its corresponding burst ($b_i \in B_q$) or non-bursty interval ($b_i \in \hat{B}_q$) and use $U_{aspect}(a_i|q, S, \delta_j)$ as before. w is then computed as follows:

$$w = \begin{cases} |p - begin(b_i)| & : p \geq t_j \\ |p - end(b_i)| & : p < t_j \end{cases}$$

In the time dimension, we need to be wary of discrediting a time interval too heavily. OnlyTime produces a result set with high temporal diversity by selecting relevant documents from important intervals and discounts those intervals heavily with the aforementioned decay function. This approach to discounting bursts doesn't consider the fact that a single burst could consist of many diverse aspects. For example, in 2000-2001 Giuliani was divorced, diagnosed with cancer and was involved in helping New York recover from 9/11. Hence unlike OnlyTime, we discount the interval in the time dimension of $d_p \in S$ by the weighted proportion of aspects covered by it. The utility of time, $U_{time}(\delta_i|q, S)$, is

$$P(\delta_i|q) \prod_{d_p \in S} \left(1 - \frac{|\,(d_{a_j,p^*} \mid a_j \in A(d_p) \wedge \Lambda(p^*) = \delta_i)|}{|(d_{p^*} \mid \Lambda(p^*) = \delta_i)|}\right) \quad (2)$$

With burst detection, for all $d_p \in S$ we simply discount all time intervals δ_i contained in it's corresponding bursty / non-bursty interval.

The essence of our approach lies in our temporal interpretation of aspect utility and aspect aware interpretation of time utility. This interpretation helps us maximize coverage in the joint space $\mathcal{AT}$, as shown in our experiments, when compared to pure aspect based diversification, pure time based diversification and multidimension diversification. Algorithm 1 shows the iterative process in which documents are selected in to result based on the utility g, where aspect and temporal utilities are traded-off by the parameter β. The parameter α trades-off the impact of document relevance $V(d|q)$ with the utility of the two dimensions.

6. TEST COLLECTION

In this section we detail the test collection which we constructed to evaluate our approach. There exist well known collections for the standard diversification task (diversity tasks in the TREC Web track), temporal information retrieval (Temporalia'14 [31]) as well as web archives [21]; however to the best of our knowledge there are no established test collections to measure the effectiveness of retrieval models designed for diversification of the $\mathcal{AT}_q$ space in news archives. Hence we choose to build our own collection guided by the target user group - historians, whose judgments will be made available to the research community.

Metrics: The evaluation metric for a search task should reflect the user's goal. For users with historical query intents the objective is to find primary sources of information from all the important aspects and time periods. The primary metric we choose to measure the effectiveness of a retrieval model is SBR (subtopic recall) in the joint aspect-temporal space $\mathcal{AT}_q$ because of the recall oriented nature of historical query intents.

Document Corpus: As a corpus we use the *Annotated New York Times* collection [2] which qualifies as a suitable news archive since

it spans for 20 years, i.e., 1987 - 2007. Although there exists larger news corpora, they span for a shorter time duration reducing the likelihood of having ample primary sources .Also, the timestamps associated with the articles are accurate and do not have to be estimated as in other web collections. The corpus consists of articles from all sections of the newspaper including the editorial desk, arts, technology and literature making it replete with various aspects interesting for historians.

Search topics: Topics, with a historical intent, for our test collections are derived from experts who held discussions with historians at the Institute of Historical Research as well as from insights in Section 2. These experts first described the intents verbosely and then proceeded to identify keywords that represent it. Since our corpus is a newspaper daily for the USA, topics are chosen from a set of historically relevant issues related mostly to the USA and a few from more global issues. To define the *subtopics* of each topic the experts were guided by the history sections from the relevant Wikipedia articles. To confirm or modify subtopics, they explored the corpus with a simple keyword search interface whenever necessary. The chosen subtopics are also qualified by a set of relevant time periods. The experts chose time periods of relevance to the subtopic by consulting relevant Wikipedia articles and defined the interval size as the period in which they found primary sources in the corpus. Depending on the type of subtopic the time interval can span months (Giuliani's efforts in the aftermath of 9/11) or years (Giuliani's senate run). Each subtopic can also have multiple time intervals like Giuliani's mayoral election campaigns which are relevant during 1989,1993 and 1997. Time intervals can also overlap each other, for instance Giulaini's personal life (struggle with cancer) and his senate run.

Listing 1: Excerpt of a topic in the workload

```
<topic>
 <query>rudolph giuliani</query>
 <desc>I want to know the history of Rudolph Giuliani</desc>
 <subtopics>
  <subtopic>
   <desc>Mayoral campaigns</desc>
   <time>[{01.01.1989 - 31.12.1989}, {01.01.1993 - 31.12.1993}, {01.01.97 -
        31.12.1997}]</time>
  </subtopic>
  <subtopic>
   <desc>Senate race</desc>
   <time>[{01.01.2000 - 31.12.2000}]</time>
  </subtopic>
  <subtopic>
   <desc>Efforts after 9/11</desc>
   <time>[{11.09.2001 - 01.04.2002}]</time>
  </subtopic>
  .
  .
  .
  .
 </subtopics>
</topic>
```

We have a total query workload of 30 topics. On average there are 5 subtopics per topic and each subtopic has at least one relevant time interval. The types of topic chosen are inspired by the characteristics defined in Section 2, i.e, broad topics related to entities like `Rudolph Giuliani` and the `Atlantic City`, major events like the `reunification of germany` and `team usa soccer world cup` as well as controversial subjects like `gay marriage` and `sarin gas`. A key assumption made when creating subtopics is the omission of historical facts that lie outside of the 20 year time period of the NYT corpus.

Pooling: We devise suitable baselines (detailed in Section7.1.2) and "submit" runs for each baseline corresponding to all possible parameter settings. By doing so we increase the coverage of documents for each topic and improve the diversity of the pool. We chose a run size of top 20 for all topics and the pool size was set to 300 documents per topic. We overall generated nine competitors which produced on average 20 runs per baseline. To gather relevance judgments we use the Cranfield paradigm [43]. Trained evaluators were instructed to assign binary relevance judgments to topic, subtopic, document triples.

Once the pools were evaluated, a standard robustness test was carried out with $\mathcal{AT}_q$ subtopic recall as the primary measure. We selected 25% of the query workload at random and split them into two equal sets. We selected 50% of the runs at random for retrieval depth 10 and calculated ranked the system runs for both sets of queries. We found that the rankings were consistent for p<=0.05.

7. EXPERIMENTAL EVALUATION

In this section, we thoroughly evaluate our HistDiv approach and the impact of its components for the historical search task using the new test collection. Before we present our results, we detail our experimental setup in which we describe how we mine our aspects and select baselines. Then to assess the effectiveness of our approach we first present in Section 7.2 the overall retrieval effectiveness across different retrieval depths, assess the impact of varying the granularity Δ and also highlight certain drawbacks. Next, in Section 7.3 we discuss interesting insights from a user study to estimate the quality of an overview produced by different rankings and finally summarize take-aways from our experiments.

7.1 Setup

7.1.1 Modeling Aspects and Time

Since we model aspects of a documents as the entities therein we consider a named-entity tagging system Aida [30] for our experiments. Aida is the state-of-art approach for *named entity disambiguation* which canonicalizes mentions of named entities (person, locations, organizations) into Wikipedia pages. Note, there are certain entities like United States that occur in nearly all documents and distort the performance of aspect based diversification methods. To overcome this we remove all entities non-salient to the collection using an IDF filter and set the threshold to 0.2. To extract the temporal references mentioned in the article text and populate $E(d)$ we use *Tarsqi*. We consider two granularities Δ of time intervals for the experiments – $\Delta = \{\text{month}, \text{year}\}$.

7.1.2 Baselines

We evaluate the effectiveness of HistDiv at diversifying the search results produced by four effective classes of baselines:

Non-temporal Baselines : The first baseline that we considered is the standard unigram language model with dirichlet smoothing (LM, $\mu = 1000$). The other approaches use the top 1000 documents returned by LM for diversification. Though not designed for the historical search task we also consider IA-Select [3] and PM2 [23] which are pure aspect-based diversification methods. It allows us to better highlight the nature of the task and the challenges faced when diversifying across a joint aspect-time space.

Temporal Diversification Baselines : First, to create baselines more suited to the task we create temporal variants of IA-Select and PM2 by (a) linearizing the aspects with the time window corresponding to the publication date of the document respectively denoted as T-IA-Select and T-PM2; (b) augmenting the aspect space by including temporal aspects. Temporal aspects are represented by time-intervals and contain documents which were published in that interval. These variations are named E-IA-Select and E-PM2 respectively. Next, we consider two approaches that take time into account directly while diversifying: OnlyTime [7] and MDiv [27]. OnlyTime [7] diversifies results purely in the time dimension, denoted by $\mathcal{T}_q$. MDiv [27] is a multi-dimensional diversification approach that treats both dimensions equivalently. Finally, we also have a variant of OnlyTime called EqT which, unlike its counterpart, assumes equal distribution for the prior and does 0/1 discount-

	k=10			k=15			k=20		
	A	T	$AT\ (W/L\%)$	A	T	$AT\ (W/L\%)$	A	T	$AT\ (W/L\%)$
LM	0.706	0.060	0.428	0.752	0.085	0.491	0.780	0.091	0.518
IA-SELECT$^{\circ}$	0.722	0.039	0.442 (23/23)	0.766	0.047	0.491 (20/26)	0.841	0.055	0.516 (20/23)
PM2*	0.707	0.069	0.429 (16/20)	0.794	0.082	0.471 (10/23)	0.817	0.097	0.509 (16/26)
TIA-SELECT$^{\bullet}$	0.614	0.039	0.380(23/36)	0.717	0.047	0.433 (20/43)	0.770	0.055	0.470 (20/26)
T-PM2$'$	0.551	**0.088**	0.308 (13/50)	0.680	0.106	0.408(20/43)	0.761	0.128	0.453 (16/33)
E-IA-SELECT‡	0.700	0.062	0.435 (23/23)	0.776	0.084	0.501 (23/23)	0.837	0.095	0.524 (23/20)
E-PM2†	0.692	0.061	0.422 (6/16)	0.766	0.083	0.469 (6/26)	0.816	0.098	0.495 (10/26)
EQT	0.714	0.076	0.440 (16/13)	0.766	0.097	0.503 (13/6)	0.802	0.117	0.542 (20/6)
MDIV$^{\blacktriangle}$	0.720	0.060	0.460 (33/33)	0.764	0.079	0.515 (23/16)	0.823	0.096	0.552 (29/3)
ONLYTIME$^{\diamond}$	0.729	0.068	0.426 (20/26)	0.807	0.092	0.497 (26/26)	0.826	0.115	0.534 (26/20)
HISTDIV	0.761$^{\diamond}$	0.07	0.497$^{\blacktriangle}$ (40/13)	0.814	0.085	0.542$^{\blacktriangle}$ (36/26)	**0.864**‡	0.101	0.583$^{\blacktriangle}$(43/13)
HISTDIV-BURST	**0.777**$^{\diamond}$	0.087	**0.509**$^{\blacktriangle}$ (33/6)	**0.830**$^{\diamond}$	**0.113**$'$	**0.560**$^{\blacktriangle}$ (46/20)	0.860‡	**0.132**	**0.601**$^{\blacktriangle}$ (43/16)

Table 1: SBR at varying depths for Δ =month. Win-Loss percentages are presented in brackets next to the $\mathcal{AT}$ scores. The superscript denotes a statistically significant difference when compared to the closest competitor (p<=0.5). For example $^{\diamond}$ represents statistically significant difference from ONLYTIME.

ing of the time windows. We do not consider xQUAD [39] due to (a) the absence of a reasonable query log for this time period and (b) the lack of clarity regarding its' adaptation without an external log for the historical search task.

HistDiv : We compare both the original HISTDIV and its burst-aware version HISTDIV-BURST to the aforementioned baselines. For HISTDIV-BURST we fix the the moving window size used to detect bursts to 24 months in all experiments.

To get the best performance and avoid over fitting we tune each variant for SBR in $\mathcal{AT}_q$ and present results using 5-fold cross validation. We evaluated all baselines for metrics mentioned in Section 6 at variable retrieval depths. Similar to the TREC diversity track we assumed equal distribution of subtopics for all topics. For the temporal space $\mathcal{T}_q$ each time window in the ground truth was divided into partitions of size Δ and then given equal importance akin to [7]. Consequently, for the $\mathcal{AT}_q$ space each subtopic and relevant time interval pair is given equal importance. Note, in the joint space we do not divide the qualifying intervals into partitions. We assumed equal distribution of the qualifying intervals associated with a single subtopic.

7.2 Results

In this section we first analyze the performance of all baselines for historical query intents using subtopic recall or SBR since our goal is to optimize SBR. Even though the measure of choice for a historical search task is SBR in the $\mathcal{AT}-$ space, observing the component spaces $\mathcal{A}$ and $\mathcal{T}$ provides a clearer explanation of our results. Table 7.2 summarizes the effectiveness of all baselines at Δ = month with respect to SBR in the aspect space $\mathcal{A}$, temporal space $\mathcal{T}$ and aspect-temporal space $\mathcal{AT}$ at $k = \{10, 15, 20\}$. Remember that we have a nested representation of important subtopics and times in our topics definition (cf. Section 6). A document d is relevant to a subtopic in the ground truth, irrespective of its corresponding time interval, is said to be relevant in aspect space $\mathcal{A}$. Similarly d, if published in a time interval I in the ground truth, is said to be relevant in the temporal space $\mathcal{T}$ irrespective of which subtopic it is relevant to.

First, we look at the performance of the non-temporal baselines. Not surprisingly, the diversity-unaware LM fares worse than most baselines. IA-SELECT performs better than LM in the aspect space $\mathcal{A}_q$ but since it does not account for time, we find that SBR in the temporal space $\mathcal{T}_q$ is lower compared to the temporal baselines. Consequently it performs poorly in $\mathcal{AT}_q$. The proportionality-based approach PM2 also performs poorly suggesting that choosing aspects proportionally is detrimental to historical query intents.

Next, we consider the temporal variants of the standard diversification approaches. We find that the linearized variant T-IA-SELECT performs consistently worse than its non-temporal counter part in all spaces across all retrieval depths. This shows that linearizing aspects with time leads to *over-specification of aspects* especially for smaller time granularities. E-IA-SELECT addresses this problem leading to significantly better results in all spaces when compared to T-IA-SELECT. It also performs significantly better in $\mathcal{T}_q$ when compared to IA-SELECT. As a consequence of the inherent temporal nature of aspects alluded to earlier, the subtopic recall of E-IA-SELECT in the $\mathcal{AT}_q$ improves and is either better or comparable to IA-SELECT along with performance in the other metrics showing marked improvement (shown in Table2).

Now we consider the other temporal baselines ONLYTIME and MDIV in comparison to the temporal baselines discussed above. We observe that MDIV has comparable to the best performance (although not the best) in both $\mathcal{T}_q$ and $\mathcal{A}_q$ and therefore is easily the best performing temporal baseline in $\mathcal{AT}_q$. Surprisingly, for Δ = month, ONLYTIME does not perform as well as we expected in $\mathcal{T}_q$ even though it optimizes for temporal coverage. However, for Δ = year, it significantly outperforms all competitors in $\mathcal{T}_q$ including HISTDIV (results omitted from Table 7.2 for space reasons) which leads us to believe that ONLYTIME is sensitive to the underlying granularity Δ. Expectedly, both ONLYTIME and EQT perform worse than MDIV. Since subtopics need not be from mutually exclusive time intervals, ONLYTIME struggles to cover the relevant $\mathcal{AT}_q$ space properly. MDIV, on the other hand, by virtue of modeling aspects (albeit independently) reconciles both dimensions better and is our closest competitor.

Lastly, we turn to HISTDIV and its burst-aware version HISTDIV-BURST. HISTDIV outperforms all classes of competitors in the $\mathcal{AT}_q$ space and is significantly different from its closest competitor MDIV by an achieved significance level of $p \leq 0.05$. The results vindicate our choice of mining aspects temporally and considering time period granularities as aspect proportions for $\mathcal{AT}_q$. A key point to note is that methods which perform comparably to HISTDIV in spaces $\mathcal{A}_q$ and $\mathcal{T}_q$ like IA-SELECT and T-PM2 are significantly outperformed in the space $\mathcal{AT}_q$ meaning that we achieve the right trade-off between these two seemingly conflicting yet interdependent dimensions. HISTDIV-BURST is the best performing variant in all spaces for all depths. The improvement over HISTDIV can be

	IAP		SBR		NDCG		IA-ERR		MAP	
	M	Y	M	Y	M	Y	M	Y	M	Y
LM	0.099	0.099	0.428	0.428	0.402	0.402	0.201	0.201	0.228	0.228
IA-SELECT$^{\circ}$	0.101	0.101	0.442	0.442	0.415	0.415	0.180	0.180	0.215	0.215
PM2*	0.100	0.100	0.429	0.429	0.388	0.388	0.213	0.213	0.241	0.241
TIA-SELECT$^{\bullet}$	**0.120**$^{\blacktriangle}$	**0.113**‡	0.380	0.361	**0.497**‡	**0.468**$^{\circ}$	0.195	0.179	0.242	0.232
T-PM2$'$	0.064	0.091	0.308	0.410	0.232	0.368	0.123	0.176	0.152	0.167
E-IA-SELECT‡	0.106	0.102	0.435	0.430	0.478	0.412	0.183	0.177	0.219	0.214
E-PM2†	0.103	0.099	0.422	0.417	0.419	0.379	0.217	0.204	0.227	0.239
EQT	0.096	0.078	0.441	0.426	0.360	0.331	0.203	0.200	0.229	0.213
MDIV$^{\blacktriangle}$	0.109	0.096	0.460	0.428	0.389	0.370	0.204	0.203	0.236	0.236
ONLYTIME$^{\diamond}$	0.089	0.076	0.426	0.415	0.354	0.297	0.196	0.189	0.236	0.220
HISTDIV	0.096	0.087	0.497$^{\blacktriangle}$	0.459°	0.383	0.339	0.229*	0.208	**0.255**$^{\bullet}$	0.231
HISTDIV-BURST	0.096	0.096	**0.509**$^{\blacktriangle}$	**0.509**$^{\circ}$	0.375	0.375	**0.231***	**0.231***	0.244	**0.244**

Table 2: Effect of granularity Δ (M = month; Y = year). The superscript denotes a statistically significant difference when compared to the closest competitor (p<=0.5). For example $^{\diamond}$ represents statistically significant difference to ONLYTIME.

attributed to the accurate identification of important time intervals pertaining to long running events. HISTDIV-BURST which also has the advantage of being granularity free and as evidenced by our results, is a better retrieval model to find relevant documents from important aspects during important time intervals. In Table 1 we also show the win-loss percentage of queries in $\mathcal{AT}$ compared to the baseline LM. We see that HISTDIV improves a sizable portion of the workload (around 43% at $k = 20$) but more importantly under performs only in a small set of queries (16%).

Unlike HISTDIV-BURST the other retrieval models that incorporate time are dependent on the granularity of the time window used to discretize the time dimension. To examine the effect of granularity on the baselines we report scores in $\mathcal{AT}_q$ at $k = 10$ in Table 2 using all intent-aware metrics mentioned earlier in Section 6 for $\Delta = \{Year, Month\}$. We find that a yearly granularity works best for certain retrieval models because, for the given workload, a smaller granularity causes temporally-aware diversification methods to *over represent* relevant time periods. HISTDIV-BURST on the other hand is more robust against over representing long running subtopics like Giuliani's mayoralty. However for queries like `reunification of germany`, temporal spread is not key to diversity but instead it is imperative to diversify within a small set of time intervals. Here T-PM2, irrespective of granularity, is able to focus on to the dominant time window "1989" and covers as many aspects as possible within this window. On the other hand, HISTDIV covers enough important aspects in the important time intervals and subsequently diversifies to find aspects in other less relevant intervals. Although with sufficient tuning HISTDIV can perform as good or if not better than T-PM2.

Interestingly, HISTDIV also performs the best in $\mathcal{A}_q$ which leads us to believe that temporal aspects guide us to make the right choice of subtopics for temporally ambiguous queries. However this is achieved at the cost of precision. In Table 2, we can also see the performance of all retrieval models for the other metrics. Historical search is cast as a recall-oriented task and hence it is not surprising to find HISTDIV is not the best in IA-P. The baseline TIA-SELECT consistently outperforms other competitors but does so by producing aspect-redundant documents indicated by the low SBR score. We also notice that IA-SELECT and MDIV achieve good precision with satisfactory subtopic recall. Balancing precision and recall is key to satisfying the target user group. In the next section we choose to focus on the intended users of our search system. We first briefly highlight the quantitative results from user-centric metrics. We then take a more qualitative approach to study if this drop in precision for HISTDIV is detrimental to user satisfaction in historical search tasks.

Finally for completeness, though not included in Table 1, we considered two other competitors – the search engine on the official New York Times portal and a commercial search engine using the appropriate site (nytimes.com) and date filters (1987-2007) on June 9, 2015. The SBR scores achieved by them were the lowest – 0.288 and 0.312 at $k = 10$ respectively. Such low recall can be attributed to the fact that commercial search engines favor recency and more popular information needs reaffirming the need for specialized retrieval models designed for historical intents.

7.3 User Study

We now turn our attention to the target users to get a deeper understanding of what they perceive is a good historical overview from search results. In this section we attempt to answer the following research question: *Despite a loss in precision, are users satisfied with the quality of the overview derived from HISTDIV when compared to its competitors?* In traditional IR, to quantify if search results are satisfactory for users (who inherently examine a result list from top to bottom) metrics such as IA-ERR and α-NDCG are used. From Table 2 we see that HISTDIV performs well in IA-ERR, which is one of the primary metrics for TREC's diversity web track, due to its ability to not only rank diverse documents higher but also cover more subtopics in the top 10 results. The lack of precision is reflected in the low NDCG(α set to 0.5) scores although HISTDIV has the highest MAP score. What these results do not indicate however is the overall satisfaction of the user when trying to discern an overview from the top-K search results as a whole. We see quantitatively that we cover more subtopics at important time periods but how good is the overview from a user perspective? Is the lack of precision detrimental to the overall user experience?

Study design: We attempt to answer the research question by comparing HISTDIV against IA-SELECT because IA-SELECT exhibits high precision coupled with good recall. It is also distinctly different from HISTDIV since it is a pure aspect based diversification algorithm. Due to the sheer number of baselines, comparing against all is prohibitive. We used a within-subjects study design to determine which approach produces a better overview. We had a total of 10 participants comprised of 3 senior historians, 1 humanities researcher and 6 computer science graduate students. The historians and the humanities researcher represent the expert opinion. Participants were required to compare and contrast both approaches for a subset of the query workload from the test collection. Of the three historians two were from the United States and the other from the United Kingdom. All of them cited previous experience using

news archives for their research. The non-expert users were from three different nationalities and possessed strong proficiency in the English language. All participated voluntarily in the study.

We selected 15 topics at random from our workload and generated the top 10 results for each from both approaches. Each participant was given between 2-5 topics to evaluate. To take into account the varying familiarity of participants with their designated topics we instructed them to first read the history section of the relevant Wikipedia page for the topic. We suggested a period of 5-15 minutes for this preparation. Following this period, participants were shown the top 10 results from both methods and asked the following question : "Which result ranking gives you a better overview of the history for the given topic and why?". The approaches were anonymized and presented to the users as "X" (IA-SELECT) and "Y" (HISTDIV). Participants were instructed to argue their choice in the form of free text. We opted against structured questions since a good overview can be subjective and many unforeseen nuances cannot be captured. Each topic from the randomly selected subset was evaluated by 3 different participants to account for inter-rater agreement. Participants were instructed to complete each topic sequentially in no particular order. Note that the same interface was used to display both result lists.

Outcomes: We obtained complete results (3 distinct raters for each topic) for 14 out of the 15 topics. The average duration was 25 minutes per topic. Participants emailed their responses individually at the end of the study. They resorted to using either bullet points or paragraphs when explaining their choices. For the evaluation, we consider an approach is superior to the other when the majority of raters vote in its favor. For the sake of readability we de-anonymize the approaches in the participants' responses in the proceeding discussion.

Overall we found that participants preferred HISTDIV's ranking in order to get a historical overview for 10 out of 14 topics. 5 out of the 14 topics had 100% agreement with 4 of those being in favor of HISTDIV. The positive comments from the participants often mentioned good coverage and a better ranking of subtopics within the top 10. Some participants explicitly stated "*good diversity in results*" while others justified their choices with examples of subtopics covered exclusively by the winning approach. HISTDIV's span of coverage was immediately apparent to the participants. One of the topics where HISTDIV out performed IA-SELECT is `Bob Dole`. Participants felt that the increased coverage provided a more rounded picture when compared to IA-SELECT ("*Histdiv mentions an article which talks about the role his wife plays which provides a more complete representation of Bob Dole*") showing that emphasis on recall is justified for this task. `Al Gore` was another such topic - "*HistDiv has articles about the Oscar winning movie but the other doesn't*" - indicating that an important subtopic was uncovered. The same participant also noted that "*Ia-Select has 7 articles from the year 2000 -> not diverse*". This redundancy is due to over-specification of the time period caused by IA-SELECT's lack of temporal awareness.

Even though participants were instructed to judge the quality of overview from the top 10 as a whole there was a tendency to be critical of the ranking. While two participants indicated that chronological ordering in the top 10 would be easier to understand, the others preferred seeing documents from the most important subtopic towards the top. For `Bob Dole`, a participant stated that "*HistDiv is better as it returns his presidential campaign as the top ranked document which highlights the pinnacle of his political career*". For the topic `Oklahoma City`, an expert remarked that the results from HISTDIV were more "*on point*". He argued that while both approaches rightly picked articles about the Oklahoma city bombing at the top, HISTDIV returned the more relevant result from the most relevant time period - "BOMB SUSPECT IS HELD, ANOTHER IDENTIFIED; TOLL HITS 65 AS HOPE FOR SURVIVORS FADES" vs "Oklahoma City, a Year Later". The former headline from HISTDIV is a vital primary source whereas the latter from IA-SELECT is a less relevant secondary source. Here HISTDIV's superior modeling of document utility helps it pick highly relevant documents from the more important historical subtopics first. This difference in importance between subtopics is not captured in the quantitative results due to our classical assumption of equally relevant subtopics.

For the topics where participants agreed that IA-SELECT did better than HISTDIV, lack of precision was cited as the main factor. Seven participants cited the number of relevant articles as the deciding factor which is a valid concern. We found from the responses that there were two types of irrelevant articles: articles completely unrelated to the topic and articles seemingly less relevant than the others that explicitly covered important historical facts. We observed this bias towards precision almost exclusively from the non-experts although for cases like `Landon Donovan` (the former captain of the U.S. mens soccer team) both sets of users agreed that articles about politicians with similar names hurt the overview. Similarly for `Charlie Sheen` participants responded with comments such as "*HistDiv includes a few articles which do not seem to be about Charlie Sheen at all or concern him only very marginally (articles ranked 2,3,7) , hard to get overview if a third of articles are not really about about Charlie Sheen*" upon encountering articles about movie listings or his father Martin Sheen.

An interesting topic that divided opinion was `Rudolph Giuliani`. In HISTDIV's ranking there was an article regarding an interview on Giuliani's personal life that non-experts considered irrelevant. An expert on the other hand acknowledged the presence of less relevant results but mentioned that for certain topics these documents are not as irrelevant as they first seem since you can find valuable contextual information – "*(For Giuliani) HistDiv is better than Ia-Select because it provides a better mix of political and personal information*". This tendency to evaluate the overview from multiple perspectives was also observed in other topics. For the topic `Atlantic City`, a participant stated that "*HistDiv gives a more well argued discussion about the political choices that have been taken in order to reinvent the city*". This shows that the lack of precision is detrimental to user satisfaction although it is highly dependent on the topic. When the inherent aspect diversity of the topic is low, HISTDIV's tendency to increase recall ranks irrelevant documents higher. An interesting direction for future work is to design methods that allow automatic adjusting of parameters to increase precision rather than recall for certain types of queries.

From the observations in both experiments we can conclude that (a) HISTDIV is consistently better at finding primary sources by best diversifying the aspect-time space indicated by high subtopic recall (b) HISTDIV shows promise for pure aspect-based diversification of temporally ambiguous queries (c) In isolated cases users feel that the lack of precision hurts HISTDIV but in the majority of cases the emphasis on recall provides a more holistic overview.

8. CONCLUSION & OUTLOOK

In this paper we introduced the notion of a historical query intent over longitudinal news collections like news archives. We cast the problem as diversification task in a new aspect-time space. To evaluate the task we built a new temporal test collection based on 20 years of the *New York Times* collection. We introduced HISTDIV which shows improvements over temporal and non-temporal methods for most of the time-aware diversification methods. We also outperform all competitors in subtopic recall over the joint space showing the suitability of our approach for historical query intents.

We observe that HistDiv works well for topics which have aspects that span across multiple time intervals and have fluctuating importance at different times. It trades-off nicely between important aspects and important times which we perceive as important in historical search. HistDiv does not perform quite as well for queries which only one dominant aspect at a certain time window. HistDiv also acheives lower precision than some of its competitors although the user study showed that only in isolated cases there is a loss in overview quality. In the future, for more practical settings where training data is little to none, we want to investigate the usage of query related features like the degree of temporal variance of aspects, the number of bursts, etc. to estimate the parameters used. This opens up exciting future work opportunities to automatically identify queries of different historical intents and evaluate them accordingly.

9. REFERENCES

[1] British newspaper archive http://www.britishnewspaperarchive.co.uk/.

[2] New york times archives, http://timesmachine.nytimes.com/browser.

[3] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM* 2009.

[4] O. Alonso, M. Gertz, and R. Baeza-Yates. On the value of temporal information in information retrieval. *SIGIR Forum*, 2007.

[5] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *CIKM* 2009.

[6] A. Anand, S. Bedathur, K. Berberich, and R. Schenkel. Index maintenance for time-travel text search. In *SIGIR* 2012.

[7] K. Berberich and S. Bedathur. Temporal diversification of search results. In *SIGIR Workshop TAIA*, 2013.

[8] K. Berberich, S. Bedathur, O. Alonso, and G. Weikum. A language modeling approach for temporal information needs. In *ECIR*, 2010.

[9] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. In *SIGIR*, 2007.

[10] A. Bingham. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History*, 2010.

[11] M. Brucato and D. Montesi. Metric spaces for temporal information retrieval. In M. d. Rijke, T. Kenter, A. P. d. Vries, C. Zhai, F. d. Jong, K. Radinsky, and K. Hofmann, editors, *Advances in Information Retrieval* 2014.

[12] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt. Survey of temporal information retrieval and related applications. *ACM Comput. Surv.*, 2014.

[13] R. Campos, A. M. Jorge, G. Dias, and C. Nunes. Disambiguating implicit temporal queries by clustering top relevant dates in web snippets. In *WI-IAT*, 2012.

[14] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

[15] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *CIKM*, 2009.

[16] D. O. Case. The collection and use of information by some american historians: a study of motives and methods. *The Library Quarterly*, 1991.

[17] J. Choi and W. B. Croft. Temporal models for microblogs. In *CIKM*, 2012.

[18] C. Clarke, N. Craswell, I. Soboroff, and E. Voorhees. Nist, overview of the trec2011 web track. In *Proceedings of TREC*, 2011.

[19] C. Clarke, N. Craswell, I. Soboroff and A. Ashkan. Nist, overview of the trec2011 web track. In *Proceedings of WSDM*, 2011.

[20] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. Tech. report, DTIC Document, 2009.

[21] M. Costa and M. J. Silva. Evaluating web archive search systems. In *WISE*, 2012.

[22] V. Dang and B. W. Croft. Term level search result diversification. In *SIGIR*, 2013.

[23] V. Dang and W. B. Croft. Diversity by proportionality: An election-based approach to search result diversification. In *SIGIR* 2012.

[24] Clarke, C. A and Kolla, M and Cormack, G and Vechtomova, O. and Ashkan, A. and Büttcher, S. and MacKinnon, I.. Novelty and diversity in information retrieval evaluation. In *SIGIR*, 2008.

[25] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz. Towards recency ranking in web search. In *WSDM*, 2010.

[26] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: Improving recency ranking using twitter data. In *WWW*, 2010.

[27] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *WSDM*, 2011.

[28] W. M. Duff and C. A. Johnson. Accidentally found on purpose: information-seeking behavior of historians in archives. *The Library Quarterly*, 2002.

[29] D. Gupta and K. Berberich. Identifying time intervals of interest to queries. In *CIKM*, 2014.

[30] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *EMNLP*, 2011.

[31] H. Joho, A. Jatowt, and R. Blanco. Ntcir temporalia: a test collection for temporal information access research. In *companion publication of the 23rd WWW*, 2014.

[32] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3), July 2007.

[33] X. Li and W. B. Croft. Time-based language models. In *CIKM*, 2003.

[34] S. Liang, Z. Ren, and M. de Rijke. Fusion helps diversification. In *SIGIR*, 2014.

[35] D. Metzler, R. Jones, F. Peng, and R. Zhang. Improving search relevance for implicitly temporal queries. In *SIGIR*, 2009.

[36] T. N. Nguyen and N. Kanhabua. Leveraging dynamic query subtopics for time-aware search result diversification. In *ECIR*, 2014.

[37] M.-H. Peetz, E. Meij, M. de Rijke, and W. Weerkamp. Adaptive temporal query modeling. In *Advances in Information Retrieval*, 2012.

[38] K. Radinsky, K. M. Svore, S. T. Dumais, M. Shokouhi, J. Teevan, A. Bocharov, and E. Horvitz. Behavioral dynamics on the web: Learning, modeling, and prediction. *TOIS*'13.

[39] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *WWW*, 2010.

[40] V. Setty, S. Bedathur, K. Berberich, and G. Weikum. Inzeit: Efficiently identifying insightful time points. *In VLDB Endow.*, 2010.

[41] C. Johnson and W. Duff. Chatting up the archivist: Social capital and the archival researcher. *The American Archivist*, 2005.

[42] H. R. Tibbo. The philosophy of information retrieval evaluation *Evaluation of cross-language information retrieval systems*, 2002.

[43] H. R. Tibbo. Primarily history in america: How us historians search for primary materials at the dawn of the digital age. *The American Archivist*, 2003.

[44] K. Zhou, S. Whiting, J. M. Jose, and M. Lalmas. The impact of temporal intent variability on diversity evaluation. In *Advances in Information Retrieval*, 2013.

[45] Y. Zhu, Y. Lan, J. Guo, X. Cheng, and S. Niu. Learning for search result diversification. In *SIGIR*'14.

[46] M. Peetz, M. Edgar and M. de Rijke. Using temporal bursts for query modeling. In the *Springer Journal of Information Retrieval*, 2014.