Refining Imprecise Spatio-temporal Events: A Network-based Approach

Andreas Spitz
Institute of Computer Science
Heidelberg University
spitz@informatik.uni-heidelberg.de

Johanna Geiß
Institute of Computer Science
Heidelberg University
geiss@informatik.uni-heidelberg.de

Michael Gertz
Institute of Computer Science
Heidelberg University
gertz@informatik.uni-heidelberg.de

Stefan Hagedorn
Dept. of Computer Science
TU Ilmenau
stefan.hagedorn@tu-ilmenau.de

Kai-Uwe Sattler
Dept. of Computer Science
TU Ilmenau
kus@tu-ilmenau.de

ABSTRACT

Events as composites of temporal, spatial and actor information are a central object of interest in many information retrieval (IR) scenarios. There are several challenges to such event-centric IR, which range from the detection and extraction of geographic, temporal and actor mentions in documents to the construction of event descriptions as triples of locations, dates, and actors that can support event query scenarios. For the latter challenge, existing approaches fall short when dealing with imprecise event components. For example, if the exact location or date is unknown, existing IR methods are often unaware of different granularity levels and the conceptual proximity of dates or locations.

To address these problems, we present a framework that efficiently answers *imprecise event queries*, whose geographic or temporal component is given only at a coarse granularity level. Our approach utilizes a network-based event model that includes location, date, and actor components that are extracted from large document collections. Instances of entity and event mentions in the network are weighted based on both their frequency of occurrence and textual distance to reflect semantic relatedness. We demonstrate the utility and flexibility of our approach for evaluating imprecise event queries based on a large collection of events extracted from the English Wikipedia for a ground truth of news events.

CCS Concepts

•Information systems → Spatial-temporal systems; Content analysis and feature selection; Information retrieval query processing;

This is the author's version of the work. It is posted here for your personal use, not for redistribution. The definitive version is published in:

GIR '16, October 31-November 03 2016, Burlingame, CA, USA

© 2016 ACM. ISBN 978-1-4503-4588-0/16/10...$15.00

DOI: http://dx.doi.org/10.1145/3003464.3003469

Keywords

Events; event representation; spatio-temporal information; information networks

1. INTRODUCTION

"Did Jimi Hendrix play in Munich in 1967? Or did he play in Berlin? When was it exactly? Did some other people play together with him at his gigs?". When historic events are concerned, people rarely remember the exact date or the exact location. Instead, we typically have a rough idea about the general location or the general timeframe, but only at a coarse-grained level. For example, *"I know that Hendrix played in Germany around 1970, but I don't recall where and when exactly"*. In such cases, the resulting information retrieval task is non-trivial and cumbersome to formulate using standard search engines. The reason being that search engines are to a large extent still unaware of granular hierarchies that allow for refinement, similar to the concept of query expansion (such as location and date information in the above example). For instance, instead of using the location "Germany" we could substitute different cities in Germany as part of the query, or even different nearby years, such as 1966 or 1968 instead of 1967. However, manually searching through numerous documents that are returned for such a coarse-grained query to find the information of interest constitutes a tedious process.

Information retrieval scenarios as outlined above are centred on the notion of an event. in this paper, adopting the definition from the Topic Detection and Tracking (TDT) task [5], we view an event as *something that happens at a given place and time between a group of actors*. Thus, an event has three core components: a geographic location, a date, and an actor. Research on event detection has a long tradition, ranging from the extraction and tracking of (new) events from news articles (e.g., [6, 8]) to more recent work on event detection in social media streams (e.g., [1, 4, 22]). Interestingly, these works are primarily concerned with the extraction of events and event information, but less with the question of how to use such extracted information for expressive event queries and event exploration tasks. A number of

studies consider visualization techniques for exploring microblogs in relation to events [10, 17], but more comprehensive frameworks for exploring repositories of event-related information in the context of IR tasks are still missing.

In this paper, we combine co-occurrence information of location, date, and actor mentions in documents with a network-based representation of these core components for the description and weighting of events, resulting in a so-called *event network*. Three base networks form the backbone of an event network: (1) a location network that models the spatial containment of geographic entities and geographic neighborhood information, (2) a date network that simply models the temporal containment of dates, and (3) an actor network that describes semantic relationships among actors, again based on co-occurrences. For the latter type of network, we employ the Wikipedia Social Network described by Geiß et al. [13]. The date network is created from scratch and models temporal containment information for months and days for an appropriate range of years. The location network is derived from publicly available geospatial data repositories and includes the spatial containment of geographic entities such as countries, states, and cities. Furthermore, information about (k-)nearest neighbors for geographic entities at different levels of granularity is included. These backbone networks are connected through location, date, actor triples such that each triple forms an *event*. The individual entity mentions as components of triples are extracted from documents and combined based on their positional neighborhood in documents. For example, the more often a triple can be found within a particular range of sentences in the documents of a collection, the more indicative this triple is for an event, thus receiving a higher weight.

In our framework, we model event triples in the form of hyperedges that connect entities from the three backbone networks, eventually forming an event network that serves as basis for our approaches to refine and explore events. Compared to existing work on spatio-temporal IR, this network not only provides a more comprehensive representation of how events are related in terms of geographic location, date, and actors involved, but it also allows for efficient approaches to answering precise and imprecise event queries. The result of an (imprecise) event query is then a ranked list of events, where each event can be explored both in terms of its relationships to other locations, dates, and actors as well as how the event is represented in the different documents that contain mentions of the respective event. We argue that such an approach is more meaningful and flexible for querying and exploring large collections of event data than existing approaches. The representation of events and their components as a network then also enables the use of network-based measures in the analysis of the data.

Contributions. In summary, we make the following contributions: (i) we present a novel framework for the representation of events and their components (geographic location, date, and actors) in the form of networks, (ii) we describe an efficient approach to answer (imprecise) event queries, resulting in a ranked list of events extracted from the event network, and (iii) we demonstrate the utility and effectiveness of our framework using a large corpus of events extracted from the English Wikipedia, which we also make available as a community resource[1].

[1]Event collection and background networks are available at: `http://dbs.ifi.uni-heidelberg.de/index.php?id=data`

Structure. The remainder of the paper is structured as follows. In Section 2, we discuss prior and related work. In Section 3, we then describe the different types of networks and how these networks are used to refine events and evaluate (imprecise) event queries. In Section 4, we present an experimental evaluation of our approach based on Wikipedia (for the construction of the networks) and external sources for event data. In Section 5, we summarize our approach and outline ongoing work.

2. BACKGROUND AND RELATED WORK

The framework presented in this paper builds on state-of-the-art methods and techniques for named entity recognition (NER), with a particular focus on the detection, extraction, and normalization of mentions of locations, dates, and persons in documents. In the following, we briefly elaborate on the respective NER techniques we use in our approach and then focus on related work in the context of (network-based) geographic and spatio-temporal IR.

Named Entity Recognition. Key to the extraction of events from documents is the detection of location, date, and actor mentions in the documents. For this, most existing approaches build their respective frameworks and methods on top of tools such as the StanfordNER [11] to detect named entities. In order to disambiguate and normalize respective mentions of locations, dates, and actors in text, subsequent techniques have to be applied. Thus, these two steps already raise several issues in terms of the quality and correctness of the extracted entities. To avoid these problems and to build our framework for querying events on high-quality named entities, we choose a different approach that is specific to Wikipedia and Wikidata as described in Section 4.

For information about geographic entities such as continents, countries and cities, we make use of the location network extracted from Wikidata, an effective approach to obtain a high-quality repository of geographic information, as described by Geiß and colleagues [14, 27]. Information about location entities is modeled in the form of a location network, which is one of three backbone networks we employ in our approach (introduced in Section 3). Similar to the location network, we employ the Wikipedia Social Network [13] as another backbone network for actor information. In this network, nodes represent persons (as found in Wikidata) and weighted edges reflect the frequency and positional distance of co-occurrences of actor pairs in a collection of documents, here the English Wikipedia. Finally, for the detection, extraction, and normalization of temporal expressions, we make use of HeidelTime [30]. In accordance with the location and actor network, we also use a generic date network, which completes the set of three backbone networks for our approach. Here, we go beyond most existing approaches to event detection and exploration that do not consider information about actors that are likely involved in an event but typically only describe it by a location and a date. These three backbone networks thus serve as lookup structures for location, date and person mentions in the documents from which we want to extract events for subsequent querying and exploration. In Section 4, we describe in more detail how these networks are constructed.

Spatio-temporal IR and Event Exploration. Once events have been determined for a collection of documents or for a stream of text data (such as Tweets), different approaches for querying, analyzing, and exploring events can

be deployed. Recently, several methods have been proposed to query and explore events in streams of microblogs. Probably the most prominent approach for detecting events in Twitter has been described by Sakaki et al. [22]. More recently, Abdelhaq et al. proposed an approach to detect localized events from Twitter, focusing on the spatial extent and description of small-scale, geographically localized events based on keywords [1]. An approach presented by Feng et al. also focuses on the detection of events from Twitter based on hashtags and also provides means to explore such hashtag clusters at different levels of spatial granularity over time. Similarly, the framework proposed by Marcus et al. [17] focuses on the (timeline-based) visualization and summarization of events on Twitter. While the above works provide rich functionality for detecting events in data streams, mostly based on the burst of keywords or in terms of activity, they do not provide any IR-style functionality to search for events, rank events based on different criteria or to correlate events.

The following methods are closer to an IR-style approach to querying and exploring events. For example, Abujabal and Berberich propose an approach to extract events from semantically annotated document collections [2]. In contrast to our approach, they construct individual events at the sentence level and do not consider events that are mentioned across several documents. Similarly, Kanhabua and Nejdl [16] restrict the detection of events as positional co-occurrences of entities and locations to those within single documents and single sentences. Adams and Gahegan explore purely the temporal and spatial dimensions in a survey of chronotopes in large document collections such as Wikipedia and observe co-occurrence patterns of temporal and geographic expressions [3]. Julinda et al. propose an approach to build a repository of events from news articles, mostly focusing on sentences that likely describe the same event [15]. In their approach, they do not consider geographic and actor information. Ceroni et al. focus on the evaluation of candidate documents for event extraction based on the entities that are involved in an event, but do not actively extract events themselves [9]. In a similar approach, Schmidt et al. present a system for page recommendations that uses entity-based search and auto-completion of possible involved entities that can be applied to event search [24]. An interesting approach for learning how to extract (local) events from web pages has recently been proposed by Foley et al. [12]. However, they consider neither IR-tasks on the extracted events nor the exploration of a repository of extracted events. In the approach presented by Mishra et al. for linking Wikipedia events to past news [19], temporal information is considered in querying for events, but no geographic or actor information is used. In a subsequent investigation, the approach is extended to include geographic and entity information [18], which equates events with entire Wikipedia pages and thus provides a view on events that is designed to be more coarse-grained than our approach.

Works that are more closely related to our approach include Nepomnyachiy et al. [20]. Based on geo-temporal stamped documents (Tweets), they provide an efficient way of searching for documents that satisfy geographic and temporal range queries. Different from our approach to explicitly deal with imprecise event specifications (including actors), they instead focus on different aspects, in particular the textual components (e.g., terms used for named events) associated with the location and date mentions in documents. Wang and Steward consider the use of spatial and temporal information in the context of events describing natural hazards in news reports. While they describe an interesting approach to integrating a hazard ontology with gazetteers, they do not describe IR-style tasks on such a framework. Quite similar to our approach is the method proposed by Strötgen and Gertz [31] as an extension to their earlier approach for querying events [29]. They present a model to rank documents according to combined textual, temporal, and geographic queries by eliminating the independence assumption between the query components through calculating proximity scores. This is similar to our approach, which assigns a weight to an event triple (location, date, actor) based on positional co-occurrence over a collection of documents. In their approach, the imprecision has to be specified explicitly in the form of time intervals and query regions, thus requiring more knowledge from the user and also not providing any event exploration capabilities.

Graph-based IR and Event Exploration. Some of the works that are most closely related to our approach use models based on graph representations of documents to facilitate IR tasks. Spitz and Gertz introduce the LOAD model for representing and browsing networks of named entities and events that are implicitly contained in large document collections [28]. In contrast to our approach, they do not extract events as triangular structures but as composites of individual relationships between entities with a focus on exploration. Das Sarma and colleagues propose an entity dynamic relation graph to determine entities that participate in (trending) events, but they do not consider the geographic aspect [23]. Blanco and Lioma use a graph-based approach that models terms as nodes in a graph and derives a strength of connection between them based on co-occurrence counts to enable document ranking, but do not include named entities [7]. Similarly, Rousseau and Vazirgiannis use directed graphs without weights to account for term order in text representation [21]. In contrast to our work, they focus on the sentence level and do not include entities, thus limiting the possibilities of this model in event exploration.

To the best of our knowledge, none of the existing approaches deal with imprecise event queries or offer efficient evaluation methods for such types of queries. Here, we thus present a new way of modelling the co-occurrences of named entities in event mentions that allows for a flexible refinement of diverse types of (imprecise) event queries.

3. EVENT REFINEMENT

In the following, we introduce our framework for the refinement of spatio-temporal events. The basis of our approach are four networks: three background networks, detailed in Section 3.1, and an event network, which is described in Section 3.2. In Section 3.3, we then describe how a list of ranked event instances for a given imprecise event query is determined using the above networks.

3.1 Background Networks

As motivated in the introduction, our approach is based on the assumption that imprecise events can be refined with prior knowledge. This knowledge includes information about events, more precisely information about how instances of locations, dates, and persons co-occur in a collection of documents and thus likely describe an event. Such co-occurrences

are determined from text corpora, such as Wikipedia or news articles, and are represented in the form of a network. Before we introduce the event network, we first describe three networks that provide background information about dates, locations, and actors.

Location Network. Geographic expressions contained in the text of documents can be of different granularities. Here, we consider locations of the types city, country, and continent. Assuming that all instances of these types have a spatial extent, a hierarchy can be formed based on spatial containment, with instances of type city as the finest granularity. Let $L := L_{con} \cup L_{co} \cup L_{ci}$ denote the set of different geographic objects of types continent, country, and city, respectively. If an object l is spatially covered by an object l', we denote this as $l \sqsubseteq l'$. For example, the following holds: `germany` $\sqsubseteq$ `europe`, `munich` $\sqsubseteq$ `germany`, and `munich` $\sqsubseteq$ `europe`. It is of course possible to alter or extend the location hierarchy as needed, for example, by including states.

Definition 1. (Location Network) Given a set L of locations of different granularities. A location network $G_L = (V_L, E_L)$ for L is determined by the node set $V_L = L$ and the set E_L of directed edges defined as $E_L := \{(l_2, l_1) \mid l_1, l_2 \in V_L, l_1 \neq l_2, \text{ and } l_1 \sqsubseteq l_2\}$.

Thus, edges are directed, going from a location of coarse granularity to a location of finer granularity. We assume that with each location entity $l \in V_L$ a list of k nearest neighbors of the same type is associated. For example, for a country, these could be the country's direct neighbors, or for a city, these could be the k nearest cities or the cities in some radius around l. We denote the list of k nearest neighbors for a location entity l as $N_k(l)$. In Section 4, we describe how such a network is in fact realized based on real geographic data.

Date Network. Similar to geographic expressions, temporal expressions can be of different granularities. Here, we consider expressions of the types date, month, and year. Further types such as week and quarter can be included, in general forming a (parallel) inclusion hierarchy among temporal types. Assuming the type day being of the finest granularity, each instance of a coarse date type includes a list of instances of finer date types. For example, the instance `2015` of type year includes twelve instances of the type month (`2015-01`, ..., `2015-12`) and 365 instances of the type day. The sets of different dates of different types are denoted T_y, T_m, and T_d for years, months, and days, respectively, with $T := T_y \cup T_m \cup T_d$. As each date in T_y and T_m can be viewed as an interval with the begin and end date being in T_d, the fact that a date t is temporally covered by a date t' is denoted as $t \sqsubseteq t'$. For example, we have `2015-03` $\sqsubseteq$ `2015`, `2015-03-01` $\sqsubseteq$ `2015-03`, and `2015-03-01` $\sqsubseteq$ `2015`. From a conceptual point of view, we assume a directed *date network* that is defined as follows:

Definition 2. (Date Network) Given a set T of dates of different granularities. A date network $G_T = (V_T, E_T)$ for T is determined by the node set $V_T = T$ and the set E_T of directed edges defined as $E_T := \{(t_2, t_1) \mid t_1, t_2 \in V_T, t_1 \neq t_2, \text{ and } t_1 \sqsubseteq t_2\}$.

Thus, edges are directed, going from a date of coarse granularity to a date of finer granularity. The network structure enables us to represent hierarchical information even in a very heterogeneous setting. For example, including temporal expression such as `Fall 2015` as a node in the network in addition to days, months and years is a simple matter if such information is available. In the following, we assume that the date network exists for a range of years, thus each year results in a connected component in G_T and each component is represented as an acyclic graph.

In our framework, we later need information about dates that are close to a given date. For example, for a given month, the k months that precede and succeed it. For each node $t \in V_T$, dates of the same type as t are determined based on the temporal distance, and with each node t a list of such k nearest dates is associated, denoted as $N_k(t)$.

Actor Network. The third background network used in our approach is the actor network, denoted $G_A = (V_A, E_A)$. It corresponds to the social network extracted from Wikipedia [13] and is based on co-occurrences of person mentions in Wikipedia documents. Person entities form the node set V_A of the actor network and undirected, weighted edges E_A describe the strength of co-occurrences between actors in the Wikipedia documents. The more often two person names co-occur in documents and the closer their positional co-occurrence in the documents, the higher the weight associated with the edge between the two actors. Due to the structure of the network, the neighborhood of an actor $a \in V_A$, e.g., the k nearest neighbors, can easily be determined using breadth-first search starting at a.

3.2 Events and Event Network

Assume a corpus $D = \{d_1, \ldots, d_n\}$ of documents. In a first step, named entity recognition is performed on D. That is, all date expressions T_D, location expressions L_D, and actor names A_D are detected, extracted and normalized for each document in D. We assume that $T_D \subseteq T$, $L_D \subseteq L$, and $A_D \subseteq A$, that is, all named entities found in D can also be found in the respective background network.

For a given document $d_j \in D$, let T_d, L_d, and A_d denote the entities detected in d_j. Furthermore, let $dist(r, r')$ denote the positional (sentence) distance of any two such entities in d_j. If two entities occur in the same sentence, then they have distance 0. From these entities, event instances are formed. Event instances $e_i = (l, t, a)$ are indexed by i, with $t \in T_d, l \in L_d$, and $a \in A_d$ and are constructed in the following way: if the pairwise distance of entities l, t, and a is less than a given parameter w, then e_i is said to be an event instance (i.e., if all three entities occur within a window size of w sentences). In a document, the same location can be part of several event instances. The same holds for a date expression t and actor a. It is also possible that the same combination (l, t, a) forms more than one instance in document d_j, e.g., when respective expressions occur at the beginning and at the end of d_j. Figure 1 (left) illustrates these aspects. In Section 4, we elaborate on different settings for the window size w. Here, we assume that w is some fixed number of sentences.

Next, the strength of an event instance $e_i = (l, t, a)$ in terms of describing an event needs to be determined. For example, if the entity combination occurs frequently in sentences, that is, l, t, and a are mentioned in several documents within a single sentence, then there is a strong evidence that this triple describes an event. On the other hand, if l, t, and a occur only once and within a very large window, they are less likely to describe an event. We thus need a measure for the triple (l, t, a) that is derived from the event instances $e_i = (l, t, a)$ found in documents in D. Recall that $dist(r, r')$

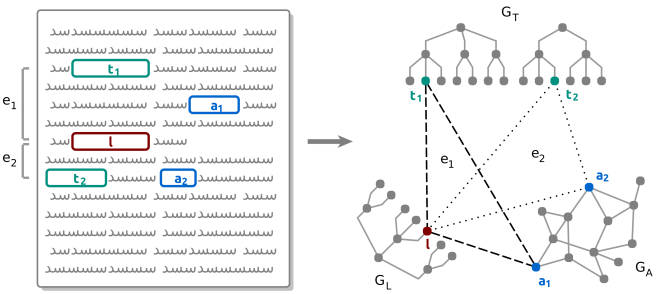

Figure 1: Mapping triples of named entities to event instances (left). Corresponding events e_1 (dashed) and e_2 (dotted) as 3-uniform edges in a hypergraph that connect the background networks.

denotes the distance between two entities r and r' in a document in terms of sentences. The further apart the two entities are, the less likely is there a (semantic) relationship between them. This can be formulated as a weight for a pair of instances of entities that decays exponentially, as already successfully employed for the construction of the above mentioned Wikipedia social network [13]:

$$\phi(r, r') := \exp\left(-\frac{dist(r, r')}{2}\right) \quad (1)$$

Thus, $\phi(r, r')$ transforms the sentence-based distance (i.e., a dissimilarity-like measure) between two entities into a measure of the strength of the relationship between them (i.e., a similarity-like measure).

Assuming that a triple (l, t, a) has been determined as an event instance e_i in a document, this raises the question of how we can determine the strength of that event instance from the strengths of the individual entity co-occurrences. One way to define this is the minimum weight of any pair of its components, i.e.,

$$\omega(e_i) = \min\{\phi(l, t), \phi(l, a), \phi(t, a)\} \quad (2)$$

This is a conservative approach to describe the likelihood of a triple forming an event instance; one could also use the average distance or maximum positional distance. Given that for a triple (l, t, a) there may be several event instances in the documents in D, we can naturally associate a set of weights with (l, t, a). From this set, a single weight is then derived. For example, the minimum or average of all weights for respective event instances can be used.

Definition 3. (Event and Event Weight) Given a combination (l, t, a) of location, date, and actor entities. Let $\omega(e_i), \ldots, \omega(e_k)$ denote the weights that have been determined for each event instance i for (l, t, a) in the documents of a collection D. Then $e = (l, t, a)$ is called an event, and its weight, denoted $\omega(e)$, is

$$\omega(e) := \sum_{i=1}^{k} \omega(e_i) \quad (3)$$

Based on this definition of an event and its weight, we now introduce the event network that links the three background networks. Intuitively, each event consists of a triple (l, t, a) from the sets of locations, dates and actors. We formalize this notion as a hypergraph in the following.

Definition 4. (Event Network) Given a set $\mathcal{E}$ of events extracted from a document collection D. The event network for $G_{\mathcal{E}} = (V_{\mathcal{E}}, \mathcal{E})$, consists of a set of nodes $V_{\mathcal{E}} = T_L \cup T_D \cup T_A$ (the entities extracted from D) and a set of hyperedges that directly correspond to the event triples $\mathcal{E}$. The weight of an edge e is then identical to the weight $\omega(e)$ of the event.

$G_{\mathcal{E}}$ therefore forms a 3-uniform, 3-partite hypergraph over the sets of locations, dates and actors, i.e., each edge consists of exactly 3 entities and includes exactly one entity from each set. Shared participation of entities in an event is then expressed through the incidence of edges. Each edge may share at most two entities since edges with three shared entities would be identical (see also Figure 1 (right)).

3.3 Event Refinement

Based on an event network $G_{\mathcal{E}}$ constructed from a document collection, we now describe our approach for the refinement of event queries. The basic idea is that the user specifies a triple ⟨`location,date,actor`⟩ for which a ranked list of events is returned that "match" the query event. Such a match can be generated or approximated by different methods, which we describe in the following. While we focus on textual location queries, a query interface that enables reverse geocoding could be used to support the input of coordinate or geometry queries as well. In general, we distinguish between two scenarios:

1. The user specifies a query event $e_q = (l_q, t_q, a_q)$ for which a ranked list $R = e_1, e_2, \ldots, e_k$ of events is returned such that for each $e_i = (l_i, t_i, a_i) \in R$, we have $l_i \sqsubseteq l_q$, $t_i \sqsubseteq t_q$, and $a_i = a_q$. That is, the events in R have the same or finer granularity with respect to the date and location specified by the user. For the purpose of this paper, we assume that the actor entity in the events has to be the same as in the query, although this assumption can be weakened.

   For the event query ⟨`Germany, 1967, Jimi Hendrix`⟩, for example, a list of events could be returned that includes events such as ⟨`Munich, 1967-05-16, Jimi Hendrix`⟩, ⟨`Germany, 1967-05, Jimi Hendrix`⟩, and ⟨`Berlin, 1967-05-15, Jimi Hendrix`⟩, as presumably respective events in the event network.

2. Assume that, for the event specified by the user, there is no event that refines the query event as above. In this case, new query events are generated "nearby" the query event, either spatially or temporally. Thus, we exploit the ordering of the geographic and temporal domain to find possible candidates for l_q and t_q.

   For example, if the query for the event ⟨`Germany, 1968, Jimi Hendrix`⟩ returns an empty list, then as nearby query events we would consider ⟨`Germany, 1967, Jimi Hendrix`⟩ and ⟨`Germany, 1969, Jimi Hendrix`⟩.

Independent of the scenario, the ranking of events can be determined based on user expectations and information needs. Since a weight is associated with each event, matching (refined) events can be ranked based on these weights. However, the user may also be interested in the most fine-grained event(s) matching the query event. In this case, respective events should appear in the ranking before coarse-grained events. Conceptually, the computation of a ranked list of matching (refined) events is straightforward. For a

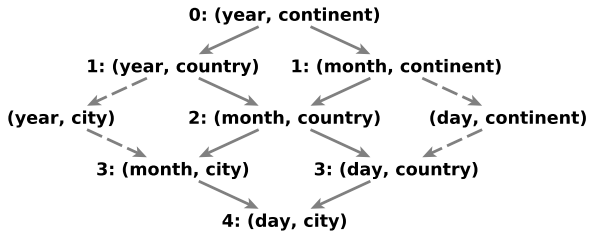

Figure 2: Stratification hierarchy for combinations of date and location types.

given event network $G_{\mathcal{E}}$ and query event $e_q = (l_q, t_q, a_q)$, a set E_q of matching (refined) events is determined as

$$E_q := \{(l, t, a) \in \mathcal{E} \mid t \sqsubseteq t_q \ \wedge \ l \sqsubseteq l_q \ \wedge \ a = a_q)\} \qquad (4)$$

The elements $e \in E_q$ of this event set are then ranked non-ascendingly according to their weights $\omega(e)$. Both the identification of candidates for E_q as well as the ranking can be performed efficiently due to the hypergraph structure that links the three background networks. The background networks for locations and dates can be used directly to obtain the lists of events whose date and location is of finer granularity than those specified in the query event (l_q, t_q, a_q) by simply following the directed edges outgoing from nodes $l_q \in G_L$ and $t_q \in G_T$, respectively.

The matching events E_q are likely of heterogeneous granularity. Note that E_q may contain at most one event that exactly matches the query event e_q. All other events have a refined location and/or a refined date. Based on this observation, we impose a stratification on the events in E_q that reflects the different types of refinements. Given the types of dates (year, month, day) and locations (continent, country, city) we consider in our approach, the stratification hierarchy illustrated in Figure 2 provides a more meaningful refinement. Independent of the actor component, the date and location pair specified in a query event e_q will always match exactly one of the nine patterns. For example, $e_q = \langle$`2016-07, Germany, ...`$\rangle$ corresponds to pattern 2 and can be refined in three ways. The list of (refined) matching events for a query event can thus always be partitioned accordingly. For example, the partition corresponding to class 4 would always include matching events of the finest granularity. Note that in each partition, matching events can still be ranked based on the events' weight.

Non-existing matches. We now consider the second scenario outlined at the beginning of this section, the case in which there is no matching event (of finer granularity) for a given query event. Following the motivation of our framework that users often do not know the exact time and location of an event, there has to be some flexibility in terms of how matches for a query event are determined. Assume, for example, a query event $e_q = (l_q, t_q, a_q)$ where no event for that exact date t_q or any finer granularity exists. Similarly, the user might specify an actor a_q for which no matching events for l_q and t_q exist. In these cases, new query events are derived that are "nearby" as follows:

- If there is no match for the location l_q, then neighboring locations l' are used to build new query events. If l_q is of type country, then for each neighboring country l' of l_q a new query event $e'_q = (l', t_q, a_q)$ is built. In case l_q is of type city, then the k nearest cities or neighboring cities in a radius of k kilometres are used. In Section 4 we elaborate on the choice of respective neighboring functions for cities.
- In case there are no matches for the date t_q, due to the ordering of the temporal domain, neighboring dates can be determined easily. If $t_q =$ `2016-01`, for example, then the neighbors would be `2015-12` and `2016-02`. This approach of building new query events is applied in all three cases, that is, if t_q is a year, month, or day. In case that no match is found for a neighboring date either, the approach is applied iteratively through neighboring dates.
- If for the actor a_q, no matching event can be found, the actor network can be employed by simply generating new event queries based on the k nearest neighbors reachable from a_q in the actor network G_A.

Since the background networks are pre-generated, such query refinements are based on local graph neighborhoods and the computational effort is low in practice. By accounting for the above cases, the model is flexible enough to handle a variety of uncertain queries even in settings where only a limited amount of information is available for a query event, as we demonstrate in the following section.

We conclude this section by discussing some important practical aspects of the proposed framework. First, beyond the ranked events of a query result, each event in the list can be linked to the context in which it appears and the documents that contain it. Thus, a user interface might highlight respective instances of a given event in documents to enable further exploration. Second, since the events are embedded in a network structure, a user can explore related events. For example, it is a simple matter to obtain events that share either the same location, date, or actor. Thus, the network structure not only provides an effective and efficient data structure for query refinement but also an interface for rich event exploration scenarios in which the user can traverse the network structure, explore paths between events, and look for patterns of interest such as shared locations.

4. EXPERIMENTAL EVALUATION

Based on the above discussion, we now show how to instantiate an event network alongside the three individual background networks, before evaluating them for the task of event refinement on a set of hand-annotated events that we extracted from news articles.

4.1 Information Extraction

For the construction of the event network and the background networks, a comprehensive source of data is required. Thus, we use the text of the English Wikipedia from the data dump of May 1, 2016. We extract about 5M content pages $(5,178,846)$, from which we remove all structured information (tables, references and info boxes). The remaining unstructured text serves as our input and is split into 97M sentences $(97,771,681)$. During content extraction, links in the text of Wikipedia are resolved to match Wikidata items.

Named entity annotation. Wikipedia links are embedded links to other Wikipedia pages, much like hyperlinks. Additionally, almost every Wikipedia page is also directly linked to a Wikidata item in the knowledge base behind

actor		location	
name	frequency	name	frequency
Barack Obama	13,187	USA	310,697
George W. Bush	10,759	France	116,307
Napoleon	9,611	Germany	105,347
W. Shakespeare	9,466	Canada	102,322
Adolf Hitler	8,700	India	95,681
Jesus Christ	8,401	UK	93,933

Table 1: The six top Wikipedia links by type alongside their frequencies in the English Wikipedia.

Wikipedia [32]. This direct link to a knowledge base allows a simplification of the named entity recognition (NER) since we only have to classify entities in Wikidata to assign class labels to the linked entities within the Wikipedia text. While such a classification of Wikidata entities is non-trivial due to the convoluted hierarchies of Wikidata [26], it is more precise than the application of NER frameworks for unstructured text. Therefore, the combination of Wikipedia links and Wikidata can serve as a unique resource for the creation of highly accurate entity annotations in the unstructured text of Wikipedia.

For the purpose of this paper, we assign entities in Wikidata to the named entity classes *locations* L and *actors* A as follows. Each Wikidata item with the statement `is instance of: human` is assigned to the actor class. While this creates some ambiguity with regard to fictional persons contained in the knowledge base, the coverage of the results is satisfying. For locations, the situation is more complicated. All items with the statement `is instance of:` and at least one of the values `continent`, `supercontinent`, `country`, `sovereign state`, `constituent country`, `city`, `town`, `village` or `capital` are assigned to the location class. Additionally, entities for which information about its `population`, `time zone`, `coordinate location` or `postal code` is available, are treated as locations. By construction, this class includes subclasses that can be used to distinguish between countries and cities, which is useful for finding neighboring locations in the background networks (see Section 4.2). In total, we find $84,110,797$ Wikipedia links on $4,813,014$ different content pages, of which $11,637,312$ link to $918,989$ different actors while $21,863,546$ link to $758,500$ unique locations. In Table 1, we show the top-referenced actors and locations by Wikipedia link frequency.

This approach, if applied without further consideration, is slightly problematic due to the Wikipedia guidelines, which state that only the first mention of an entity on a Wikipedia page should be linked to its respective Wikipedia page. Therefore, subsequent mentions of an entity may not be found by our approach, which could lead to incomplete co-occurrence information and thus incomplete events or networks. To account for this, we also employ a string search for the cover texts, parts of the cover text and the Wikidata label of named entities that we locate on any given page. Further mentions of these strings are then directly linked to the previously disambiguated entity.

Temporal annotation. For the annotation of dates, a different approach is required. While some dates have Wikipedia pages and are (sometimes) linked in the text, the vast majority of dates is not supported by Wikipedia links and there are no pages for each individual day. Furthermore, since Wikipedia texts are generally written in a narrative style, many temporal taggers that are trained for the news domain are ill equipped to handle this input data. Therefore, we use HeidelTime as a domain-sensitive temporal tagger that can be adjusted to narrative texts [30]. Here, we limit the extraction to temporal expressions that can be normalized to a granularity of year, day or months (i.e., no intervals). In total, we find $43,873,004$ such temporal expressions that we use for the construction of the networks.

4.2 Background Networks

Based on the annotations of entities in the text, we now construct the background networks. Intuitively, to keep the network structures natural, we construct the temporal and geographical networks in a hierarchical structure, while the actor network is designed to resemble a social network.

Location network. In order to construct the network from the identified locations, we extract further information from Wikidata to better distinguish them and assign to them a level in the hierarchy. Specifically, we identify countries from the list of recognized UN countries (i.e., member states of the United Nations). This distinction is necessary to avoid the inclusion of historic countries that do not exist any longer as well as ambiguities between historic and present-day versions of the same country. For each of these countries, we extract lists of bordering countries from Wikidata and include this relationship in the network. Furthermore, countries are linked to the continent on which they are located in a hierarchical relationship. For each city, we extract the coordinate information from Wikidata (central point coordinates are used as an approximation for the geographic extension). We calculate great circle distances between all cities in the data set with the Haversine formula [25]. For any given city, we then rank neighboring cities based on this distance to extract the k closest cities or all cities within a radius of m kilometres as neighbors. For the following evaluation, we let $k = 25$, i.e., we extract the 25 closest cities and consider them to be neighbors.

Temporal network. The construction of the temporal network is fairly straightforward since the hierarchy of days, months and years is already given. For any date of granularity `yyyy-mm-dd`, we link it to the corresponding month `yyyy-mm`. Similarly, for months, we link them to the corresponding year. The resulting network can thus be used to identify temporal neighbors of given dates as well as contained dates for the evaluation of event refinement.

Actor network. The network of actors is different from both the geographic as well as the temporal network since there is no clear hierarchy. Conceptually, such a hierarchy might exist in settings that are very specific to a domain. For example, members of a given government can usually be arranged in a hierarchical fashion with the head of the state at the top of the hierarchy, the ministers at the next level, followed by emissaries, and so on. A similar case can be made for companies, of course. However, the extraction of such hierarchies for would be prohibitively expensive for a data set that is as large as Wikipedia, and even if it could be done, there are many domains where such information is unavailable (e.g., film actors, athletes, etc.). Therefore, we employ a different approach and suggest the use of a social network that is based on relationship strength between individual actors in the network. Since such a network can be extracted from entity co-occurrences [13], this is a natural candidate for the background network that contains relations between the set of all persons mentioned in Wikipedia.

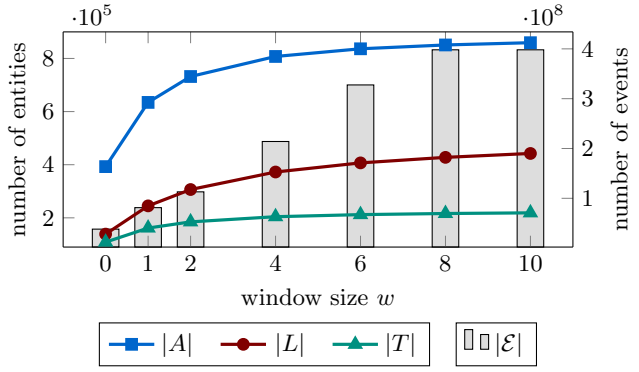

Figure 3: Number of Events ($|\mathcal{E}|$) as a bar chart and number of unique entities participating in events for different window sizes w.

4.3 Events and Event Network

The event network can be extracted in much the same way as the background networks. Since the set of nodes is already known from the background networks (i.e., the locations, dates, and actors), it is just a matter of linking them by extracting the hyperedges that correspond to the events. To this end, we scan the entity-annotated text of Wikipedia and extract as event instances all co-occurrences of entity triples that occur within a given window w, before aggregating them into events as described in Section 3.2. While the intuition behind the window size w is simple, the choice of an optimal size is difficult. Ideally, events should be described within one sentence that contains both time and location as well as the involved actor(s). However, this may not occur in reality and we thus consider different window sizes from $w = 0$ (which correspond to intra-sentence co-occurrences) up to $w = 10$ (which indicates that entities may be up to 10 sentences apart). In Figure 3, we show the number of identified events by varying window size as well as the number of entities that participate in these events. The number increases strongly up to a window size of $w = 2$, which corresponds to the case of the entities being spread across three consecutive sentences. Afterwards, the number increases much more slowly, indicating that $w = 2$ is likely a good candidate for the window size. While higher values seem to be a rather drastic assumption and stretch the concept of semantic relatedness by proximity, we find that this may occur occasionally as we demonstrate in the evaluation.

A second, less obvious effect of the window size is the influence on the participating entities of events. In Table 2, we show the overall most frequent locations, dates, and actors when limited to entities that occur in an event, for two values of w. It is evident that for lower window sizes, Indian actors play a much more important role than they do for large window sizes. This effect is not observable for locations or dates, however, where an increasing window size has little effect and the focus remains strongly on western locations and dates past 2000, respectively. While we are not certain what the reasons behind this phenomenon are, we note that the English Wikipedia is collaboratively edited by users worldwide. This spike in co-occurrences may thus indicate a different structure in the description of events in localized versions of English as spoken in different parts of the world and warrants further investigation.

w	actor	date	location
1	Gautama Buddha	2004	USA
1	Aurangzeb	2002	London
1	Akbar	2010	France
1	Shah Jahan	2006	India
1	Mahmud of Ghazni	2005	Paris
8	Napoleon	2006	USA
8	Elizabeth II	2005	London
8	Queen Victoria	2004	New York City
8	Barack Obama	2010	France
8	George W. Bush	2002	United Kingdom

Table 2: The five most frequent entities that participate in events for two different window sizes w.

4.4 Evaluation of Refinement

Before we proceed to the evaluation of the event refinement procedures, we now introduce the set of queries, their ground-truth answers, and the evaluation measure.

Ground truth and queries. To obtain a set of queries for evaluation as well as corresponding ground-truth answers to these queries, we turn to a source outside of Wikipedia in order to avoid evaluating on the same data that we used to construct the networks. Therefore, we look at a selection of news articles (56,008 in total) from British and U.S. newspapers. Specifically, we consider news articles with a political content (world news, UK or US news, UK or US politics) that were published between April 1999 and May 2016 in the newspapers New York Times, The Independent, The Guardian and the Reuters news agency.

We extract event candidates by automatically annotating person and place mentions in the text of the news articles with the Stanford NER toolset [11] and dates with HeidelTime [30] to identify triples of entities. Based on this output, we then check these event candidates manually to ensure that they are meaningful real-world events and that the entities occur in Wikidata. The latter restriction serves to avoid an evaluation of the comprehensiveness of Wikipedia. In total, we identify 31 events that we use for the evaluation.

For each of these events, we then generate a number of queries by reducing them to a more coarse-grained level in up to two of the dimensions time and location to create uncertain queries. For dates, we replace the day with the corresponding month. For locations, we move up in the spatial hierarchy, i.e., we replace a mention of a city by the corresponding country or the country by the continent, respectively. To increase the difficulty, we also consider the case of queries that are uncertain in both time and location. Additionally, we build queries to test for the neighborhood refinement by replacing days with a neighboring day. Similarly, for cities and countries, we replace them in the event by a random neighboring city or country. Again, we also consider the case where both the date and location are uncertain and only a neighboring date or a neighboring location is given. Since not all locations in the queries represent a city or a country and we focus on these location types in our location network, we do not have information about neighboring locations for all queries. Take for example the Mount Everest, for which the set of neighbors (cities, countries and mountains) is very heterogeneous. Thus, we limit the selection of locations to the hierarchy of continents, countries and cities and include only 18 queries for neighboring locations. Finally, we combine both approaches to create queries

refine	T	L	T & L				T	L
neighbor				T	L	T & L	L	T
$\|Q\|$	31	31	31	31	18	18	18	31

Table 3: Size of the set of uncertain queries Q for modifications along the dimension of date T or location L. Shown are coarse-grained queries for refinement (left), uncertain neighborhood queries (middle) and their combination (right).

in which either the date or the location are more coarse and the other is replaced by a neighboring entity in the hierarchy. In Table 3, we show the total number of queries for each possible modification (i.e., reduction in certainty) that can be applied.

In each of the cases, our goal is the identification of the original, certain event in the event network from the uncertain, more coarse-grained input query event. In the following, we show that we are able to reliably find the original event in the event network, even for uncertain queries.

Evaluation measure. For each type of refinement or combinations of refinements, we obtain a list of queries. Each of these queries then produces a ranked list of resulting events that we have to evaluate. In this setting, the mean averaged precision (MAP) is a solid evaluation measure, which locally aggregates the results of each query and then averages over all queries. Formally, we first compute the average local precision for each query q as

$$AvgP(q) = \frac{\sum_{k=1}^{n} P(k) \cdot rel(k)}{m} \tag{5}$$

Here, m denotes the number of relevant results (in our case, $m = 1$ since there is one correct event for each query), n is the number of retrieved events (i.e., the size of the ranked list), $P(k)$ denotes the precision at rank k, and for the relevance we let

$$rel(k) = \begin{cases} 1 & \text{if event at rank } k \text{ is correct} \\ 0 & \text{otherwise} \end{cases}$$

These average precision scores are then combined into the mean averaged precision by averaging the local results for all queries in the set of queries Q as

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} AvgP(q) \tag{6}$$

We compare the resulting MAP scores for different event refinement schemes and a number of window sizes.

Results. In Figure 4 (top), we show the results of our evaluation for different extraction window sizes and combinations of the two query modifications. As expected, the precision drops with growing window size due to the overall growing number of events (and events that are returned for each modified query). For the queries where the information was made uncertain, the precision is lower than for the neighbor queries. Here, the explanation is given by the observation that a reduction of entities to a coarse-grained level in the refinement step also includes all entities with finer granularity in the query result. For example, a reduction of the date `2013-08-09` to `2013-08` in the search process causes all dates in August 2013 to be retrieved. If the date is instead set to `2013-08-10` as a neighbor, the network is only searched for the two directly adjacent dates.

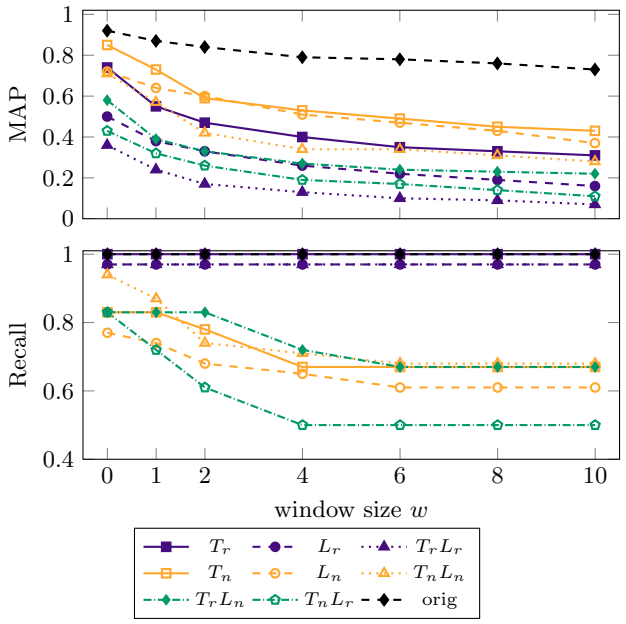

Figure 4: MAP(top) and Recall (bottom) for different extraction window sizes and combinations of the two query modifications. The two possible modifications (refinement uncertainty r and neighborhood search n) are applied to the geographic dimension L, the temporal dimension T, and their combination. The unmodified queries are denoted as *orig*.

In our evaluation of the precision, we only include query events that are present in the event network with window size $w = 0$. A full evaluation of the recall based on all news events would not be meaningful since we would end up evaluating the comprehensiveness of Wikipedia and the recall of the NER tools. However, a relative evaluation of the recall based on window size is more sensible. As shown in Figure 4 (bottom), we find that the recall is steady over growing window size for the unmodified query and the uncertain queries. This is unsurprising due to the increased number of events in the network. For neighbor queries, the recall decreases with growing window size, which we attribute to the fact that the search for neighboring entities is only performed if no results are found after refining the entities. With larger window size the probability for finding events by refining the query increases, and therefore the search for neighboring entities is performed less often.

In summary, we find that the choice of an optimal window size depends on the application, as the approach is most precise for window size around $w = 1$, yet the recall suffers for low values of w. Furthermore, we expect that the effects of the window size likely to depend on the language of the document collection.

5. CONCLUSIONS AND ONGOING WORK

In this paper, we presented a novel approach for the extraction and representation of spatio-temporal events from large corpora of unstructured text. Based on three classic background networks that represent the relationships between locations, dates and actors as participating entities of events, we connected the three distinct entity types in a regular hypergraph model that captures the described events.

Using this model, we highlighted the versatility of such a network representation in the answering of event queries as well as the applicability to the task of imprecise event retrieval. Here, we found that the underlying graph structure allows for an efficient retrieval of adjacent entities as well as entities with refined granularity for event queries that are based on imprecise information. An evaluation of our approach on a query set of news events suggests that it is highly precise for small windows sizes in the event extraction phase. As a result, we find that the proposed network combination of event and background networks can serve as a valuable tool for the retrieval and exploration of both precise and imprecise events from collections of unstructured text, where the networks of entities induce a navigable structure.

Ongoing work. While we observe that most events can be found in a window of $w = 2$ sentences, the restriction to triples in the extraction phase is fairly strict and does not account for events that are missing one dimension (actor, location or date information) or are spread across multiple documents. To this end, we are working on methods for the imputation of event hypergraphs from basic co-occurrence graphs through edge aggregation techniques. This will then support the extraction and refinement of even those events that are not explicitly given as triples in the text.

Acknowledgements. The authors would like to thank Christian Kromm and David Stronczek for their assistance in preparing query and ground-truth data.

6. REFERENCES

[1] H. Abdelhaq, M. Gertz, and A. Armiti. Efficient Online Extraction of Keywords for Localized Events in Twitter. *GeoInformatica*, 2016 (online April 2016).

[2] A. Abujabal and K. Berberich. Important Events in the Past, Present, and Future. In *TempWeb*, 2015.

[3] B. Adams and M. Gahegan. Exploratory Chronotopic Data Analysis. In *GIScience*, 2016.

[4] C. C. Aggarwal and K. Subbian. Event Detection in Social Streams. In *SDM*, 2012.

[5] J. Allan. *Topic Detection and Tracking: Event-Based Information Organization*, volume 12. Springer Science & Business Media, 2002.

[6] J. Allan, R. Papka, and V. Lavrenko. On-line New Event Detection and Tracking. In *SIGIR*, 1998.

[7] R. Blanco and C. Lioma. Graph-based Term Weighting for Information Retrieval. *Information Retrieval*, 15(1), 2012.

[8] T. Brants, F. Chen, and A. Farahat. A System for New Event Detection. In *SIGIR*, 2003.

[9] A. Ceroni, U. Gadiraju, J. Matschke, S. Wingert, and M. Fisichella. Where the Event Lies: Predicting Event Occurrence in Textual Documents. In *SIGIR*, 2016.

[10] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, and J. Huang. STREAMCUBE: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration Over the Twitter Stream. In *ICDE*, 2015.

[11] J. R. Finkel, T. Grenager, and C. D. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL*, 2005.

[12] J. Foley, M. Bendersky, and V. Josifovski. Learning to Extract Local Events from the Web. In *SIGIR*, 2015.

[13] J. Geiß, A. Spitz, and M. Gertz. Beyond Friendships and Followers: The Wikipedia Social Network. In *ASONAM*, 2015.

[14] J. Geiß, A. Spitz, J. Strötgen, and M. Gertz. The Wikipedia Location Network: Overcoming Borders and Oceans. In *GIR*, 2015.

[15] S. Julinda, C. Boden, and A. Akbik. Extracting a Repository of Events and Event References from News Clusters. In *AHA! Workshop on Information Discovery in Text*, 2014.

[16] N. Kanhabua and W. Nejdl. On the Value of Temporal Anchor Texts in Wikipedia. In *TAIA*, 2014.

[17] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. TwitInfo: Aggregating and Visualizing Microblogs for Event Exploration. In *CHI*, 2011.

[18] A. Mishra and K. Berberich. Event Digest: A Holistic View on Past Events. In *SIGIR*, 2016.

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

[20] S. Nepomnyachiy, B. Gelley, W. Jiang, and T. Minkus. What, Where, and When: Keyword Search With Spatio-temporal Ranges. In *GIR*, 2014.

[21] F. Rousseau and M. Vazirgiannis. Graph-of-Word and TW-IDF: New Approach to Ad Hoc IR. In *CIKM*, 2013.

[22] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *WWW*, 2010.

[23] A. D. Sarma, A. Jain, and C. Yu. Dynamic Relationship and Event Discovery. In *WSDM*, 2011.

[24] A. Schmidt, J. Hoffart, D. Milchevski, and G. Weikum. Context-Sensitive Auto-Completion for Searching with Entities and Categories. In *SIGIR*, 2016.

[25] R. W. Sinnott. Virtues of the Haversine. *Sky and Telescope*, 68(2):158, 1984.

[26] A. Spitz, V. Dixit, L. Richter, M. Gertz, and J. Geiß. State of the Union: A Data Consumer's Perspective on Wikidata and Its Properties for the Classification and Resolution of Entities. In *Wiki Workshop at ICWSM*, 2016.

[27] A. Spitz, J. Geiß, and M. Gertz. So Far Away and Yet so Close: Augmenting Toponym Disambiguation and Similarity with Text-based Networks. In *GeoRich Workshop at SIGMOD*, 2016.

[28] A. Spitz and M. Gertz. Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events. In *SIGIR*, 2016.

[29] J. Strötgen and M. Gertz. Event-centric Search and Exploration in Document Collections. In *JCDL*, 2012.

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

[31] J. Strötgen and M. Gertz. Proximity^2-aware Ranking for Textual, Temporal, and Geographic Queries. In *CIKM*, 2013.

[32] D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10), 2014.