Time will Tell: Temporal Linking of News Stories

Thomas Bögel
Institute of Computer Science
Heidelberg University, Germany
thomas.boegel@informatik.uni-heidelberg.de

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

ABSTRACT

Readers of news articles are typically faced with the problem of getting a good understanding of a complex story covered in an article. However, as news articles mainly focus on current or recent events, they often do not provide sufficient information about the history of an event or topic, leaving the user alone in discovering and exploring other news articles that might be related to a given article. This is a time consuming and non-trivial task, and the only help provided by some news outlets is some list of related articles or a few links within an article itself. What further complicates this task is that many of today's news stories cover a wide range of topics and events even within a single article, thus leaving the realm of traditional approaches that track a single topic or event over time.

In this paper, we present a framework to link news articles based on temporal expressions that occur in the articles, following the idea "if an article refers to something in the past, then there should be an article about that something". Our approach aims to recover the chronology of one or more events and topics covered in an article, leading to an information network of articles that can be explored in a thematic and particular chronological fashion. For this, we propose a measure for the relatedness of articles that is primarily based on temporal expressions in articles but also exploits other information such as persons mentioned and keywords. We provide a comprehensive evaluation that demonstrates the functionality of our framework using a multi-source corpus of recent German news articles.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA.

ACM 978-1-4503-3594-2/15/06 ...$15.00.
DOI: http://dx.doi.org/10.1145/2756406.2756919.

Keywords

Information networks; news networks; document similarity

1. INTRODUCTION

Many of today's news stories report on current events and topics that are complex in terms of the background the reader has to have to fully get a picture of the context of the story. Some news outlets, for select articles, provide links to related articles either within the news text itself or through an extra list of links, aiming to give the reader some more background information. However, such links often seem to be ad-hoc and do not follow any structure, such as a chronology. Take, for example, a news story that covers a recent meeting of the board of inquiry dealing with the NSA affair in Germany. Understanding statements made or actions decided by that board often require a good understanding of what happened in the past. Such historic information ideally should be covered as additional information in the article as well. In practice this is done, if at all, in a very short form and then mostly hinting to some event(s) in the past. More importantly, relevant historic information can be of diverse types and not only be related to a single topic. A good example for this is the pro-Russian unrest in the southern and eastern Ukraine with all its implications and precursors in the areas of politics and economy, to name but a few. A news article covering a recent development in that context can be related to many different topics, events, and previous stories. Getting a grasp of how it came to a recent event is difficult when only a single news article and, optionally, some ad-hoc links are provided.

Approaches that aim to recover a chain or thread of events for a news story typically focus on events within a single topic. Links to be discovered between news articles then have to be salient and coherent regarding that topic (see, e.g., [Feng and Allan, 2009, Shahaf and Guestrin, 2012]). Hierarchical topic models and document clustering approaches, e.g., [Nallapati et al., 2004], also only focus on single (composite) topics and are not able to relate individual articles that cover diverse types of events that belong to different topics. What becomes very clear for the two example stories mentioned above, however, is that a temporal ordering or a chronology of events play a crucial role in detecting link structures among news articles, an aspect mostly neglected in related approaches.

In this paper, we present a novel framework to extract links between news articles based on the "temporal relatedness" of articles. A key feature we build on for this are *temporal expressions* that occur in an article and, when nor-

malized appropriately, hint to points in time where older articles (published at that date or timeframe) might give more details about what is being covered in the article. Especially the extraction and normalization of temporal expressions is something rarely employed in standard approaches for determining document similarity or relatedness. Our conjecture is that temporal expressions provide a good focus on what articles should be investigated regarding their relatedness to a given article, an aspect that holds in particular true for the mostly very dense style in which today's news articles are written.

We propose an approach to construct a directed *information network* of news articles from a multi-source, heterogeneous collection of news articles that capture a broad range of topics. The construction of the network, which can be performed in an iterative fashion, e.g., for a stream of news articles, is primarily based on exploiting temporal expressions, but subsequently also uses person names and keywords occurring in parts of news articles. Furthermore, instead of comparing articles in their entirety as done in traditional approaches for document or topic similarity, we exploit the structure of typical news articles (lead paragraph, explanatory paragraphs, and additional information paragraphs). A network of articles extracted that way provides an ideal tool for the exploration and analysis of articles related to a given article, allowing the user to chronologically explore links between articles.

Because of the typical network structure built from a collection of news articles, different measures known for networks can be employed to further support exploration tasks. These include aspects like centrality (hinting to articles often referenced latently through temporal expressions), connected components (chronologically sorted collections of related articles), and bibliographic coupling (hinting to duplicate articles from different news outlets). In our experimental evaluations based on a large corpus of news articles from different German news outlets, we demonstrate the functionality of the proposed framework. In particular, we show that temporal link structures indeed provide an effective means to organize and explore news articles in a chronological manner not offered by related approaches.

The remainder of the paper is structured as follows. After a review of related work in the following section, we present our model for news articles and the approach for linking articles in Section 3 and 4, respectively. In Section 5, we evaluate our system and demonstrate the functionality and utility of our framework using large collections of recent German news articles.

2. RELATED WORK

Linking news stories has been addressed from a variety of directions. Most approaches in this area tackle the problem from a viewpoint of topic detection and tracking based on various similarity metrics: [Allan, 2002] and [Nomoto, 2010], for instance, use tf-idf and language models for content words, while [Shahaf and Guestrin, 2012] determine important phrases. [Vaca et al., 2014] employ collective factorization to model the temporal evolution of news. All of these approaches compute pairwise document similarity without taking into account the thread structure of news or giving explicit background information for complex news stories.

Besides topic tracking, many approaches focus on events for document linking (e.g., [Brants et al., 2003, Kumaran and Allan, 2005, Zhu and Oates, 2012, Zhu and Oates, 2013]). This is problematic as the definition of event is vague and a story thread might consist of multiple events covering different topics. In addition, each distinct definition of event requires laborious manual annotations.

[Wang and Li, 2011] use geo-spatial information to tackle the huge amount of news stories by coarse-grained geo-spatial clustering. Geographical information in isolation might, however, not be informative enough to model the fine-grained temporal development of news stories. Creating news summaries [Yan et al., 2011] is another approach to bring structure into large document collections but – by definition – neglects detailed information about relationships between news stories. As illustrated in Section 5.5.2, creating a story thread for central articles yields both a summary of a complex news story and also a fine-grained analysis of the temporal progression of a story.

The systems described in [Nallapati et al., 2004] and [Gillenwater et al., 2012] are most similar to ours as they create a linked, threaded structure of news. [Shahaf and Guestrin, 2012] also explicitly take the link structure of news into account by using activation patterns across multiple documents for document linking. In contrast to them, we do not rely on the notion of topical salience but our network is based on temporal links prevalent in news articles to allow for a fine-grained analysis. We will show in our evaluation that salience and document similarity get less and less reliable for connecting complex stories spanning a larger time frame. [Shahaf et al., 2013] use the metaphor of metro maps for representing complex threads of stories and their relationships. While their system also tries to combat information overload by representing news in a graph-like structure, the precondition is different: our system operates on a collection of documents that cover different topics and thus might be relevant to each other or not, while their approach assumes a collection of documents for a specific user query (e.g., "Middle East"), thus pre-filtering articles and inducing structure on articles for which a certain relatedness is already known.

In addition, while there are coarse-grained multi-lingual systems (e.g., [Pouliquen et al., 2008]), to our knowledge, there is no previous system that explicitly provides a fine-grained analysis of German news texts.

3. DOCUMENT MODEL AND TIMELINES

We assume a given collection A of news articles. To apply our approach, the collection does not have to be static but new (recent) articles can be added over time. Each article $a \in A$ has a timestamp $a.ts$, a title $a.ti$, and an abstract $a.ab$. Furthermore, an article can have one or more blocks $\langle b_1, \ldots, b_n \rangle$, given in document order, where a block is a paragraph or section of the article.

Basis for our framework to link articles using temporal information are *timelines*, which can be of different granularities. The timeline of the finest granularity is of type *date*, denoted T_{date}. An element of this timeline would be a fully specified date, e.g., "2014-08-30". Coarser timelines are T_{month} (with elements such as "2014-05") and T_{year} (with elements such as "2013") and are employed as well. The document timestamp of each news article (i.e., the date when an article has been published) is anchored at T_{date}. Furthermore, with each date $t \in T_{date}$, a set of articles A_t can be associated, i.e., $A_t = \{a | a \in A, a.ts = t\}$.

We now turn to three types of functions we employ to extract information from articles that are later used in our article linkage model.

Temporal Information Extraction. Let $T = T_{date} \cup T_{month} \cup T_{year}$ be the domain of all normalized time elements of different granularities. Assume some text corresponding to the title, abstract or some block of an article. Then, the function $t : text \rightarrow 2^T$ returns a set of normalized time elements. A temporal tagger in combination with a normalization component is a realization of such a function, where the tagger discovers explicit, implicit, and relative temporal expressions in a text and maps them to a normalized, standard format (see, e.g., [Strötgen and Gertz, 2013]).

Person Names. A common task in Named-Entity Recognition is the detection of person mentions in a text and mapping respective expressions to normalized person names. Assume a set P of normalized person names. The function $p : text \rightarrow 2^P$ returns a set of person names that have been detected in a text.

Keywords. The third function we employ returns a set of keywords that occur in a text. In our approach, we consider nouns, adjectives, and verbs, all in their lemmatized form, as keywords. The function $k : text \rightarrow 2^W$ returns a set of keywords for a text, with W simply being some kind of dictionary of words.

4. ARTICLE LINKAGE MODEL

In this section, we present our model to link news articles based on temporal expressions occurring in articles. After giving the problem statement in the following Section 4.1, in Sections 4.2 and 4.3, we detail how links between pairs of articles are determined and how an information network of interlinked news articles is incrementally built. In Section 4.4, we then discuss typical properties of such an information network, tailored to news articles.

4.1 Problem Statement and Objective

Given an article $a \in A$ describing some news story, we want to determine those articles in the collection A that provide further information about what is being covered in a. While we are not looking for a story chain describing the evolution of a single event or topic over time, we are interested in what other (older) articles in A are related to a in the sense that these articles provide further information about what is being covered in article a. This notion thus subsumes a story line or chain of events. Although we do not make use of an explicit notion or representation of an event, the key assumptions underlying our framework are that (1) there is a temporal relationship between events and, more importantly, (2) such relationships are often made explicit in news articles through temporal references. That is, if an article refers to "something" in the past by means of a temporal expression, then there should be an article related to that "something". We make such relationships between articles explicit through weighted links that suitably take the different types of the information (latently) embedded in and extracted from news articles into account.

Given a corpus of news articles, through the above type of relationship a directed information network of interlinked articles is formed. For example, for a given (recent) article a directed acyclic graph can be built that shows how a (recent) story or event is related to previous events that are both "temporally related" and "relevant" with respect to the current article. How individual links are determined, and how eventually a network is built and explored will be discussed in the following.

4.2 Pairwise Article Linkage

Given an article a with its components $\langle a.ts, a.ti, a.ab, a.b_1, \ldots, a.b_n \rangle$. For the sake of explanation, we first only consider temporal expressions of type T_{date} that can be extracted from the abstract $a.ab$ and blocks $a.b_i$. Let $t(a) = t(a.ab) \cup t(b_1) \cup \ldots \cup t(b_n)$ denote all normalized date expression extracted from these $n+1$ components of the article a. For each date $t \in t(a)$, we determine the set of articles $A_t \subset A$ that have the timestamp t. We thus get a set of all articles that have a document timestamp corresponding to some temporal expression in article a. Note that not every component in a has to have a temporal expression, and the same (normalized) temporal expression can occur in more than one component. For each date $t \in t(a)$, one can consider A_t as a set of articles that are "temporally related" to the article a.

Given an article $a' \in A_t$ with timestamp t, we now determine whether also the content of a' is "relevant" to the article a, assuming that a' might provide some background information about what is covered in a as a' is already temporally related to a. For this, we do not compare the two full articles a and a' but their components as follows. For two text components c_1, c_2, the similarity is defined as follows:

$$sim(c_1, c_2) := \frac{1}{4}(\alpha_1(Jac(p_{c_1}, p_{c_2}) + Cos(p_{c_1}, p_{c_2})) + \alpha_2(Jac(k_{c_1}, k_{c_2}) + Cos(k_{c_1}, k_{c_2}))) \quad (1)$$

where p_{c_i} and k_{c_i} denote the set of persons names and keyword, respectively, extracted from text components c_i and c_j (see Section 3). Jac and Cos denote the Jaccard and Cosine similarity, respectively. We choose this formulation of a similarity measure because both metrics have successfully been used in previous work for story linking of English texts [Kumaran and Allan, 2005, Wang and Li, 2011]. We also employ a weighting of person and keyword similarity using the weights α_1 and α_2 with $\alpha_1 + \alpha_2 = 1$. This allows us to employ, e.g., a more person-centric similarity by choosing appropriate values for these weights.

To determine whether an article a' that is temporally related to an article a is also relavent from a content point of view, we introduce a measure for relevance as follows. Let $a.B_t = \{b_k, \ldots, b_l\}$ denote all components in article a that have a temporal expression mapped to time point t (the timestamp of article a'). The relevance measure $link$ is then defined as follows:

$$link(a, a') := \max_{a.b_i \in a.B_t} sim(a.b_i, a'.ab) \quad (2)$$

That is, we determine the relevance of an article a' with respect to an article a based on only the maximum similarity between the abstract of a' and each block in $a.b_i$ of a that has a temporal expression mapped to t. The rationale for choosing only the abstract of a' for comparison is that we assume the abstract of an article gives a succinct summary of an article and it mentions important terms and persons.

4.3 Network Construction, Pruning, and Support Paths

For a collection of articles, the above computation of pairwise links can be done efficiently, starting with the most recent article(s) and then proceeding in a chronological order of article timestamps. It should be noted that if for an article we consider only temporal expressions whose normalized values are less (earlier) than the article's timestamp, we obtain a directed acyclic graph, with paths leading from recent to older articles. If we also consider temporal expressions in an article a that refer to the future, then one can also compute links between a and articles whose timestamp is in $t(a)$, i.e., articles more recent than a. Such (potential) forward links can be maintained over time and investigated once new articles are added to the collection. If both types of temporal expressions are considered, then the resulting graph structure is likely to be not acyclic anymore. However, as our experiments will confirm, the amount of temporal expressions referring to the past is much larger than those referring to the future.

How to handle temporal expressions in articles coarser than of type T_{date}? Expressions of type T_{month} and the like result in time intervals. If an article contains such coarse-grained temporal references, then a pairwise linkage to those articles needs to be investigated whose timestamp lies in that temporal interval. Thus, all types of temporal expressions can be considered to build an initial *article network* $N = (A', L)$ of articles $A' \subseteq A$ linked by directed edges L between pairs of articles in A' according to Eq. (2) above.

Obviously, not every link is meaningful, namely when the value for $link(a, a')$ is below a certain threshold, say γ. Setting such a threshold is not trivial, and we will elaborate on this in Section 5. It would be straightforward to prune the network N by deleting all edges whose link value is below γ, leading to a reduced network, denoted N_γ. Instead, we make use of so-called *support paths* to keep some links with $link(a, a') < \gamma$. The idea is as follows: Assume a path $p = \langle a_1, \ldots, a_n \rangle$ where all edges have a value greater γ. All articles along that path are pairwise temporally related and also relevant with respect to their content. Assume another path $p' = \langle a_i, \ldots, a_k \rangle$ such that $a_i, a_k \in p$, that is, the first and last article in p' can be found along the path p. Figures 1(a) and (b) illustrate this case. Now assume that some links in p' have a value less than γ. Given that p provides evidence that a_i and a_k are related (because all link values along p are greater than γ), the path p' is fully kept, despite having (some) pairwise links below the threshold. In such a scenario, we call path p the *support path* for p'.

Based on the idea of support paths, given a threshold γ, a pruned network N_γ thus is obtained as follows: all edges are deleted from N that (1) are not part of a path being supported and (2) whose link value is less than γ. In Section 5, we will show that for a fairly heterogeneous collection of news articles, support paths typically consist of just a few edges (for the most simple example, see Figure 1(c)).

4.4 Properties of Article Networks

It is obvious that the larger the value for the threshold γ, the more components the network N_γ will have, because the number of support paths is reduced. The initial network N itself may have many components, namely when many articles neither contain a temporal expressions nor are temporally related to other articles. Each such article that is not linked forms a component. In a practical setting, one would choose some initial value for γ and let the user interactively increase or decrease that value to see how the network changes, e.g., what components develop when γ becomes larger.

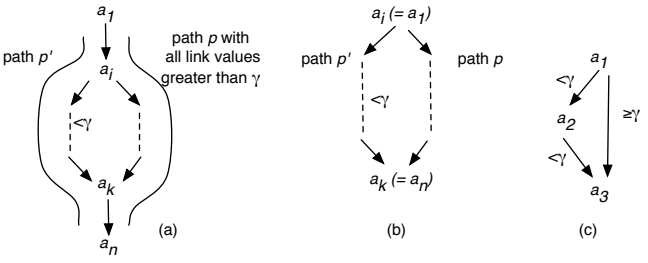

Figure 1: Examples of support paths scenarios ((a) and (b)) and most simple support path configuration found in a collection of news articles (c).

In general, an article network N_γ can be viewed as an information network specified in the form of a directed graph that may have several components. In particular, N_γ can be represented as an adjacency matrix $\mathbf{N}_\gamma$. Standard network measures [Newman, 2010] then can be obtained through matrix operations, such as in-degree centrality, which gives those articles that are often referred to by other articles (temporally and content-wise) and thus are likely to cover some key event or initial story.

Of particular interest is the *bibliographic coupling* measure known from citation networks [Newman, 2010]. In our context, bibliographic coupling of two articles a_i and a_j is the number of articles to which both are directly linked, described as $B_{ij} = \sum_{k=1}^{n} N_{ki} N_{kj} = \sum_{k=1}^{n} N_{ik}^T N_{kj}$, where N_{ij} is the ij-th element of the $(n \times n)$-matrix $\mathbf{N}_\gamma$. Articles that have a high bibliographic coupling thus are likely to be about the same event or story, or even duplicate articles, because they are related to many of the same (older) articles. In Section 5, we will show how bibliographic coupling is exploited in the context of a corpus of news articles from different news outlets.

5. EXPERIMENTS AND EVALUATION

Having presented the theory behind temporal linking and network construction, we will now present the details about our implementation, as well as optimization strategies for the resulting network in Section 5.3. Finally, the system will be evaluated formally and empirically in Section 5.5).

5.1 Description and Statistics of the Data Set

To determine the influence of data quality and size of the data set, we use two different data sets for our experiments:

NSA: A manually created data set covering a specific topic – the NSA spying scandal. We collected all articles related to the topic between 06-2013 and 08-2014 from two major German news sites, *Spiegel Online*[1] and *FAZ*[2].

[1] `www.spiegel.de/`
[2] `www.faz.net/`

	NSA	A5
articles	1155	13945
time range: begin	06/06/13	05/25/2014
time range: end	08/08/14	08/08/2014
news sources	2	5
sentences per article	41.2	32.3
exact dates (anchors)	2189	27703
articles: >0 exact dates	68%	69%
future references	5.9%	6.0%

Table 1: Data set statistics and comparison between NSA and A5 data set.

A5: A large data set automatically assembled from five major news sites (*Tagesschau*[3], *Spiegel Online*, *FAZ*, *Welt*[4], *SZ*[5]) by continuously pulling RSS streams in the categories *Politics* and *Economy* and parsing the raw HTML source using manually written rules.

A comparison of the statistics relevant to our system is given in Table 1. As our approach only links articles containing at least one temporal expression, 68% to 69% of all articles can *instantiate* a link. Note, however, that the remaining articles can be link *targets*. With only about 6% (relative to all references) being future references, only considering references to the past for linking (see Section 4.2) is a reasonable approach.

Comparing both data sets, one can see that documents in the NSA data set comprise more sentences and slightly more articles with exact dates than in the automatic data set. This is due to noise introduced by the automatic extraction process with few articles containing only photo galleries etc.

5.2 Preprocessing Pipeline

To extract temporal expressions, persons and keywords from our texts, we first apply a modular pipeline for robust information extraction at a fine-grained level based on Apache UIMA[6], performing annotations with increasing levels of complexity and allowing for easy adaptation and exchange of different base components. The pipeline consists of off-the-shelf NLP tools, such as the TreeTagger for part-of-speech tagging [Schmid, 1995] and Stanford NER for extracting person names. After person extraction, we cluster all mentions of persons that denote the same entity in the real world using similarities between names obtained by NESim [Do et al., 2009]. Our candidates for keywords are nouns, adjectives and verbs. Stop words and auxiliaries are filtered, and the lemmatized version of a token is used for a broader generalization and reduction of sparsity.

Extracting temporal expressions.

As our system strongly relies on temporal references in texts, we employ HeidelTime [Strötgen and Gertz, 2013], a multi-lingual temporal tagger that extracts and normalizes temporal expressions. The latter step is especially crucial as most of the temporal expressions are ambiguous, such as *"yesterday"* or *"last weekend"* that are only meaningful with respect to their specific context. The normalization component in HeidelTime correctly converts context-dependent temporal expressions to the concrete calendaric date (e.g., "2014-08-24") based on multiple heuristics.

Splitting documents.

The final component in our pipeline splits news articles into blocks using structural and paragraph information from the original article, such as tags within the HTML code or empty lines. Based on the information obtained by the pipeline, articles are linked as described in Section 4.2.

As a result, for an article consisting of the blocks $\langle b_1, \ldots, b_n \rangle$, the preprocessing pipeline creates a set of occurring persons $\langle p_{a_1}, \ldots, p_{a_n} \rangle$ and keywords $\langle k_{a_1}, \ldots, k_{a_n} \rangle$ for each block.

5.3 Network Optimization

After the network structure has been created as described in Section 4.2, we will now first study the influence of different similarity metrics and the pruning threshold γ. Having obtained a reasonable setting, we analyze the static structure of the network and validate our assumptions about necessary properties we consider important for the network to be used as an exploration tool.

Person-centric vs. balanced similarity.

To measure the effect of different weightings for person and keyword similarities – as introduced in Section 4.2 – we experiment with two different settings: a *balanced* version that assigns equal weights to person and keyword similarity ($\alpha_1 = \alpha_2 = 0.5$), as well as a *person-centric* scenario that favors person similarity ($\alpha_1 = \frac{2}{3}$, $\alpha_2 = \frac{1}{3}$).

Finding the optimal pruning threshold (γ).

As described in Section 4.2, all edges below a certain threshold γ are pruned to filter out weak links. After each pruning step, we also remove all nodes that are not connected. Figure 2 shows a histogram of edge weights – corresponding to document similarities – in the network when using the balanced and person-centric similarity metric, respectively. Most of the weights are very small, which is due to cosine similarity as a similarity metric for sparse vectors. In general, the person-centric metric yields more evenly distributed similarities. The average similarity for the balanced metric is 0.053, while it is 0.058 for the person-centric similarity metric.

To find the optimal threshold γ, we investigate its influence on the network structure, which is presented in Fig. 3 for the A5 data set. The numbers are similar for the NSA data set. Obviously, the number of edges decreases substantially with an increasing threshold, while the number of nodes decreases more slowly. This indicates that the pruning step reduces weak links between nodes but many of the nodes are still connected to other nodes. Comparing the two pairwise similarity metrics, the person-centric similarity metric yields higher overall weights and thus prunes less edges for the same threshold compared to the balanced similarity.

Setting γ to the average edge weight, more than 70% of all edges are removed from the network but only less than 6% of all the nodes. This means that removing many potentially weak edges still retains most of the nodes and thus information in the network. For the following experiments, we therefore use the respective average edge weight as our pruning threshold. The difference between the two similar-

[3] www.tagesschau.de/
[4] www.welt.de/
[5] www.sueddeutsche.de/
[6] http://uima.apache.org/

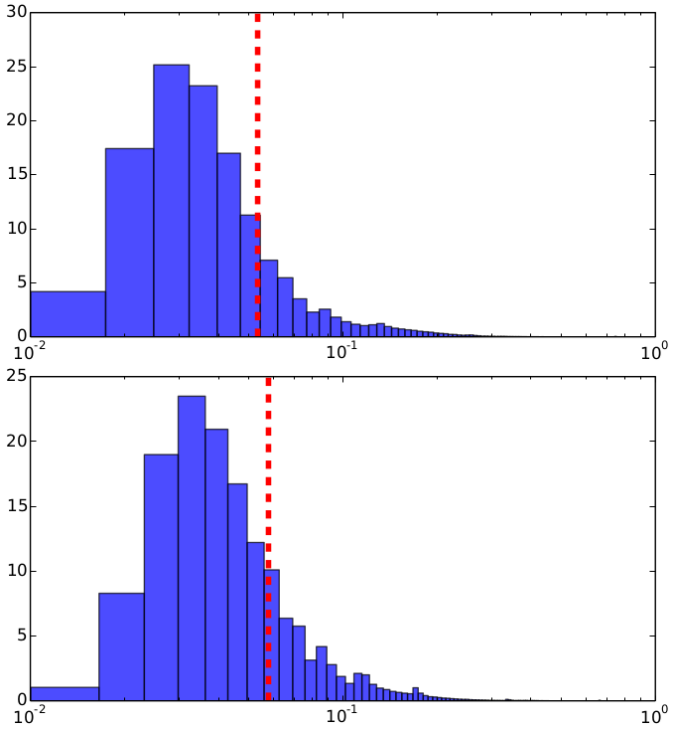

Figure 2: Distribution of edge weights (x-axis, log-scale) in the *A5* network for balanced (top) and person-centric (bottom) similarity. The dashed line represents the average similarity.

ity metrics when using the average similarity as a threshold is minor: nonetheless, the number of nodes is slightly higher for the person-centric metric, while the number of edges is slightly lower. As a loss of nodes represents a loss of information, we use the pruned network with person-centric similarities and support paths as the foundation for further data analysis. Note, however, that the choice of thresholds and similarity metrics is completely flexible and changeable by the user in an exploration interface, depending on the desired application.

Exploiting support paths as a quality indicator to block pruning (see Section 4.4) yields 30% more edges in the final network consistently across all data sets and similarity metrics. This shows that relying on single edge weights alone leads to the removal of potentially useful links as it neglects the additional information provided by the global network structure. Inspection of the data revealed that only less than 2% of all paths consist of more than 3 nodes (see Table 2).

Network statistics.

Table 2 lists statistics about the resulting networks when using the parameters described in the previous section: γ is set to the average similarity score and support paths are exploited. It is evident that the average size of components is much larger for the A5 data set (280 nodes per component) than for the NSA data set (23 nodes per component). This is mostly caused by the larger number of news sources that implicitly cite each other, resulting in more news articles that cover the same story.

	NSA	A5
nodes	668	11353
edges	1913	180138
components	63	86
cluster (communities, Sec. 5.4)	73	106
avg. nodes per component	23	280
avg. nodes per community	7	131
nodes in largest community	80	430
support paths (3 nodes)	203	19038
support paths (4 nodes)	46	3589
highest out-degree	34	230
highest in-degree	70	247

Table 2: Statistics for the information network obtained after pruning weak links as described in Section 5.3.

5.4 Investigating Story Threads

Due to the high connectedness in our information network – evident by the number of nodes per connected component – we employ strategies for clustering the entire network into smaller story threads that cover the progression of single story lines over time.

Community detection for clustering.

The number of nodes in the largest component is quite high for both data sets (cf. Table 2). To investigate the structure, we plotted the nodes and edges of the largest component of the NSA data in Figure 4. The network consists of densely connected sub-structures as well as single connections between these dense structures.

As the network structurally resembles the structure of a social network [Wasserman, 1994], we apply **community detection** to the network based on modularity optimization [Blondel et al., 2008] to cluster the network. The resulting communities – indicated by different colors in Fig. 4 – reflect these highly connected sub-structures and their relationships very well. Regarding a community as a **story thread**, the result allows a user to find pivot articles connecting two story threads.

As Table 2 shows, using community detection for clustering yields more clusters in comparison to components (106 vs. 86 for the A5 network) while maintaining connections between them at the same time.

Temporal evolution of a story thread.

Instead of using a model based on document similarity for linking stories – such as topic models or relevance models (e.g., [Lavrenko et al., 2002]) – we argue that explicit temporal expressions model the flow of complex story threads more naturally. To validate our hypothesis, we investigate the similarity between all documents within a story thread depending on the time span between pairs of articles. Our assumption is that the larger the time difference between articles within a story thread, the lower the document similarity. Document similarity is measured as described in Section 4.2.

Figure 5 illustrates the average similarity for article pairs covering a specific time span. As expected, the average document similarity decreases with increasing time spans. This shows that document similarity in isolation is not sufficient to cover evolving aspects of a more complex story but instead

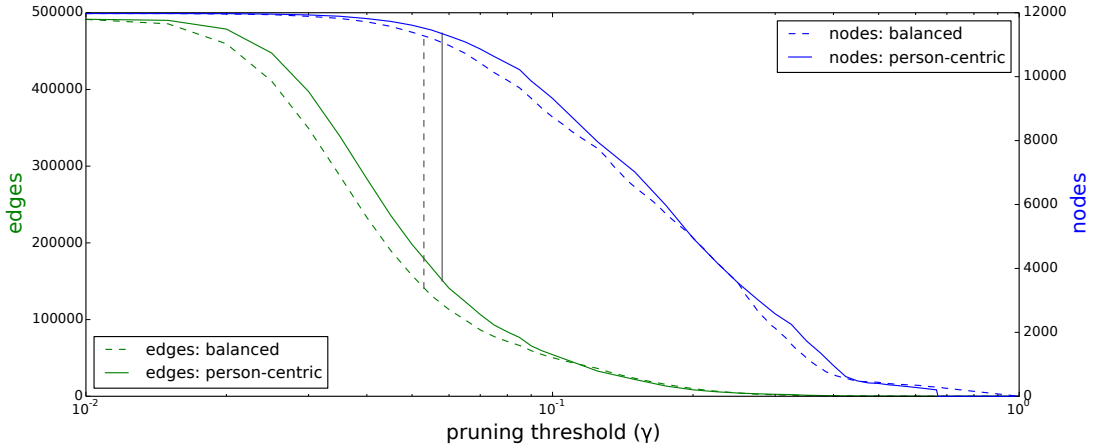

Figure 3: Influence of pruning: nodes and edges depending on γ for balanced (dashed, $\alpha_1 = \alpha_2 = 0.5$) and person-centric (solid, $\alpha_1 = \frac{2}{3}$, $\alpha_2 = \frac{1}{3}$) metric. Horizontal bars indicate the respective average similarity.

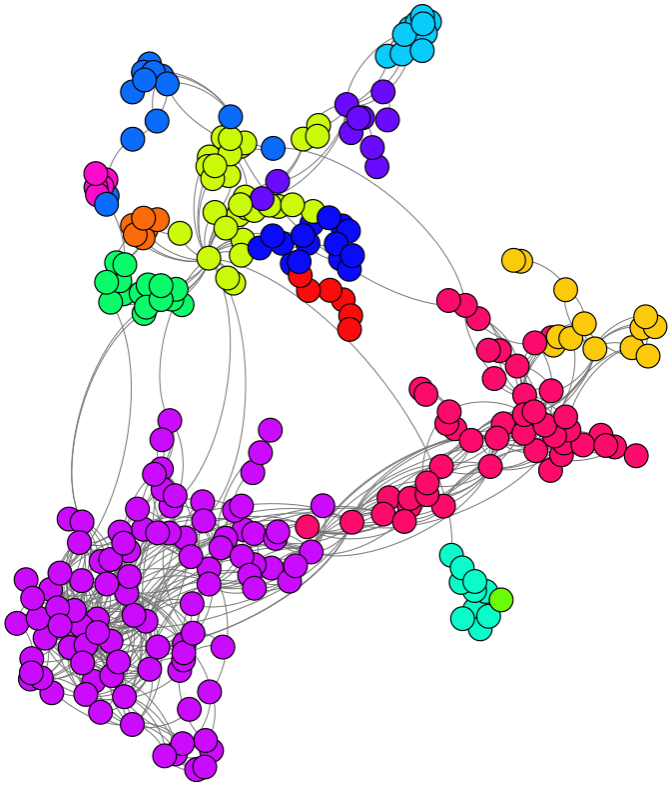

Figure 4: Structure of news articles and their connections in the largest component of the NSA information network. Colors represent different clusters based on community detection.

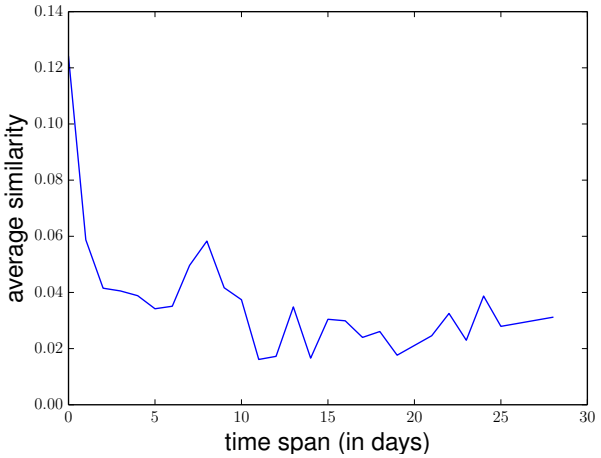

Figure 5: Average similarity within all articles of a community plotted against time span (in days) within a community.

yields story threads comprising time spans with a narrow, topically highly related collection of articles.

5.5 Evaluation and Data Exploration

As noted by [Shahaf and Guestrin, 2012], a quantitative evaluation of story linking in general is challenging due to the lack of training and evaluation data. This factor is owed to the highly subjective nature of the endeavor to link stories: similar to the field of Information Retrieval, there is no definite answer to the question of relevance. While we think that a formal evaluation is crucial, in our opinion, a more intuitive evaluation of the plausibility and practicality for using our system as an exploration interface is equally important.

Thus, we will first propose a setting to formally evaluate story linking approaches, followed by examples showing whether our system fulfills desired properties of a news exploration interface in Section 5.5.2.

	temporal links	explicit links
linked nodes	11353	4165
% of articles (coverage)	81.41%	29.87%
edges	180138	4432
↳ same news outlet	22.01%	95.17%

Table 3: Comparison of the article coverage of temporal vs. explicit links and the fraction of edges connecting articles from the same news outlet.

5.5.1 Comparison to explicit links

In order to evaluate whether our system links related articles that are of interest to the user, we propose a formal evaluation consisting of a comparison between the links obtained by our system and manual, **explicit** links added by the authors of an article.

As indicated in Section 1, explicit links are links within article pointing to other articles that might be of interest to the user reading the current story. As this is the typical (and oftentimes only) way of exploring related articles and because the links are manually added by the authors of an article, these links provide a good estimator of the quality of our linking approach. All comparisons between explicit and temporal links are performed with the entire A5 data set.

Article coverage.

As we regard our system as a news exploration tool that presents other, potentially interesting articles to the reader of a current news article, the overall coverage of news articles is very important. We regard as coverage the number of nodes that are connected to at least one other node (i.e. with a degree ≥ 1). If a node in the network is not connected, no valuable information about related articles can be found.

Table 3 compares how many articles are covered by the information network when using temporal links and explicit links. Using explicit links, only about one third of all articles is linked to any other article, meaning that explicit links do not provide information about related articles for the vast majority of articles. In contrast, the network constructed with our temporal linking approach covers more than 80% of all articles. But external links raise another important issue: 95% of all explicit links connect news articles published by the same news outlet – thus ignoring competitive outlets. Temporal linking circumvents this issue by linking articles to potential targets published by all news outlets on a specific date.

Evaluation of link recall.

Explicit links only represent a fraction of articles that are related. Besides the issue of competitive news outlets mentioned above, if there exists a manually created, explicit link between two articles, they are most certainly relevant to each other and should thus also be linked by our temporal linking system. It should be noted, however, that this evaluation scenario neglects near-synonym articles reporting the same news item published by different news sites: while there might not be a temporal link between two explicitly linked articles from the *same* news outlet, there might exist a link *across different* outlets.

Treating explicit links as a gold standard, we measure link recall of the temporal linking system based on the number of articles that exist in both networks. As the result in Table 4 reveals, 82% of all explicit links also exist when applying temporal links.

	temporal links
link recall	82.62%

Table 4: Evaluation of link recall for temporal linking with respect to explicit links.

Ability to group related articles.

Due to the small fraction of explicitly linked nodes, measuring precision is not very meaningful. Recall in isolation, however, does not account for erroneous edges. Thus, we measure whether our approach groups related articles together. Taking explicitly linked articles as a ground truth for a subset of articles that are related, all pairs of connected articles should be assigned to the same story thread, i.e., cluster by our system.

In fact, 87% of all articles that are connected by explicit links are assigned to the same community. This shows that our approach does not yield arbitrary links but groups related articles.

Error analysis.

Manual error analysis revealed that most of the missing links between articles are due to non-existent or unrecognized temporal expressions. While the temporal tagger HeidelTime recognizes most of the temporal expressions, certain German compounds are currently not yet recognized, such as "Montagmorgen" (Monday morning). Most of these issues will be resolved in an upcoming version of the temporal tagger. Documents without any temporal expression are more challenging to account for. Experiments to link these articles with highly similar articles that were published within the same week yield first promising results.

5.5.2 Usage as a data exploration tool

As already noted, we regard our system as an exploration tool for end users who can adjust the parameters for similarities and the pruning threshold γ. To judge the usefulness of the system, we thus analyze the resulting information network obtained from our experiments and check whether it fulfills the required properties of a news exploration tool.

Central articles.

Intuitively, articles with a high out-degree centrality should be overview articles that mention many other articles and thus present an overview of developments. The article with the highest out-degree in the NSA data set is titled "Timeline of a scandal"[7] and was written at the end of October 2013. It links to 34 articles in the network. The overview aspect is well-represented in the link structure as the articles comprise all major events in the course of this story, beginning with the discovery of Prism in June 2013. Again, our system is able to link articles of different aspects and topics in one story line. This shows that central articles are suitable for a quick overview of the development of a story over time and are thus a suitable starting point. In the au-

[7] `http://www.faz.net/-gpf-7irh3`: October 25, 2013

tomatic data set, there are also numerous central overview articles: the article "Conflict between the Ukraine and Russia"[8], for example, reviews the major developments in the conflict between the Ukraine and Russia.

Bibliographic coupling.

Viewing our network of news articles as a citation network with links representing explicit citations of previously written articles, we compute the co-citation between articles. Intuitively, if two articles are both cited more often in the same article, they have a strong resemblance and can thus be regarded as near-duplicates. Our information network supports this hypothesis. In the NSA data set, the two articles titled "Merkel and the spying scandal" (FAZ) and "Obama and the chancelor's cellphone" (Spiegel) describe the same story covered by two different news sites and are co-cited 30 times within following news articles. In the NSA data set, about 35% of all articles are co-cited in more than 3 articles, whereas the proportion of frequently co-cited articles is higher in the automatic data set (about 50%). This is not surprising as the automatic data set comprises more news sites with potentially redundant articles from different outlets (near duplicates).

The degree of similarity between co-cited articles varies, as some of the articles are actually paraphrased copies of a news agency report, whereas other pairs describe the same story from different points of views. Bibliographic coupling and co-citation can thus be used for an analysis of news and information flow across different news sites with respect to time (e.g., what outlet published a story first).

5.5.3 Usage scenario: background information

Another application for our exploration tool is to get background information for a specific, complex news article. Figure 6 shows a shortened version of a news network for the starting article "Argument about EU posts". The original network consists of 13 news stories. The initial article (highlighted in green) mentions multiple aspects of the debate about political EU posts that are not understandable without the corresponding background knowledge. The network gives an overview of topics related to the entire story, including, among others, the stability pact. This helps to uncover dependencies across different news stories and to better understand and judge the story at hand.

5.6 Summary of our Experiments

Having applied our document linking model (Section 4.2) to real-world data, the findings of our experiments can be summarized as follows. We first investigated the influence of the *pruning threshold* γ as well as support paths to obtain an information network for our evaluation in Section 5.3. To cluster the network structure into *story threads*, we then motivated and applied *community detection* and showed why document similarity in isolation is insufficient to model the temporal evolution of complex stories (Section 5.4).

Comparing our information network to a network structure based on explicit links in Section 5.5.1 revealed that: (a) our network captures more articles than explicit links in isolation and interlinks different news outlets, (b) temporal links reproduce the vast majority of explicit links and (c) our similarity metric in combination with community detection is suitable for grouping related articles into the same story thread. Finally, we showed several applications to demonstrate the usefulness our news information network as an exploration tool in Section 5.5.2.

Discussion.

While the presented evaluation clearly shows the functionality and performance of our approach in comparison to explicit links, we are aware of the fact that additional experiments are necessary in future. To reliably determine the applicability of our system as a news exploration tool, user studies comparing our approach with alternative methods of document linking should be employed.

6. CONCLUSIONS AND ONGOING WORK

This paper presented an approach to link news articles by taking temporal expressions in articles as starting points for linking articles. The link strength between articles is determined using document similarity metrics applied to keywords and persons. We evaluated and validated the resulting network structure for a German news corpus and presented several usage scenarios for data exploration.

We are currently working on an interface that lets users explore a news network either for a specific news article or an overview over a complex news-topic in its entirety. In addition, the current architecture provides a good starting point to exploit properties of the global information network in a joint-inference setting to enhance performance, as well as integrating additional and alternative similarity metrics based on activation patterns.

7. REFERENCES

[Allan, 2002] Allan, J. (2002). *Topic detection and tracking: event-based information organization*, volume 12. Springer.

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[Brants et al., 2003] Brants, T., Chen, F., and Farahat, A. (2003). A system for new event detection. In *SIGIR '03*, pages 330–337.

[Do et al., 2009] Do, Q., Roth, D., Sammons, M., Tu, Y., and Vydiswaran, V. (2009). Robust, light-weight approaches to compute lexical similarity. *Computer Science Research and Technical Reports, University of Illinois*.

[Feng and Allan, 2009] Feng, A. and Allan, J. (2009). Incident threading for news passages. In *CIKM'09*, pages 1307–1316.

[Gillenwater et al., 2012] Gillenwater, J., Kulesza, A., and Taskar, B. (2012). Discovering diverse and salient threads in document collections. In *EMNLP-CoNLL 2012*, pages 710–720.

[Kumaran and Allan, 2005] Kumaran, G. and Allan, J. (2005). Using names and topics for new event detection. In *HLT/EMNLP 2005*, pages 121–128.

[Lavrenko et al., 2002] Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., and Thomas, S. (2002). Relevance models for topic detection and tracking. In *Human Language Technology Research*, pages 115–121.

[8] `http://sz.de/1.2021629`: June 30, 2014

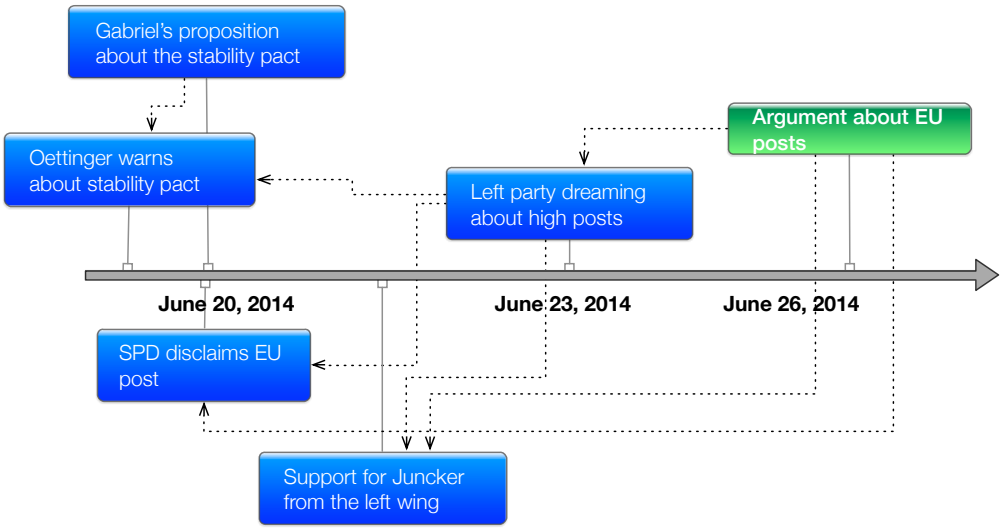

Figure 6: Sample information network demonstrating the usage of our linking approach to obtain background information for a complex topic.

[Nallapati et al., 2004] Nallapati, R., Feng, A., Peng, F., and Allan, J. (2004). Event threading within news topics. In *CIKM 2004*, pages 446–453.

[Newman, 2010] Newman, M. (2010). *Networks - An Introduction*. Oxford University Press.

[Nomoto, 2010] Nomoto, T. (2010). Two-tier similarity model for story link detection. In *CIKM'10*, pages 789–798.

[Pouliquen et al., 2008] Pouliquen, B., Steinberger, R., and Deguernel, O. (2008). Story tracking: linking similar news over time and across languages. In *MMIES '08*, pages 49–56.

[Schmid, 1995] Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *ACL SIGDAT-Workshop*.

[Shahaf and Guestrin, 2012] Shahaf, D. and Guestrin, C. (2012). Connecting two (or less) dots: Discovering structure in news articles. *TKDD*, 5(4):24.

[Shahaf et al., 2013] Shahaf, D., Yang, J., Suen, C., Jacobs, J., Wang, H., and Leskovec, J. (2013). Information cartography: Creating zoomable, large-scale maps of information. In *Knowledge Discovery and Data Mining*, KDD '13, pages 1097–1105.

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

[Vaca et al., 2014] Vaca, C. K., Mantrach, A., Jaimes, A., and Saerens, M. (2014). A time-based collective factorization for topic discovery and monitoring in news. In *WWW '14*, pages 527–538.

[Wang and Li, 2011] Wang, L. and Li, F. (2011). Story link detection based on event words. In *Computational Linguistics and Intelligent Text Processing*, pages 202–211. Springer.

[Wasserman, 1994] Wasserman, S. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge University Press.

[Yan et al., 2011] Yan, R., Wan, X., Otterbacher, J., Kong, L., Li, X., and Zhang, Y. (2011). Evolutionary timeline summarization: A balanced optimization framework via iterative substitution. In *SIGIR '11*, pages 745–754.

[Zhu and Oates, 2012] Zhu, X. and Oates, T. (2012). Finding story chains in newswire articles. In *Information Reuse and Integration*, pages 93–100.

[Zhu and Oates, 2013] Zhu, X. and Oates, T. (2013). Finding news story chains based on multi-dimensional event profiles. In *Open Research Areas in Information Retrieval*, pages 157–164.