EXPOSÉ: EXploring Past news for Seminal Events

Arunav Mishra
Max Planck Institute for Informatics
Saarbrücken, Germany
amishra@mpi-inf.mpg.de

Klaus Berberich
Max Planck Institute for Informatics
Saarbrücken, Germany
kberberi@mpi-inf.mpg.de

ABSTRACT

Recent increases in digitization and archiving efforts on news data have led to overwhelming amounts of online information for general users, thus making it difficult for them to retrospect on past events. One dimension along which past events can be effectively organized is time. Motivated by this idea, we introduce **EXPOSÉ**, an exploratory search system that explicitly uses temporal information associated with events to link different kinds of information sources for effective exploration of past events. In this demonstration, we use Wikipedia and news articles as two orthogonal sources. Wikipedia is viewed as an event directory that systematically lists seminal events in a year; news articles are viewed as a source of detailed information on each of these events. To this end, our demo includes several time-aware retrieval approaches, that a user can employ for retrieving relevant news articles, as well as a timeline tool for temporal analysis and entity-based facets for filtering results.

Categories and Subject Descriptors

H.3.3 [**Information Search & Retrieval**]: Search process

Keywords

Exploratory Search; Temporal Information Retrieval; Linking; Past Events; Wikipedia

1. INTRODUCTION

In today's digital age, news media has witnessed a drastic evolution with the Internet as of its key elements. Digitization of both recent and archived news and their availability as portals, online newspapers, and blogs facilitate affordable, effective, and efficient dissemination of news to a wider audience. On one hand, this ease of access to information presents new opportunities to retrospect on past news events and, on the other hand, overwhelmingly large amounts of information make it difficult to do so.

WWW '15 May 18-22, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742844 .

Recently, many alternatives have emerged as prominent sources of information on events. One such source that is particular interesting is the free encyclopedia Wikipedia. Among many initiatives taken to enrich the collaboratively authored Wikipedia, the WikiProject Years[1] organizes events into year, decade, century, and millenium articles. An event in these articles is characterized with a short *textual description* and a temporal expression referring to a specific *date* indicating when it happened. Consider as a concrete example, the article `http://en.wikipedia.org/wiki/1987`, which lists the following as seminal event in that year,

> *October 11 1987* : The first National Coming Out Day is held in celebration of the second National March on Washington for Lesbian and Gay Rights.

To design an effective exploratory search system for seminal past events, one could think of connecting the Wikipedia events to relevant news articles. With these badly-missing connections in place, a user with an unclear information need can start by browsing through the Wikipedia lists organized collectively as a *directory of events* and jump to connected news articles for more fine-grained details. In addition, flexibility to submit typed description and dates extends the scope of exploration to unlisted events or enables the user to simply refine a listed event. Further, visualization tools like facets and timelines offer valuable learning assistance while exploring past events.

Time becomes an interesting dimension of past events and can help to effectively explore their ramifications [9]. However, incorporating time into exploration is not straightforward. General users may refer to a specific event date to explore news articles by having different temporal intent. For example, a user could choose to see news articles that were published close to the indicated date, or be interested in news articles that discuss the dates in their content, or a combination of both. Decoupling text and time for exploration of temporal events and giving explicit handles for each would result in more effective and efficient exploration. Consider the following use cases for further motivation:

- Tony Stark, a computer scientist, wants to understand the ramifications of the *Dot-com bubble burst in January 2000*. For this, he intends to explore all articles that are published around the event time period or

[1] `http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Years`

discuss the time period in their content. In this regard, he sets his preference to view relevant news for exploration.

- Melita Garner, a news curator, researching on *1987 Maryland train collision* wants to better understand how the story presented by the media unfolded as the event had happened. For this, she wants to explore news article that are published around the event time period and sets her preference accordingly.

Both users performing exploratory search over news have different temporal intents. Motivated by the above scenarios, our system offers five modes of temporal exploration: **1)** *relevant news* – automatically generates interesting time period for an event and identifies relevant and contemporary news articles that discuss the date in their content; **2)** *news published around the event date* – uses the event date to identify relevant and contemporary news articles; **3)** *news that mention the event date* – identifies relevant news articles that discuss the event date in their content; **4)** *news published around the event date or mention the event date* – identifies relevant and contemporary news articles that also discuss the event date in their content; **5)** *news by a standard text search* – identifies relevant news articles with the given textual description and completely ignores the event date. Depending on information needs, a user may choose to use one of the above modes to explore past events. Each mode uses a different temporal retrieval method in the background that we discuss in the next section.

State-of-the-art vertical search engines like Google news that provide only lookup services do not help in exploration [3]. To cater to the temporal needs, they provide simple filtering on publication dates and do not explicitly regard the temporal content of news articles. In addition, an event description issued as a keyword query to such systems suffers from verbosity which leads to topic drifts. However, our system is based on retrieval methods developed in [4] that address these challenges.

For time-aware exploratory search, Odijk et al. [5, 6] design interfaces to visualize document temporal distribution and word clouds. These visualizations are useful for analyzing word meanings and usage change over time but do not seem to work for event exploration. As an extension, Reinanda et al. [7] presented a system to explore entity associations in longitudinal document collections. Though interesting, this system becomes difficult to extend to events.

Contributions. We present a prototype time-aware exploratory search system for exploring events that were seminal in *years 1987 to 2007*. We implement methods developed in our prior work [4] on time-aware retrieval models as different modes of exploration.

Organization. In Section 2, we point out the advantages of our retrieval methods with supporting examples. Section 3 discuss the details of our prototype along with a demonstration scenario.

2. ALGORITHMIC BUILDING BLOCKS

In this section we describe five approaches with example queries for which they become most effective. For full details on individual approaches and an evaluation of their performances, we refer to our prior work [4]. Our EXPOSÉ system encodes each of the approaches as a search mode to make exploration with different temporal intents possible.

To develop our time-aware approaches, we adopt a simplified representation for temporal expression as given in Berberich et al. [2]. Temporal expressions as modeled simply as time intervals $[b, e] \in T \times T$ with begin time point b and end time point e with the assumption $b \leq e$. Any given event, like the one in our previous example, is treated as a query with two parts, q_{time} that is generated from the date; and q_{text} that is generated from textual description. Analogously, we represent a document (news article) d with publication date t as a combination of a bag of temporal expressions d_{time} and a bag of textual terms d_{text}. Sometimes, we refer to the entire collection as a coalesced document D with its textual part as D_{text} and temporal part as D_{time}.

Text-Only Approach. As a simple approach, we compare the query description q_{text} to the textual terms in a document d_{text}. For this, we use a query-likelihood approach with Dirichlet smoothing that ranks documents according to

$$P(q_{text} \mid d_{text}) = \prod_{v \in q_{text}} \frac{P(v \mid d_{text}) + \mu \cdot P(v \mid D_{text})}{|d_{text}| + \mu} \quad (1)$$

where μ is the document-length based Dirichlet prior.

Intuitively, the text-only approach is most effective for short event descriptions with discriminative terms. Further, this approach is also useful when a user does not have any information on the event time period. Consider the following event as an example where this approach work well.

December 24 1992 : President George H. W. Bush pardons 6 national security officials implicated in the Iran-Contra affair, including Caspar Weinberger.

The example above refers to a specific event about the Iran-Contra scandal (of 1986) that happened during the final days of the presidency of George H. W. Bush. The relatively short textual description contains many terms as surface form of entities that makes it selective. This example retrieves relevant articles in the top-10 with our simple text-only approach.

Publication-Date Approach. Newsworthy events receive quick media attention with articles reporting on the details as the events happen. Intuitively, the news articles that are published around the event date are more likely to discuss the event in more detail. This approach ranks documents by relating their publication dates to the event date by assuming recency as the user's temporal information intent. That is, a user is satisfied if the publication date of a news article is close to the event date and quickly becomes dissatisfied if the two dates grow apart.

To rank documents, the publication-date approach first assumes independence between the temporal and the textual part of a given query. It then estimates the likelihood of generating the query from the document as

$$P(q \mid d^t) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid t) \quad (2)$$

where the first factor is the probability of generating the q_{text} from d_{text} and is estimated according to Equation 1. The second factor is the probability of generating the publication date t from q_{time} estimated as

$$P(q_{time} \mid t) = \frac{1}{1 + e^{\,r|q_{time} - t|}} \; . \quad (3)$$

This approach is most effective for events with short ramifications. As an example query for which this approach becomes effective, consider the following

September 16 2000 : Winnie Mandela is convicted of kidnapping. On May 14, she is sentenced to 6 years in prison.

This example can be considered as a facet of a larger event that spans from the kidnapping to trial. Here, the date becomes a strong signal to identify relevant news articles.

Temporal-Expression Approach. News articles that are not contemporary to an event may still be relevant if they provide information pertaining to the event. Such articles often refer to the time period of the event in their content. Motivated by this, the temporal-expressions approach ranks an article higher if many temporal expressions in its content refer to the time period when the event happened.

To rank documents, our approach first assumes independence between the textual and temporal part of a query and ranks documents according to

$$P(q \mid d^t) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid d_{time}) \quad (4)$$

where the first factor is estimated as per Equation 1. The second factor is the probability of generating the q_{time} from the temporal part of the document d_{time} as

$$P(q_{time} \mid d_{time}) = \frac{1}{|d_{time}|} \sum_{[b,\, e] \in d_{time}} \frac{\mathbb{1}(q_{time} \in [b,\, e])}{e - b + 1} . \quad (5)$$

This approach is most effective for events that did not receive adequate media coverage and have few relevant contemporary articles. Consider the following example,

September 16 2000 : Ukrainian journalist Georgiy Gongadze is last seen alive; this day is taken as the commemoration date of his death.

In this example, though journalist Georgiy Gongadze was last seen alive on the given date, his death was confirmed months later. Thus, we find that news articles that elaborate on his death are published much later to the event date, however mention this time period in their content.

Publication-Date + Temporal-Expression Approach. This approach combines the publication-date and temporal-expression approaches to retrieve relevant news articles as

$$P(q \mid d^t) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid t) \cdot P(q_{time} \mid d_{time}) \quad (6)$$

Intuitively, this approach performs best for events that have larger ramifications or are discussed in the media at diverse time points. This approach is effective for ambiguous event descriptions. Consider the following example.

February 27 1991 : President Bush declares victory over Iraq and orders a cease-fire.

The description shown above contains references to a President Bush and a war in Iraq which are ambiguous since there were multiple presidents named Bush and multiple wars in Iraq. For this, any news article that is retrieved by syntactic matching, and has publication date close to the event date or mentions the event time period turns out to be relevant.

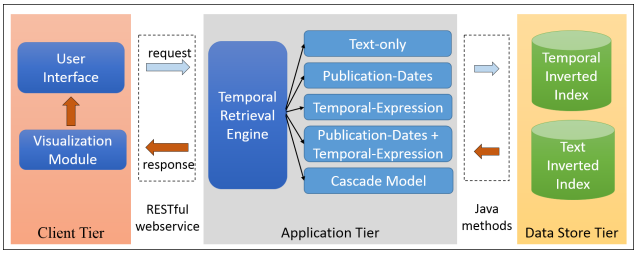

Figure 1: System architecture of EXPOSÉ

Two-Stage Cascade Approach. The temporal part of a given query refers to a single day and becomes overly specific to indicate when an event happened. This is often seen for events that spanned over a larger time period. The two-stage cascade approach aims at automatically identifying additional time points (days) that are seminal to the event. As a key difference to other approaches, the two-stage cascade approach uses relevance feedback to expand the q_{time} and shift the focus to a larger time period.

In *Stage 1*, it performs an initial round of retrieval with the text-only approach as shown in Equation 1, and estimates a temporal query model using the temporal expressions in the retrieved articles, thus treating them as pseudo-relevant. We estimate the probability of generating a time point τ from query model Q_{time} as

$$P(\tau \mid Q_{time}) = \sum_{d \in R_k} \frac{P(q_{text} \mid d_{text})}{\sum_{d' \in R_k} P(q_{text} \mid d'_{text})} \cdot P(\tau \mid d_{time}) \quad (7)$$

where the first factor is the normalized text score of the document. The second factor is the likelihood of generating the time point from a temporal part of document and is estimated as per Equation 5.

In *Stage 2*, our approach re-ranks the initially retrieved candidate articles by taking into account the Kullback-Leibler (KL) divergence between the query and document temporal models. To generate the final document score we additionally combine the divergence between the originally given event date and publication date of the document, and the text parts of the query and document to preserve the textual relevance. Formally, this is denoted as

$$\alpha \cdot KLD(Q_{time} \| d_{time}) + \beta \cdot KLD(q_{time} \| t) + \gamma \cdot KLD(q_{text} \| d_{text})$$

where α, β, and γ are tunable parameters.

This approach is most effective for events that did not receive extensive coverage when they happened or that spanned over a long time period. Consider the following example,

March 19 2002 : US war in Afghanistan: Operation Anaconda ends after killing 500 Taliban and Al-Qaeda fighters, with 11 allied troop fatalities.

The example above refers to the end of Operation Anaconda which itself lasted for a longer time period. For this, building a temporal model, with feedback expressions as salient time points, shifts the focus from the overly specific event date to a larger time period.

3. DEMONSTRATION

Implementation. The responsive web interface is implemented using the Twitter Bootstrap toolkit. The server side

Figure 2: Preference options

is implemented in Java and deployed in Tomcat server. We use the Stanford CoreNLP to tokenize documents, extract entities and annotate temporal expressions in their content.

News Collection. For the prototype, we use The New York Times Annotated Corpus [1], which contains about 2 million documents published between 1987 and 2007.

Directory Events. We use The English Wikipedia dump released on February 3rd 2014 and generate 3076 Wikipedia events as queries from the articles for the years 1987 to 2007. These events are provided in the prototype as a directory.

System Architecture of EXPOSÉ is illustrated in Fig 1. A user submits a preference option, an event description, and a date as a RESTful web-service request to temporal retrieval engine. This engine implements five retrieval methods, one of which is selected by the user. The retrieval engine uses inverted text and temporal indexes to efficiently retrieve and rank relevant news articles. The visualization module receives a ranked list of top-10 news articles as a response. Finally, the visualization module performs facet extraction and updates the user interface.

Exploratory Interface. The Web interface exhibits a three column layout that we describe next.

The left-most column holds two components: first, a Search Box panel where a user can type in an event description and a date; second, an Event Directory panel that contains 3,076 clickable events extracted from Wikipedia. To select a particular mode for time-aware exploration, a user can choose one the following options in Preference tab as illustrated in Fig 2: **1)** *Relevant news* – is the default option and uses the two-stage cascade approach; **2)** *News published around the event date* – regards event date and uses publication-date approach; **3)** *News that mention the event date* – considers the temporal expressions in article content and uses the temporal-expression approach; **4)** *News published around the event date or mention the event date* – regards both temporal expressions and the publication date, and uses the publication-date + temporal-expression approach; **5)** *News by a standard text search* – activates the text-only approach.

The middle column contains the Search Results panel that displays the relevant news articles after processing the user input. For each identified news article, its title and the publication date are displayed as a ranked list.

A user can then click on a title to read its content which opens in the News Article panel aligned to the right-most column.

The Timeline panel visualizes the publication dates distribution of the retrieved news articles for a given event. The timeline plots titles of the news articles which when clicked opens the corresponding content in the News Article panel. Assuming that the distribution represents salient time points for the event, this visualization for publication date-based methods helps to explore the temporal evolution of the event. For methods exploiting temporal expressions, this visualization is useful to identify related events whose corresponding news articles refer to the current event in their content. This feature can be activated or deactivated in the timeline panel by toggling the Timeline button which is shown in Fig 2.

The Facet Panel displays three sets of entity labels semantically representing persons, locations, and organizations that are extracted from relevant news articles. The aggregated term frequency of each label is also displayed which gives an understanding of their prominence. The three sets are organized into sub panels, and can be selected in a desired combination to dynamically filter news articles. This feature can be activated or deactivated in the facet panel by toggling the Facet button as shown in Fig 2.

Demonstration Scenario. Let us consider the example event in Section 1. To explore, a user begins by looking for this event in the directory using a directory search (or simply types in the search box). By setting the preference to *News published around the event date* option, she quickly identifies relevant contemporary news articles that elaborate on the event. Further, the user uses a timeline tool in the system to visualize and learn salient time periods when this event was discussed in the media, like *November 1987* and *April 1988*. Entities like *Pope John Paul* (of type person), *United States of America* (of type geographic location) that are mentioned in the identified articles are extracted as facets for further filtering. For further exploration, the user also has the flexibility to submit a typed description of a related event that is not listed in the directory, or is simply a refinement for an existing description.

A video showcasing EXPOSÉ is available at:

`http://youtu.be/t4frzvvonqE`

4. REFERENCES

[1] The New York Times Annotated Corpus `http://corpus.nytimes.com`.

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

[3] G. Marchionini. Exploratory search: From finding to understanding. *Commun. ACM*, 49(4), 2006.

[4] A. Mishra, D. Milchevski, and K. Berberich. Linking Wikipedia Events to Past News. In *TAIA*, 2014.

[5] D. Odijk, O. de Rooij, M.-H. Peetz, T. Pieters, M. de Rijke, and S. Snelders. Semantic document selection. In *TPDL*, 2012.

[6] D. Odijk, G. Santucci, M. de Rijke, M. Angelini, and G. L. Granato. Time-aware exploratory search: Exploring word meaning through time. In *TAIA*, 2012.

[7] R. Reinanda, D. Odijk, and M. De Rijke. Exploring entity associations over time. In *TAIA*, 2013.

[8] C. Zhai. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1, 2008.

[9] J. Strötgen, O. Alonso, M. Gertz. Identification of Top Relevant Temporal Expressions in Documents. In, *TWAW*, 2012.